ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

PATRIMONIO CULTURALE NELL'ECOSISTEMA DIGITALE

Ciclo 38

Settore Concorsuale: 11/A4 - SCIENZE DEL LIBRO E DEL DOCUMENTO E SCIENZE STORICO RELIGIOSE

Settore Scientifico Disciplinare: M-STO/08 - ARCHIVISTICA, BIBLIOGRAFIA E BIBLIOTECONOMIA

RAPPRESENTARE IL DIGITALE D'AUTORE: DAL FILE SYSTEM AL KNOWLEDGE GRAPH

Presentata da: Lucia Giagnolini

Coordinatore Dottorato

Francesca Tomasi

Supervisore

Francesca Tomasi

Co-supervisori

Paolo Bonora

Paola Italia

Esame finale anno 2026

Finanziato dall'Unione europea- Next Generation EU, Missione 4, Componente 2, Investimento 3.3 (D.M. 117/2023) CUP J33C22001820002

Alle radici che mi tengono salda

e alle foglie che sono cadute e hanno saputo germogliare ancora,

perché è nella loro leggerezza che l'albero ha imparato a crescere.

Ringraziamenti

Ringrazio Francesca Tomasi per aver supervisionato questo lavoro con rigore e visione, per avermi offerto gli strumenti per costruire una ricerca autonoma – conoscerne i limiti, sfidarli, e accettarli.

Ringrazio Paolo Bonora per la co-supervisione tecnica, precisa e generosa, che mi ha permesso di alzare l'asticella del progetto dove non avrei saputo arrivare da sola, e la professoressa Paola Italia per gli spunti preziosi che hanno allargato le prospettive della ricerca.

Ringrazio i revisori Emmanuela Carbé e Pierluigi Feliciati per la lettura attenta e per i suggerimenti che hanno contribuito a migliorare questa tesi.

Ringrazio /DH.arc e ADLab per il supporto alla ricerca, e in particolare Paolo Bonora, Tommaso Vitale, Silvia Samorì, Marco Serra e Matteo Pascoli per il supporto costante, tecnico e umano, nella gestione dell'Archivio Valerio Evangelisti.

Ringrazio il Centro Manoscritti dell'Università di Pavia, Giuseppe Antonelli, Sara Pizzi e Chiara Andreatta per avermi aperto questo campo di ricerca, affidandomi il progetto Pavia Archivi Digitali (PAD) e i suoi preziosi archivi.

Ringrazio l'Associazione Valerio Evangelisti – Il Sol dell'Avvenire per la fiducia accordata, per l'incredibile opportunità di lavorare sul fondo Evangelisti e di continuare a collaborare per la sua valorizzazione.

Ringrazio l'Information Systems Research Group (InfoLab) della Facoltà di Ingegneria (FEUP) dell'Università di Porto e, in particolare, le professoresse Carla Teixeira Lopes e Cristina Ribeiro, per avermi accolta a braccia aperte nel loro gruppo di ricerca. Un ringraziamento speciale a Inês Koch per la sua costante disponibilità nel confronto delle nostre ricerche e per aver contribuito a rendere il mio periodo all'estero accademicamente e umanamente meraviglioso.

Ringrazio amicə e collegə del gruppo di Digital Humanities per aver tenuto insieme le idee e le persone, dentro e fuori la ricerca – *once in the DH.ark*, *always*.

Ringrazio le mie case – Pesaro, Bologna, Porto – e chi le ha abitate con me, e chi, abitandole, mi abita ancora.

Ringrazio i miei genitori, per tutto quello che non ha bisogno di essere scritto.

Abstract

Nel contesto digitale, gli autori contemporanei si configurano come soggetti produttori di ecosistemi documentari complessi, distribuiti e stratificati su più ambienti. Le proprietà intrinseche ed estrinseche dei materiali nativi digitali, la loro articolazione strutturale e relazionale rendono insufficiente la descrizione archivistica tramite strumenti tradizionali. Mentre la comunità archivistica internazionale si orienta progressivamente verso l'adozione dei Linked Open Data (LOD), mancano modelli semantici in grado sia di rispondere alle specificità del digitale d'autore, sia di valorizzare le potenzialità di automazione e interoperabilità offerte dalla natura stessa di questi materiali.

La presente ricerca sviluppa modalità di rappresentazione dedicate al digitale d'autore attraverso un approccio integrato ed articolato su quattro livelli: indagine delle pratiche autoriali contemporanee; progettazione di un modello semantico; implementazione di processi automatizzati di descrizione basati sul modello proposto; individuazione di strumenti di analisi e consultazione.

L'indagine fenomenologica ha coinvolto cinquanta finalisti dei Premi Strega e Campiello e l'analisi dell'Archivio Valerio Evangelisti, il più ampio caso d'uso di archivi digitali d'autore finora documentato in Italia, comprendente 2,1 TB di materiali distribuiti su supporti eterogenei. Sono stati identificati pattern gestionali ricorrenti, strategie di autorappresentazione digitale e forme emergenti di "volontà d'archivio", insieme a una limitata percezione del valore documentario e del ruolo cruciale degli autori nella preservazione futura.

Lo studio dello stato dell'arte e l'analisi fenomenologica hanno permesso di individuare cinque requisiti per la modellazione del digitale d'autore: rappresentare la stratificazione fisica, logica e concettuale dei materiali digitali; introdurre misure di integrità crittografica; adottare un approccio scalabile ai metadati; documentare *provenance* e relazioni contestuali. A partire da una ricognizione di modelli esistenti (RiC-O, ArCo, ArchOnto, PREMIS), è stata sviluppata BoDi (Born-Digital Ontology), un'estensione ontologica di RiC-O che integra elementi da PREMIS, PROV-O e LRMoo con nuove classi e proprietà specifiche in grado di soddisfare i requisiti individuati e di conciliare standard di preservazione e descrizione.

Per automatizzare la descrizione dei materiali digitali, è stato implementato un processo che traduce strutture e contenuti dell'archivio in grafi RDF conformi a BoDi, articolato in cinque fasi: estrazione automatizzata di strutture e metadati; validazione; arricchimento tramite modelli generativi e conoscenza specialistica; ricostruzione dei contesti originari; anonimizzazione dei dati personali. Il workflow è stato

testato sull'Archivio Valerio Evangelisti, generando oltre 60 milioni di triple per la rappresentazione di 78.211 file organizzati in 11.135 cartelle.

L'analisi e la visualizzazione dei dati sono state realizzate attraverso query SPARQL e mediante la piattaforma ResearchSpace, traducendo il grafo in forme gerarchiche integrate con prospettive contestuali, accessibili anche a utenti non specialisti.

Nel complesso, i risultati mostrano che la convergenza tra modellazione semantica, automazione e visualizzazione consente di affrontare in modo sistematico le sfide della descrizione del digitale d'autore. L'approccio proposto offre un contributo metodologico e applicativo replicabile, capace di integrare la curatela archivistica con le potenzialità dei materiali digitali e delle tecnologie semantiche, aprendo nuove prospettive per la rappresentazione e la valorizzazione del patrimonio digitale contemporaneo.

Sommario

Indice delle figure VIII
Indice delle tabelle XII
Indice dei listati XIII
Introduzione 1
Parte I - Fenomenologia 4
1. Il digitale d'autore 4
1.1 Definizione 4
1.2 Progetti e linee di ricerca 6
1.3 Il problema della descrizione 14
2. Il soggetto produttore nel contesto digitale: cinquanta interviste d'autore 21
2.1 Introduzione 21
2.2 Metodologia 24
2.3 Risultati 27
2.3.1 Contesti di produzione documentale nell'era digitale 28
2.3.2 Modalità di gestione e archiviazione documentale 36
2.3.3 Strategie di conservazione e trasmissione 41
2.3.4 Nuove forme d'archivio e di scrittura 47
2.4 Discussione 55
3. L'Archivio Valerio Evangelisti 58
3.1 Nota biografica 58
3.2 L'archivio ibrido 60
3.3 Storia archivistica 66
3.4 "Finché c'è lotta informatica c'è speranza": Evangelisti fra tecnologia e letteratura 70
Parte II - Modellazione 81
4. Elementi costitutivi per la descrizione del digitale d'autore 81
4.1 Caratteristiche del documento digitale 81
4.2 Metadati 86
4.3 Provenance 90
4.4 Contesti 97
5. La rappresentazione semantica nella pratica archivistica 102
5.1 Ontologie per gli archivi 103
5.2 Prospettive comparative sulla modellazione semantica negli archivi 126
5.2.1 RiC-O e ArCo 127

5.2.2 RiC-O e ArchOnto ... 146
5.3 Migrazioni semantiche di inventari tradizionali ... 165
5.3.1 Approccio rule-based ... 169
5.3.2 Approccio generativo ... 173
6. Modellazione del digitale d'autore ... 185
6.1 Metodologia e approccio alla ricerca ... 185
6.2 Il modulo Born-Digital (BoDi) ... 188
6.3 Validazione del modello ... 213
Parte III - Automazione ... 218
7. Automazione del processo di rappresentazione ... 218
7.1 Sistematizzazione dell'informazione esistente ... 220
7.2 Validazione e inferenze ... 235
7.3 Arricchimento della descrizione ... 251
7.4 Ricostruzione della storia conservativa ... 264
7.5 Anonimizzazione dei dati ... 268
Parte IV - Visualizzazione ... 273
8. Scalable reading d'archivio: ipotesi e risultati ... 273
8.1 Statistiche generali ... 274
8.2 Analisi strutturale ... 276
8.3 Considerazioni sui codici hash ... 281
8.4 Tipologie di file ... 285
8.5 Distribuzione temporale delle attività ... 292
8.6 Distribuzione dei materiali relativi ai romanzi ... 304
9. Mediare il grafo: accesso e visualizzazione ... 308
9.1 La piattaforma sperimentale ... 308
9.2 Prime sperimentazioni di visualizzazione ... 311
Conclusioni ... 321
Bibliografia ... 327

Indice delle figure

Figura 2.1 Distribuzione dei generi e delle tipologie testuali prodotte dagli autori oltre alla narrativa. 29
Figura 2.2. Distribuzione delle tipologie di processo di scrittura. 30
Figura 2.3. Distribuzione delle tipologie di processi di scrittura per fasce d'età. 30
Figura 2.4. Distribuzione delle tipologie di processi di scrittura per fasce d'età (espressi in percentuale). 31
Figura 2.5. Distribuzione delle tipologie di dispositivi utilizzati per la scrittura. 34
Figura 2.6. Distribuzione dei dispositivi utilizzati per la scrittura per fasce d'età. 35
Figura 2.7. Distribuzione del numero di dispositivi utilizzati per fasce d'età espresso in percentuale. 35
Figura 2.8. Distribuzione delle strategie utilizzate per il trasferimento di file fra dispositivi. 36
Figura 2.9. Distribuzione delle finalità di stampa. 38
Figura 2.10. Distribuzione delle modalità di backup. 42
Figura 2.11. Distribuzione delle modalità di backup utilizzate suddivise per fasce d'età (gli autori hanno espresso una o più preferenze). 43
Figura 2.12. Distribuzione dell'utilizzo di uno o più dispositivi espressa in percentuale. 43
Figura 2.13. Distribuzione delle risposte riguardo la disponibilità effettiva a procedere ad un conferimento archivistico. 45
Figura 2.14. Distribuzione delle risposte sulla possibilità di adozione di un testamento digitale. 47
Figura 2.15. Distribuzione della presenza di blog o sito web d'autore. 48
Figura 2.16. Distribuzione della presenza di blog o sito web d'autore in base alle fasce d'età. 49
Figura 2.17. Distribuzione della presenza di blog o sito web d'autore in base alle fasce d'età, espresso in percentuale. 49
Figura 2.18. Distribuzione della presenza degli autori sui social media. 51
Figura 2.19. Distribuzione della presenza degli autori sui social media in base alle fasce d'età. 52
Figura 2.20. Distribuzione della presenza degli autori sui social media in base alle fasce d'età (espresse in percentuale). 52
Figura 2.21. Distribuzione delle piattaforme social adottate dagli autori. 53
Figura 3.1. Distribuzione dei materiali digitali dell'Archivio Valerio Evangelisti per supporto di memorizzazione. 63
Figura 3.2. Contenitori 3M in cui sono stati rinvenuti parte dei floppy. 68
Figura 5.1. Panoramica di RiC-CM pubblicata da ICA-EGAD nel settembre 2023. 106
Figura 5.2. Rappresentazione della struttura di ArchOnto elaborata da Inês Koch. 113

Figura 5.3. Esempio di rappresentazione della descrizione di livello “Fondo” del fondo Juízo da Índia e Mina. 115
Figura 5.4. Rappresentazione grafica del modulo Archives di ArCo. 120
Figura 5.5. PREMIS Data Model (PREMIS Editorial Committee 2015a, 6). 123
Figura 5.6. Fasi del processo di comparazione fra RiC-O e ArchOnto e arricchimento semantico. ... 147
Figura 5.7. Rappresentazione della nascita mediante ArchOnto. 148
Figura 5.8. Prima opzione per rappresentare un evento di nascita mediante RiC-O. 149
Figura 5.9. Seconda opzione per rappresentare un evento di nascita mediante RiC-O. 149
Figura 5.10. Rappresentazione dell'attività di battesimo mediante ArchOnto. 152
Figura 5.11. Prima possibilità di rappresentazione dell'attività di battesimo mediante RiC-O. 152
Figura 5.12. Seconda possibilità di rappresentazione dell'attività di battesimo mediante RiC-O. 153
Figura 5.13. Rappresentazione della risorsa archivistica e dei suoi contesti mediante ArchOnto. 155
Figura 5.14. Possibile rappresentazione della risorsa archivistica e dei suoi contesti mediante RiC-O. 155
Figura 5.15. Rappresentazione dell’estrazione e della conversione dei dati in ArchOnto e CRMdig. 160
Figura 5.16. Possibile rappresentazione dell'estrazione e della conversione dei dati in RiC-O. 160
Figura 5.17. Entità riconosciute nel testo e relativo classificazione. 175
Figura 5.18. Dipendenze sintattiche relative alla prima frase. 175
Figura 5.19. Grafo dell’interpretazione delle condizioni di nascita elaborato da FRED (sulla base del testo tradotto in inglese). 176
Figura 5.20. Grafo dell’interpretazione delle condizioni di nascita elaborato da FRED con il testo originale in italiano. 176
Figura 5.21. Output di GPT-4o relativo alle condizioni di nascita di Andrea Costa (visualizzazione creata da mermaid.live). 178
Figura 5.22. Diagramma di flusso della pipeline implementata. 180
Figura 6.1. Modellazione grafica dei tre livelli di astrazione (contenutistico, logico, fisico). 194
Figura 6.2. Modellazione grafica del concetto di Fixity. 197
Figura 6.3. Modellazione grafica dei metadati nativi cp:revision e FileSize (senza dichiarazione di provenienza). 200
Figura 6.4. Modellazione grafica della chain of custody attraverso relazioni di derivazione, storage locations e storage media. 203
Figura 6.5. Modellazione grafica della provenance in relazione all’estrazione dei metadati. 209

Figura 6.6. Modellazione grafica dell'integrazione tra contesto archivistico verticale (gerarchia file-cartella) e contesto bibliografico (collegamento all'opera letteraria Tortuga). 212
Figura 6.7. Schema delle classi introdotte da BoDi e delle ontologie di riferimento RiC-O, PREMIS, PROV-O e LRMoo. Per chiarezza, sono mostrate solo le classi di RiC-O estese da BoDi. 213
Figura 7.1. Workflow in cinque fasi per la trasformazione di archivi nativi digitali in una knowledge base RDF conforme al modello BoDi. 219
Figura 7.2. Prima fase del workflow: sistematizzazione dell'informazione esistente. 220
Figura 7.3. Rappresentazione grafica degli step previsti dalla prima fase del workflow. 221
Figura 7.4. Esempio di struttura della knowledge base al termine della prima fase del workflow, che integra la rappresentazione semantica della struttura archivistica con metadati tecnici, relazioni e informazioni di provenance, ossia la baseline per le fasi successive. 234
Figura 7.5. Seconda fase del workflow: validazione e inferenze. 235
Figura 7.6. Terza fase del workflow: arricchimento della descrizione mediante conoscenza specialistica e modelli generativi. 251
Figura 7.7. Quarta fase del workflow: ricostruzione della storia conservativa. 264
Figura 7.8. Quinta fase del workflow: anonimizzazione dei dati. 268
Figura 8.1. Distribuzione delle macro-tipologie documentarie individuate nell'Archivio Evangelisti. 288
Figura 8.2. Distribuzione temporale dei file basata sulle date di creazione. 292
Figura 8.3. Distribuzione temporale dei file creati da Valerio Evangelisti basata sulle date di creazione. 293
Figura 8.4. Distribuzione temporale dei file dell'Archivio Valerio Evangelisti: confronto tra il corpus totale e il dataset attribuibile all'autore sulla base delle date di creazione. 294
Figura 8.5. Heatmap della distribuzione temporale dei file dell'Archivio Valerio Evangelisti basata sulle date di creazione, con dettaglio mensile. 295
Figura 8.6. Heatmap della distribuzione temporale dei file attribuibili a Valerio Evangelisti, basata su creazione e ultima modifica, con dettaglio mensile. 296
Figura 8.7. Distribuzione temporale dei file basata sulle data di modifica. 297
Figura 8.8. Heatmap della distribuzione temporale dei file dell'Archivio Valerio Evangelisti basata sulle date di creazione, con dettaglio mensile. 298
Figura 8.9. Distribuzione temporale dei file creati da Valerio Evangelisti basata sulle date di modifica. 298

Figura 8.10. Distribuzione temporale dei file dell'Archivio Valerio Evangelisti: confronto tra il corpus totale e il dataset attribuibile all'autore sulla base delle date di modifica. 299
Figura 8.11. Heatmap della distribuzione temporale dei file dell'Archivio Valerio Evangelisti basata sulle date di modifica, con dettaglio mensile. 300
Figura 9.1. Esempio di visualizzazione di un risultato di ricerca contestualizzato all'interno della struttura gerarchica. 313
Figura 9.2. Struttura informativa dei materiali nativi digitali di Evangelisti, con distinzione tra computer principale, hard drive esterno e riversamenti da floppy disk. 316
Figura 9.3. Struttura ad albero dei materiali digitali nativi di Evangelisti e finestra di dettaglio sui materiali dell'hard drive del computer di Evangelisti. 317
Figura 9.4. Visualizzazione delle informazioni di provenance e dei processi di migrazione tra supporti. 317
Figura 9.5. Scheda descrittiva del file "00-Indice.Rtf" con focus su informazioni descrittive di anteprima, riferimento al codice hash e all'opera correlata. 318
Figura 9.6. Scheda descrittiva del file "00-Indice.Rtf" con focus sui metadati tecnici. 319
Figura 9.7. Informazioni di provenance relative al processo di estrazione dei metadati dal file "00 - Indice.rtf". 319

Indice delle tabelle

Tabella 5.1 Confronto dei principali modelli di riferimento per la rappresentazione archivistica: RiC-O, ArCo Archive, ArchOnto e PREMIS. 125
Tabella 5.2. Allineamenti diretti fra classi di RiC-O e ArCo. 129
Tabella 5.3. Allineamenti gerarchici fra classi di RiC-O e ArCO. 129
Tabella 5.4. Allineamenti diretti fra object property di RiC-O e ArCo. 131
Tabella 5.5. Allineamenti gerachici fra object property di RiC-O e ArCo. 132
Tabella 5.6. Allineamenti diretti fra data property di RiC-O e ArCo. 133
Tabella 5.7. Comparazione delle caratteristiche generali di ArchOnto e RiC-O. 163
Tabella 7.1. Allineamento tra metadata types provenienti da strumenti di estrazione diversi. 245
Tabella 7.2. Media types presenti nell'Archivio Valerio Evangelisti e relative frequenze. 248
Tabella 7.3. Relazioni tra cicli narrativi e romanzi nell'opera di Valerio Evangelisti. 254
Tabella 8.1. Distribuzione dei record e dei record set per tipologia di supporto. 275
Tabella 8.2. Distribuzione delle entità bodi:TechnicalMetadata per strumento di estrazione e tipologia di supporto. 275
Tabella 8.3. Distribuzione del numero di tipologie di metadati (bodi:TechnicalMetadataType) per strumento di estrazione e tipologia di supporto. 276
Tabella 8.4. Distribuzione dei file per livello di profondità, supporto e grafo di appartenenza. 277
Tabella 8.5. Distribuzione delle cartelle per livello di profondità, supporto e grafo di appartenenza. . 279
Tabella 8.6. Distribuzione media type per supporto. 287
Tabella 8.7. Distribuzione dei file e delle cartelle correlate ai romanzi di Valerio Evangelisti nei diversi supporti di memorizzazione (hard disk principale, hard disk esterno e floppy disk). I valori percentuali indicano la proporzione di ciascun supporto rispetto al totale dei materiali associati a ogni opera. 306
Tabella 8.8. Distribuzione dei file e delle cartelle riferibili ai cicli e alle trilogie di Valerio Evangelisti nei diversi supporti di memorizzazione. I valori percentuali indicano la proporzione di ciascun supporto rispetto al totale dei materiali associati a ogni ciclo o trilogia. 306

Indice dei listati

Listato 5.1. Registrazione di battesimo di Maria da Silva in XML EAD. 171
Listato 5.2. Registrazione di battesimo di Maria da Silva in XML con tag aggiuntivi derivati dal processo di knowledge extraction. 172
Listato 5.3. Rappresentazione RDF dell'evento di nascita di Andrea Costa secondo RiC-O che formalizza soggetti, data, luogo e relazioni a partire dalla descrizione testuale. 182
Listato 6.1. Modellazione espressa in Turtle dell'istanza "Tortuga.docx" nel contesto del file system. 192
Listato 6.2. Modellazione espressa in Turtle dell'istanza "Tortuga.docx" nel contesto del file system. 194
Listato 6.3. Modellazione espressa in Turtle del concetto di Fixity. 196
Listato 6.4. Modellazione espressa in Turtle dei metadati nativi cp:revision e FileSize (senza dichiarazione di provenienza). 199
Listato 6.5. Modellazione espressa in Turtle della chain of custody attraverso relazioni di derivazione, storage locations e storage media. 202
Listato 6.6. Modellazione espressa in Turtle della provenance in relazione all'estrazione dei metadati. 209
Listato 6.7. Modellazione espressa in Turtle dell'integrazione tra contesto archivistico verticale (gerarchia file-cartella) e contesto bibliografico (collegamento all'opera letteraria Tortuga). 212
Listato 7.1. Pseudocodice della pipeline per la prima fase del workflow. 232
Listato 7.2. Pseudocodice per la generazione del grafo RDF delle opere narrative di Valerio Evangelisti e il loro collegamento con i materiali dell'archivio. 255
Listato 7.3. Esempio di metadati tecnici estratti da un file RTF. 261
Listato 7.4. Pseudocodice per la generazione automatica di descrizioni tecniche tramite RAG con modello LLM locale, gestione cache persistente e inserimento incrementale nel triplestore. 263
Listato 7.5. Pseudocodice per la documentazione della chain of custody e delle relazioni di derivazione tra istanziazioni dei supporti dell'archivio. 267
Listato 7.6. Pseudocodice rappresentante l'algoritmo di anonimizzazione selettiva per le entità archivistiche. 272

Ogni vita è un'enciclopedia, una biblioteca,
un inventario d'oggetti, un campionario di stili,
dove tutto può essere continuamente rimescolato
e riordinato in tutti i modi possibili.

Italo Calvino

Introduzione

Gli archivi nativi digitali d'autore si configurano come un terreno di indagine complesso, al crocevia di problematiche concettuali, gestionali e conservative. La natura digitale della produzione documentaria contemporanea, generata in ambienti eterogenei e caratterizzata da multiple stratificazioni, non può essere adeguatamente descritta con i soli strumenti tradizionali. In questo quadro, la rappresentazione semantica, basata sui principi dei Linked Open Data (LOD), offre la possibilità di descrivere documenti, contesti e provenienze in modo strutturato e interoperabile.

Questa ricerca indaga le modalità di rappresentazione degli archivi nativi digitali d'autore attraverso modelli a semantica esplicita, con l'obiettivo di restituire in modo adeguato la complessità del loro ecosistema, superando descrizioni statiche e monodimensionali e valorizzando, al contempo, le potenzialità di automazione dei processi di descrizione rese possibili dalla natura digitale dei materiali.

La tesi ha l'obiettivo di rispondere a quattro domande di ricerca (*research questions* - RQ):

- RQ1 - Fenomenologia: come si configura un archivio d'autore contemporaneo? Quali pratiche adottano gli autori nella produzione, gestione e conservazione dei propri materiali?
- RQ2 - Modellazione: come rappresentare il digitale d'autore, analizzando la complessità e le relazioni contestuali?
- RQ3 - Automazione: in che modo le proprietà intrinseche del digitale possono essere sfruttate per automatizzare e facilitare i processi di descrizione archivistica?
- RQ4 - Visualizzazione: come presentare e analizzare i dati ottenuti? Quali possibili soluzioni per mostrare all'utente finale i risultati della ricerca?

Parte I - Fenomenologia

Il capitolo 1 introduce il tema del digitale d'autore, soffermandosi sulla definizione (cap. 1.1), sulle linee di ricerca che si sono sviluppate sul tema (cap. 1.2) e sul problema della descrizione in particolare (cap. 1.3).

Il capitolo 2 è destinato all'analisi del punto di vista dei soggetti produttori attraverso l'elaborazione di un'indagine su cinquanta finalisti dei Premi Strega e Campiello (1985-2024) (cap. 2.1). La metodologia si basa sulla raccolta di interviste strutturate (cap. 2.2) per indagare dispositivi utilizzati, gestione e archiviazione dei documenti, strategie di conservazione e nuove forme di scrittura digitale (cap. 2.3). L'analisi evidenzia la centralità del digitale, pur con una minima persistenza del cartaceo, e mostra pratiche eterogenee di gestione, backup e trasmissione dei materiali (cap. 2.4).

Il capitolo 3 introduce il caso di studio, l'Archivio di Valerio Evangelisti (1952-2022), che presenta tanto una consistente sezione digitale (2,1 TB di dati distribuiti su supporti eterogenei), quanto una componente analogica (5 buste, 0,5 metri lineari). Dopo la presentazione della nota biografica, (cap. 3.1) si descrive il fondo (cap. 3.2) e se ne ricostruisce la storia archivistica (cap. 3.3). Il capitolo si conclude con un'analisi del rapporto di Evangelisti con la tecnologia, che ricopre un ruolo determinante nel processo creativo dell'autore (cap. 3.4).

Parte II - Modellazione

Il capitolo 4 identifica le dimensioni del digitale d'autore con implicazioni dirette sulla pratica descrittiva. Si intende distinguere, in particolare, tra caratteristiche intrinseche ed estrinseche dei documenti (cap. 4.1), discutendo il ruolo dei metadati (cap. 4.2), del principio di provenienza (*provenance*) (cap. 4.3) e dei contesti (cap. 4.4).

Il capitolo 5 individua gli strumenti per rappresentare la fenomenologia identificata nei capitoli precedenti. Dopo aver delineato le premesse teoriche dei LOD e il loro ruolo nella descrizione archivistica, vengono confrontati i principali modelli a semantica esplicita già utilizzati in ambito archivistico (RiC-O, ArCo, ArchOnto, PREMIS) (cap. 5.1) e proposte ipotesi di allineamento fra modelli (cap. 5.2). Una sezione è dedicata alle sperimentazioni di migrazione da inventari tradizionali a *knowledge base* semantiche, condotte sia attraverso strumenti *rule-based*, sia mediante *Large Language Models* (LLMs), evidenziando i benefici e i margini di innovazione che ne derivano (cap. 5.3).

Nel capitolo 6 viene sviluppato nello specifico il tema della modellazione del digitale d'autore. Dopo aver delineato i requisiti a cui un tale modello dovrebbe rispondere (cap. 6.1), vengono presentate le scelte di design che hanno portato allo sviluppo dell'ontologia formale BoDi (Born-Digital Ontology), un'estensione dell'ontologia RiC-O che integra elementi delle ontologie PREMIS, PROV-O e LRMoo, arricchita da nuove classi e proprietà specificamente pensate per gli archivi digitali d'autore (cap. 6.2). Infine, si presenta la valutazione del modello attraverso verifiche formali e sperimentazioni applicative (cap. 6.3).

Parte III - Automazione

L'applicazione concreta del modello proposto al caso Evangelisti viene illustrata in un *workflow* descritto nel capitolo 7, che a partire dalle cartelle dell'archivio produce una *knowledge base* in forma di grafo RDF. Il *workflow* è strutturato nelle seguenti fasi: sistematizzazione delle informazioni delle *directory* dell'archivio (con estrazione automatizzata della struttura e dei metadati) (cap. 7.1); validazione *rule-based* (cap. 7.2); arricchimento della descrizione grazie a conoscenze specialistiche e modelli generativi

(7.3); ricostruzione dei contesti di provenienza originari (7.4). Infine, il *workflow* prevede una fase di preparazione della *knowledge base* ai fini della pubblicazione gestendo l'anonimizzazione dei dati personali (7.5).

Parte IV - Visualizzazione

Il capitolo 8 presenta i risultati ottenuti e ipotesi di *distant* e *scalable reading* d'archivio, ovvero di analisi complessiva dei dati prodotti. Grazie al grafo RDF generato, composto da oltre sessanta milioni di triple, è stato possibile: condurre analisi statistiche generali (cap. 8.1) e strutturali (cap. 8.2); presentare la distribuzione delle tipologie di file (cap. 8.3); individuare i codici hash come chiavi di esplorazione dell'archivio (cap. 8.4); e infine analizzare l'attività di Evangelisti nel tempo, ovvero la creazione di contenuti digitali (cap. 8.5), con un focus sulle risorse legate alla produzione dei romanzi (cap. 8.6).

Il capitolo 9 presenta un'applicazione di consultazione dei dati descrittivi dell'Archivio Valerio Evangelisti tramite la piattaforma ResearchSpace, illustrandone l'architettura, le funzionalità principali (9.1) e le prime sperimentazioni di visualizzazione dei dati (9.2), che evidenziano come i LOD possano essere accessibili anche attraverso modalità di esplorazione più tradizionali e intuitive.

Infine, le conclusioni discutono l'impatto e le ricadute della proposta nella pratica archivistica, le limitazioni metodologiche e tecnologiche riscontrate, e le prospettive future di sperimentazione e ricerca.

Parte I - Fenomenologia

1. Il digitale d'autore

Storicamente, la memoria letteraria si è sedimentata attraverso documenti materiali e tangibili. Con uno "strappo nel cielo di carta", il digitale ha ridefinito gli archivi d'autore, sollevando questioni inedite e multidisciplinari. Per comprendere come si configura un archivio d'autore contemporaneo e quali pratiche caratterizzano la produzione, gestione e conservazione dei materiali (RQ1), è necessario innanzitutto definire il campo d'indagine e ricostruire lo stato dell'arte. Partendo dalla definizione di "digitale d'autore", questo capitolo traccia una panoramica dei progetti e delle linee di ricerca che si sono sviluppate sul tema, concludendo con un focus sulle pratiche di descrizione.

1.1 Definizione

Nel 2010 l'archivista Ricky Erway pubblica il breve saggio "Defining Born Digital", in cui propone una classificazione delle diverse tipologie di materiali nativi digitali e ne analizza i principali rischi di perdita, dovuti al degrado delle informazioni e all'obsolescenza tecnologica (2010). Nel suo vademecum, Erway introduce una definizione sintetica e incisiva: «Born-digital resources are items created and managed in digital form» (2010, 1).

Il concetto di "digitale d'autore" nasce successivamente, nell'ambito degli studi italiani sugli archivi nativi digitali. Emmanuela Carbé lo conia nel 2017 come traduzione e rielaborazione del termine inglese *born-digital (literary) archives*:

> un'entità o un'aggregazione di entità istanziate su supporto digitale da un'autrice o da un autore, e/o da altri soggetti produttori in interrelazione con l'autrice o l'autore all'interno di un determinato contesto; le entità, così come le eventuali relazioni tra entità, possono costituire gli eventi dell'archivio. Si configurano quindi come digitale d'autore: materiali digitali, per esempio memorizzati in piattaforme cloud o nella memoria di massa di dispositivi, originati o modificati dall'autore (es. foto, documenti di testo, audio, video, configurazioni ecc.); materiali originati o modificati da terzi quando entrano in relazione con l'autore (es. la bozza di stampa con correzioni di un autore, una stesura d'autore con commenti di un editore ecc.); materiali scaricati dal web su un dispositivo (es. la pagina di un'enciclopedia online salvata localmente); materiali digitali prodotti online, che possono anche includere conversazioni con altri soggetti (thread in siti web o social network, e-mail, messaggi istantanei ecc.) (2023, 10).

In una definizione più canonicamente archivistica, il digitale d'autore identifica l'insieme dei materiali prodotti, raccolti e utilizzati da un autore all'interno di ambienti digitali. Questa produzione rappresenta un'evidenza documentaria stratificata, che include non solo il contenuto, ma anche i metadati, i contesti di produzione e le tracce materiali del processo creativo. Carbé sottolinea, inoltre, come ogni entità dell'archivio sia istanziata con specifiche referenze contestuali e temporali, anche se non sempre correttamente preservate. Ne consegue che «i soggetti conservatori (intesi come archivi, biblioteche e musei), nel prendere in carico il digitale d'autore conservano la fotografia di quel dato momento, l'hic et nunc: in linea di principio un processo non dissimile a quello di un archivio cartaceo, ma che nei fatti produce degli archivi profondamente diversi» (Carbé 2023, 10).

Per questo motivo, il digitale d'autore rappresenta un ambito disciplinare in fase di sviluppo che pone importanti sfide metodologiche e operative alle istituzioni culturali non soltanto dal lato della conservazione, ma anche nelle dimensioni di gestione, descrizione e accessibilità (Ries e Palkó 2019). Le complessità specifiche derivano da una molteplicità di fattori interconnessi che si manifestano su diversi livelli concettuali e operativi. Innanzi tutto, è difficile identificarne i confini: questi archivi sono caratterizzati da una dispersione naturale, diffusi tra molteplici dispositivi e ambienti digitali. I documenti non risiedono più in luoghi tangibili, ma sono disseminati tra *file system* locali, servizi di *cloud storage* (OneDrive, Dropbox, Google Drive ecc.), social media, applicazioni di messaggistica e altre piattaforme (Allegrezza 2020, 305–9).

Questo ecosistema digitale, in costante evoluzione, contribuisce a rendere l'archivio personale una realtà fluida, difficilmente delimitabile a priori. A differenza dei sistemi di gestione documentale istituzionali, caratterizzati da protocolli di produzione e gestione standardizzati e definiti ex ante[1], gli archivi digitali personali si presentano come entità idiosincratiche strutturate secondo logiche spesso estranee alle buone pratiche. Ogni autore sviluppa un rapporto peculiare con le tecnologie, modellato da abitudini personali e influenzato dai cambiamenti del panorama digitale. Buona parte delle tecnologie che usano e che hanno utilizzato sono destinate ad evolversi se non a sparire: basti pensare alla parabola dei blog che, dopo un periodo molto popolare nella produzione autoriale, hanno visto un declino quasi totale con l'affermarsi dei social media (Mohammed 2017).

Parallelamente, anche il progressivo abbandono di alcuni supporti fisici, come CD-ROM e floppy disk, ha comportato l'inaccessibilità di molte tracce documentarie a causa dell'obsolescenza dei supporti e dei

[1] Si vedano, ad esempio, le *Linee guida sulla formazione, gestione e conservazione di documenti informatici* prodotte dall'Agenzia per l'Italia Digitale (AgID) (2021).

relativi dispositivi di lettura. La questione dell'obsolescenza tecnologica è chiaramente centrale e riguarda tanto i supporti fisici quanto i formati dei file, talvolta proprietari e soggetti a rapida dismissione. In questo quadro, senza una gestione preventiva e consapevole, interi fondi rischiano di diventare inaccessibili (Allegrezza 2021). A questo scenario si aggiungono le criticità legate ai processi di autenticazione: la perdita delle credenziali di accesso, come nome utente e password, può rendere irrecuperabili i contenuti conservati su piattaforme digitali, e anche l'intervento di specialisti in informatica forense non garantisce sempre esiti risolutivi (d'Arminio Monforte 2025).

Questo quadro impone l'elaborazione di metodologie differenti da quelle tradizionalmente applicate ai materiali cartacei. Gli operatori del settore devono adottare un approccio curatoriale dinamico, fondato sull'aggiornamento delle competenze e sull'implementazione di strategie mirate all'acquisizione, gestione, conservazione e accesso di archivi digitali complessi. Nei capitoli successivi saranno esaminati progetti e sperimentazioni che mostrano come istituzioni, comunità e ricercatori abbiano raccolto questa sfida.

1.2 Progetti e linee di ricerca

La comunità archivistica non è mai stata insensibile all'impatto del cambiamento tecnologico: già negli anni Ottanta le problematiche legate ai documenti elettronici erano oggetto di dibattito in conferenze e nella letteratura specialistica (Langdon 2016). I primi progetti dedicati al tema degli archivi privati digitali si sono concentrati, comprensibilmente, sulla preservazione nell'ottica di scongiurare la paventata *Digital Dark Age* (Kuny 1997), ossia il fenomeno per cui informazioni storiche digitali rischiano di andare perdute a causa di formati di file, software o hardware obsoleti che diventano corrotti, difficilmente reperibili o inaccessibili con l'evoluzione delle tecnologie e il degrado dei supporti[2].

[2] Un contributo fondamentale alla riflessione teorica e metodologica sulla preservazione degli archivi digitali, seppure con un'attenzione rivolta principalmente agli archivi istituzionali e ai documenti prodotti in contesti governativi, è rappresentato dal progetto InterPARES (International Research on Permanent Authentic Records in Electronic Systems), avviato nel 1999 da Luciana Duranti presso l'Università della British Columbia come consorzio internazionale tra studiosi di archivistica, esperti di ingegneria informatica, università e istituzioni. La fase 1 del progetto (1999-2001) ha affrontato la preservazione a lungo termine dei documenti creati e gestiti in database e sistemi di gestione documentale; la fase 2 (2002-2007) si è concentrata sui documenti prodotti in ambienti dinamici e interattivi, nell'ambito di attività scientifiche, artistiche e governative; la fase 3 (2007-2012, *TEAM*) ha applicato i risultati delle prime due fasi in istituzioni archivistiche di piccole e medie dimensioni, verificandone la trasferibilità operativa; la fase 4 (2013-2019, InterPARES Trust) ha analizzato i documenti digitali nei contesti della rete, approfondendo le questioni di affidabilità e responsabilità negli ambienti online. Infine, la fase 5 (2021-2026, InterPARES Trust AI) è dedicata allo sviluppo e all'applicazione di tecnologie di intelligenza artificiale per sostenere l'accessibilità e la gestione dei documenti pubblici digitali. Pur essendo orientato prevalentemente verso archivi istituzionali, i risultati delle varie fasi del progetto sono un punto di riferimento per l'archivistica digitale e hanno fornito una base teorica applicabile anche agli archivi privati e d'autore.

Carbé ha tracciato un quadro articolato delle iniziative nazionali e internazionali specificamente dedicate agli archivi letterari *born-digital* (Carbé 2023, 24-43)[3]. Tale mappatura evidenzia come sul piano internazionale, gli Stati Uniti e la Gran Bretagna rappresentino indiscutibilmente il contesto più avanzato e strutturato. Le iniziative pionieristiche avviate tra il 2009 e il 2010 sotto l'impulso di Matthew Kirschenbaum hanno dato vita a una rete di collaborazione tra la Stuart A. Rose Library dell'Emory University[4], l'Harry Ransom Center della University of Texas[5] e il centro Digital Humanities della University of Maryland (MITH)[6], stabilendo per la prima volta metodologie gestionali condivise che hanno aperto la strada a un progressivo coinvolgimento di numerose istituzioni americane nell'acquisizione di materiali digitali[7].

Il progetto *Born Digital Collections: An Inter-Institutional Model for Stewardship* (AIMS) (2012), a cura della University of Virginia Library, con la collaborazione della Stanford University, the University of Hull e Yale University, ha delineato uno dei primi modelli condivisi per la gestione e conservazione delle collezioni digitali. Basato su un approccio interistituzionale e orientato alle buone pratiche, il progetto ha definito principi operativi per una s*tewardship* efficace, sottolineando al contempo la necessità di soluzioni flessibili e l'importanza della formazione professionale degli archivisti nel contesto della gestione del digitale d'autore.

Sempre nel contesto statunitense, un importante contributo metodologico è rappresentato dal progetto *Demystifying Born Digital* dell'Online Computer Library Center (OCLC), nato dalla constatazione che

[3] Di seguito si richiamano i casi più emblematici, rimandando allo studio di Carbé per un quadro più dettagliato fino al 2023.

[4] L'archivio di Salman Rushdie presso la Stuart A. Rose Library è diventato paradigmatico per l'implementazione di strategie che combinano emulazione e migrazione. Il fondo, costituito da materiali cartacei e digitali provenienti da computer Mac degli anni Novanta, ha permesso di sviluppare un flusso di lavoro completo che va dall'acquisizione alla consultazione virtuale. La metodologia adottata prevede sia la ricostruzione dell'ambiente informatico originale attraverso l'emulazione, sia la conversione dei file in formati aperti e duraturi come PDF/A e TIFF, garantendo un duplice livello di accesso e preservazione (Carroll et al. 2011; Alexander 2015).

[5] Attualmente l'Harry Ransom Center conserva 138 collezioni con materiali *born-digital*, per un totale di 8TB di materiali processati disponibili agli utenti e 21TB di copie complete archiviate presso il Texas Advanced Computing Center (TACC). Il centro ha sviluppato un workflow sofisticato basato su strumenti come BitCurator e GuyMager per l'imaging, con accesso attraverso laptop dedicati nella *reading room* e partnership innovative come quella con EaaSI (Emulation-as-a-Service Infrastructure) per l'emulazione di software obsoleti, permettendo casi d'uso pionieristici come l'accesso al videogioco "Puppet Motel" di Laurie Anderson e progetti didattici di *digital forensics* (Edwards 2025).

[6] Il MITH ha sviluppato competenze specifiche nella preservazione di letteratura elettronica attraverso l'archivio di Deena Larsen, autrice di ipertesti e collezionatrice di archivi contenenti letteratura elettronica (Stollar Peters 2006).

[7] La rete statunitense si è progressivamente estesa ad altre istituzioni. Carbé (2023, 34-37) riporta la Yale University (con acquisizioni di archivi ibridi dal 2007, a partire dai materiali di George Whitmore e James Welch (Forstrom 2009)); la Houghton Library di Harvard (con 50 floppy disk da 5¼ pollici di John Updike («John Updike Papers, 1940-2009 (MS Am 1793)», s.d.)); la UCLA (con hard disk di Susan Sontag e un progetto di *virtual reading room* per i materiali di Richard Rorty (Light 2014; Carbé 2023, 37)); la Princeton University (che conserva il fondo ibrido di Toni Morrison dal 2014 (Colón-Marrero e Hughes 2015)).

le biblioteche statunitensi stavano rimanendo pericolosamente indietro nell'acquisizione e nel processamento di materiali digitali. Il progetto ha prodotto una serie di report pratici e formativi, tra cui *Defining Born Digital* (Erway 2010), *You've Got to Walk Before You Can Run: First Steps for Managing Born-digital Content Received on Physical Media* (Erway 2012) e *Walk This Way: Detailed Steps for Transferring Born-digital Content from Media You Can Read In-house* (Barrera-Gomez e Erway 2013). Coordinato da un gruppo di lavoro che includeva esperti come Matthew Kirschenbaum, Gabriela Redwine, Ricky Erway e altri professionisti, il progetto ha avuto un impatto importante sul settore, ispirando iniziative come la *SAA Jump In Initiative*[8] e fornendo linee guida pratiche per istituzioni che si trovavano ad affrontare per la prima volta il problema.

Il Regno Unito ha contribuito considerevolmente allo sviluppo del settore, a partire dal progetto PARADIGM (The Personal Archives Accessible in Digital Media)[9]. Il progetto nasce dalla collaborazione tra le principali biblioteche di ricerca delle Università di Oxford e Manchester per esplorare le questioni relative alla conservazione degli archivi privati digitali attraverso l'esperienza pratica di acquisizione dei documenti di alcuni politici britannici. A PARADIGM si deve la pubblicazione del manuale *PARADIGM Workbook on Digital Private Papers*, contenente linee guida riguardanti le problematiche connesse all'archiviazione di documenti personali in formato digitale: acquisizione, gestione, scarto, preservazione e catalogazione (Bodleian Libraries and Gardens e John Rylands Library 2007).

Di lì a poco, presso la British Library, Neil Beagrie e Jeremy Leighton John otterranno un finanziamento per *Digital Lives: Research collections for the 21st Century* (2007-2009). Il contributo più significativo del progetto ai lavori della British Library è stato quello di delineare, per la prima volta, un flusso di lavoro basato sulla combinazione di metodologie di *digital forensics* per la cattura, l'autenticazione e per la migrazione dei formati di file finalizzata a una interoperabilità duratura dei materiali (Leighton John 2009). A partire da questo progetto, la British Library assume un ruolo determinante nel proseguire la ricerca su questo tracciato, strutturando nel tempo un approccio metodologico particolarmente sofisticato, che Callum McKean suddivide in tre periodi di sviluppo: il periodo 2005-2015, caratterizzato dal confronto con la *Digital Dark Age*, con l'avvio del progetto di Beagrie e Leighton John; il periodo 2016-2018, che ha affrontato quello che Jonathan Pledge e Eleanor Dickens hanno definito il «somewhat

[8] https://www2.archivists.org/groups/manuscript-repositories-section/jump-in-initiative.

[9] Il sito ufficiale del progetto è stato dismesso a causa dell'obsolescenza delle tecnologie su cui era basato, ma una versione archiviata rimane accessibile tramite Internet Archive all'indirizzo: https://wayback.archive-it.org/org-467/20170930070055/http://www.paradigm.ac.uk/.

intractable problem» dell'accesso (2018, 59), sviluppando un workflow basato, da un lato, sulla cattura forense per la preservazione e, dall'altro, sulla migrazione verso PDF/A per garantire l'accesso agli utenti della British Library[10]. L'evoluzione più recente (2019-2024) ha segnato una svolta decisiva verso l'automazione, superando i limiti del workflow del 2017 che, pur rappresentando una novità nel panorama britannico, risultava sostenibile solo per collezioni di piccole dimensioni. Centrale, in quest'ultima fase, lo sviluppo dell'Automated Migration Workflow for Personal Digital Archives (AMW), che ha rivoluzionato i tempi di elaborazione, riducendo drasticamente la necessità di interventi manuali per operazioni quali la cattura dei contenuti, l'elaborazione dei metadati e la migrazione verso formati di accesso, che rappresentavano le fasi più dispendiose in termini di tempo e maggiormente soggette a errori umani del *workflow* precedente (McKean 2025, 109-18). Grazie all'utilizzo del *Rapid Assessment Model* della Digital Preservation Coalition (DPC)[11], il Born Digital Cataloguing Working Group della British Library ha ottenuto, nel 2021, l'inclusione degli archivi nativi digitali di persona nel *Minimum Preservation Tool* (MPT) della biblioteca[12] (McKean 2025, 110).

Sempre nel contesto anglosassone, si segnala anche il British Archive for Contemporary Writing (BACW) dell'University East Anglia (UEA) che ha sviluppato un modello *storehouse* per la gestione di archivi letterari contemporanei, implementando un'infrastruttura di preservazione digitale basata sul software Preservica[13] e sui *toolkits* della DPC, ma calibrata su vincoli di risorse e capacità istituzionali ridotte (Gooding et al. 2019; Busby 2024)[14] .

[10] Il *workflow* sviluppato in questa fase è stato sperimentato sull'archivio della poetessa Wendy Cope, composto da 76 floppy disk, materiali prodotti con computer Amstrad e una chiavetta USB da 16 GB (Pledge e Dickens 2018).

[11] https://www.dpconline.org/digipres/implement-digipres/dpc-ram. La Digital Preservation Coalition (DPC), fondata nel 2002, è un'organizzazione internazionale no-profit che riunisce istituzioni pubbliche e private per affrontare le sfide della conservazione digitale. Il suo scopo è sviluppare risorse, strumenti e standard volti a garantire l'accesso sostenibile e duraturo ai contenuti digitali, superando problemi legati all'obsolescenza tecnologica, al degrado dei supporti e ai cambiamenti organizzativi. La DPC promuove le buone pratiche attraverso la pubblicazione dei *Technology Watch Reports* dal 2004 e la formazione professionale, attraverso il *Digital Preservation Handbook* e workshop dedicati. Inoltre, svolge attività di *advocacy* e sensibilizzazione ai temi legati al digitale, organizza *Digital Preservation Awards* biennali e riconosce contributi individuali tramite la *DPC Fellowship*. La *membership* della DPC, oggi composta da oltre 150 organizzazioni tra biblioteche, archivi, università ed enti internazionali, mantiene una posizione neutrale rispetto ai fornitori tecnologici e promuove la collaborazione globale per la conservazione digitale. Per approfondimenti si rimanda al sito: https://www.dpconline.org/

[12] Il MPT è una raccolta di *utility* in Python che fornisce una soluzione provvisoria di conservazione dei file, garantendo backup remoti *off-site* e controlli regolari di integrità, eliminando così la dipendenza da hard disk esterni e backup locali (Beaman 2020).

[13] https://preservica.com/.

[14] Il British Archive for Contemporary Writing è stato lanciato presso la UEA nel 2015, con collezioni di autori Premio Nobel come Doris Lessing e Nadine Gordimer, oltre a materiali di JD Salinger. Più recentemente, sono stati acquisiti i fondi di Naomi Alderman e Richard Beard, che contengono anche materiali nativi digitali (Gooding et al. 2019).

Si evidenziano, inoltre, le ricerche di Lise Jaillant (Loughborough University) (Jaillant 2019, 2022b, 2022a, 2024; Jaillant et al. 2022). Tra i suoi lavori si segnala il progetto *Survival of the Weakest: Preserving and Analysing Born-Digital Records to Understand How Small Poetry Publishers Survive in the Global Marketplace*[15], finanziato dall'UK Arts and Humanities Research Council, volto ad analizzare e preservare gli archivi digitali di piccole case editrici di poesia. Fra il 2020 e il 2025, Jaillant ha coordinato quattro progetti internazionali su archivi e intelligenza artificiale (IA) che intersecano esigenze di archivi analogici e *born-digital*, soprattutto nell'ottica di miglioramento della loro accessibilità: LUSTRE[16], EyCon[17], AEOLIAN[18] e AURA Network[19].

In Francia l'ITEM (Institut des textes et manuscrits modernes) ha sviluppato tecniche particolarmente raffinate per la gestione e lo studio dei materiali digitali di Jacques Derrida conservati presso l'IMEC (Institut mémoire de l'édition contemporaine)[20] attraverso il progetto Derrida Hexadecimal, coordinato da Aurèle Crasson e sostenuto dall'EUR Translitterae e dal DIM Sciences du texte et connaissances nouvelles. Il progetto, che riunisce un'équipe multidisciplinare composta da Laurent Alonso, Jean-Gabriel Ganascia, Jean-Louis Lebrave e Jérémy Pedrazzi, rappresenta un caso di studio particolarmente ricco ed esteso dell'applicazione di tecniche di informatica forense a supporto della *critique génétique*, per ricostruire il processo creativo del filosofo[21].

Il panorama internazionale si arricchisce ulteriormente con il progetto Bit Philology[22] dell'Università di Berna, diretto da Elena Spadini e avviato nel marzo 2025 con il finanziamento dal Swiss National Science

[15] https://gtr.ukri.org/projects?ref=AH%2FR00773X%2F1.

[16] https://lustre-network.net/.

[17] https://eycon.hypotheses.org/.

[18] https://www.aeolian-network.net/.

[19] https://www.aura-network.net/.

[20] L'archivio digitale di Derrida conservato presso l'IMEC comprende circa 400 floppy disk, 3 hard drive, 5 Iomega Zip drive, 2 Un Syquest drive, un CD-ROM e tre dei suoi sei computer, coprendo il periodo 1986-2004 della produzione intellettuale del filosofo. Si tratta complessivamente di 20.000 file di testo, per un totale di circa un gigabyte di dati, elaborati prevalentemente tramite l'applicativo MacWrite.

[21] La riflessione sull'applicazione dell'informatica forense ai materiali letterari prende avvio dal contributo pionieristico di Matthew Kirschenbaum, "Mechanisms: New Media and the Forensic Imagination" (2007), che ha posto le basi teoriche e metodologiche per una nozione di digital materiality intesa come insieme di tracce tecniche e materiali costitutive dei documenti nativi digitali. A questa prima sistematizzazione hanno fatto seguito le ricerche di Christopher (Cal) Lee (2012), volte a definire metodologie di preservazione e di accesso forense agli archivi digitali (anche in collaborazione con lo stesso Kirschenbaum, Kam Woods e Alexis Chassanoff (C. A. Lee et al. 2013)) e quelle di Thorsten Ries (2018, 2022, 2025), che hanno contribuito a consolidare il campo nell'ambito delle digital literary forensics. In Italia esperienze pionieristiche sono state condotte da Emmanuela Carbé e da Mariangela Giglio (Carbé 2025; Giglio 2025), che hanno applicato tecniche di indagine forense ai floppy disk del Fondo Franco Fortini, offrendo uno dei primi esempi di filologia forense su materiali d'autore.

[22] https://boris-portal.unibe.ch/entities/project/ce44b9b7-ea0c-4f49-a9d0-41e84e8e2619.

Foundation (SNSF Starting Grant, 2025-2030). Il progetto si propone di sviluppare un *toolkit* metodologico e tecnico per lo studio filologico di fonti letterarie born-digital create prima dell'avvento del cloud computing. L'iniziativa intende adottare un approccio fortemente interdisciplinare che combina *digital humanities*, filologia d'autore, critica genetica, filologia materiale, *digital forensics* e *media studies*, con l'obiettivo di trasformare le fonti *born-digital* in oggetti di studio filologico attraverso metodologie innovative per la descrizione, l'edizione e l'analisi di archivi letterari digitali (Spadini 2025).

Nel contesto italiano, il progetto Pavia Archivi Digitali (PAD), avviato nel 2009 presso l'Università di Pavia, rappresenta la prima iniziativa in Italia dedicata esclusivamente alla preservazione degli archivi nativi digitali di scrittori, giornalisti e intellettuali contemporanei (Paul Gabriele et al. 2017; Weston et al. 2019, 2020)[23]. Il digitale d'autore trova qui la sua prima sistematizzazione metodologica nazionale, muovendo i primi passi nello stesso periodo in cui il tema era agli albori anche in realtà internazionali più strutturate. A partire dal 2023, il progetto è stato sottoposto a operazioni di reingegnerizzazione a cura del Centro Manoscritti, adottando un nuovo workflow strutturato sul modello OAIS (Open Archival Information System)[24] (Giagnolini e Baldini 2025; Giagnolini 2025a). Al centro del rinnovamento si trova il *Digital Preservation Workflow* (DPW), un applicativo sviluppato in Java e basato su librerie *open source*, progettato per integrare le operazioni di acquisizione, conservazione, analisi e consultazione in un solo sistema. Il DPW è stato sviluppato grazie ad un'intensa collaborazione con il tecnico Primo Baldini che ha permesso di migliorare l'efficienza e l'organizzazione delle precedenti procedure, così da ottimizzare la gestione degli attuali undici fondi[25] nativi digitali del Centro e aprire la prospettiva a nuove acquisizioni (Giagnolini e Baldini 2025; Giagnolini 2025a, 85–92).

[23] Nel corso degli anni, il progetto è stato seguito con continuità dall'informatico Primo Baldini, affiancato da un comitato gestionale e da un comitato scientifico coordinato, dal 2014, da Paul Gabriele Weston, con i contributi di collaboratori quali Grazia Bruttocao, Emmanuela Carbé, Annalisa Doneda, Nicoletta Leone e Laura Pusterla. Dal 2021 PAD è confluito tra le attività e il patrimonio del Centro per gli studi sulla tradizione manoscritta di autori moderni e contemporanei (noto anche come Centro Manoscritti), presieduto da Giuseppe Antonelli e ha visto la collaborazione di chi scrive e Adele Gorini, affiancate da Primo Baldini per lo sviluppo software.

[24] http://www.oais.info.

[25] Si tratta dei fondi di Silvia Avallone, Franco Buffoni, Gianrico Carofiglio, Mario Desiati, Paolo Di Paolo, Jolanda Insana, Valerio Magrelli, Francesco Pecoraro, Laura Pugno, Beppe Severgnini, Sandro Veronesi. Le acquisizioni condotte nell'ambito del progetto PAD sono state realizzate attraverso la collaborazione diretta con gli autori. Questa scelta metodologica ha risposto alla duplice esigenza di avviare una sperimentazione sistematica sul digitale d'autore e di superare le criticità legate all'accesso forense ai materiali digitali protetti. Questo approccio comporta, tuttavia, diverse limitazioni epistemologiche: i fondi acquisiti si configurano necessariamente come raccolte parziali, rappresentando istantanee della produzione autoriale in specifici momenti cronologici, senza possibilità di documentare l'evoluzione diacronica del processo creativo. La parzialità è inoltre determinata dalla mediazione selettiva degli autori, i quali influenzano la composizione archivistica attraverso criteri

Al centro di progetti di curatela attivi in Italia troviamo anche il Centro Interdipartimentale di Ricerca Franco Fortini, che nel 2022 ha avviato una campagna di recupero e di messa in sicurezza dei supporti digitali riconducibili all'archivio dell'autore attraverso la realizzazione di copie forensi[26]. Il progetto si inserisce nell'iniziativa SOS-DH. Strumenti operativi e scientifici nell'ambito delle Digital Humanities, finanziata dall'Università di Siena nel quadro del programma F-LAB 2022 e coordinata da Emmanuela Carbé.

Tra le attività condotte presso il Centro Franco Fortini, si cita la ricerca di Mariangela Giglio sul fondo Fortini che ha condotto, tra gli altri esiti, alla realizzazione di una virtualizzazione integrale dell'ultimo computer appartenuto all'autore, ottenuta a partire da un insieme selezionato di copie forensi caratterizzate dal minor grado di manipolazione successiva. Nell'ambito dello stesso progetto sono stati inoltre sviluppati strumenti computazionali specifici, tra cui un software per il clustering di supporti basato su *hashing* (Giglio 2025) e eFFeQuadro, un applicativo di prossima pubblicazione dedicato all'analisi forense di supporti Macintosh obsoleti.

Carbé ha rintracciato ulteriori enti conservatori di digitale d'autore sul territorio italiano[27], ma in un quadro ancora molto frammentato, poco documentato e senza procedure di gestione e conservazione strutturate (Carbé 2023, 39-42). È molto probabile, infatti, che esista un ampio sommerso di fondi nativi digitali nelle biblioteche e negli archivi italiani: floppy disk, CD e pennette USB potrebbero essere passati inosservati, "nascosti" tra i faldoni degli archivi analogici per il momento ignorati a causa della carenza di risorse o della scarsa sensibilità verso il tema del nativo digitale. In quest'ottica, il progetto *ALDiNa:*

soggettivi di inclusione ed esclusione dei materiali. Per approfondimenti sulle scelte metodologiche e informazioni più dettagliate sui fondi si vedano (Carbé 2023, 51-112; Giagnolini e Baldini 2025).

[26] Si tratta di 54 floppy disk di Franco Fortini, conservati presso la Biblioteca di Area Umanistica dell'Università di Siena (Carbé 2025). Nell'ambito del progetto, il servizio di salvataggio e migrazione dei dati da floppy disk del Fondo Franco Fortini è stato affidato, in outsourcing, alla società specializzata AudioInnova, attiva nel settore della rimediazione e della conservazione di supporti digitali obsoleti (Carbé 2025, 203).

[27] Carbé riporta: un computer di Imre Toth, conservato dalla stessa Biblioteca Umanistica di Siena; floppy disk e CD nei fondi di Francesca Sanvitale, Mario Luzi, Cristina Campo, Enzo Siciliano e Claudio Magris presso il Gabinetto Vieusseux; l'archivio di Nanni Balestrini, che raccoglie floppy disk e altri materiali nativi digitali; email e file di Luciano Marrocu presso il Fondo Autografi Scrittori Sardi (FASS) dell'Università di Sassari; materiali nativi digitali di autori come Flavio Santi e Nicola Lagioia presso il Centro Interdipartimentale di Ricerca su Tradizione e Traduzione (CIRTT) dell'Università di Cassino. Si segnala anche la recente acquisizione di due computer di Silvia Giacomoni al Centro Apice. La notizia, datata 4 ottobre 2024, si apprende dal sito web del Centro Apice https://www.apice.unimi.it/attivita/larchivio-silvia-giacomoni-apice; ulteriori informazioni sono state fornite da Roberta Cesana durante il convegno "Il Futuro della Memoria. Dove, Come, Cosa Salvare", svoltosi a Milano, presso la Fondazione Arnoldo e Alberto Mondadori, il 5 novembre 2024 (Cesana e Desideri 2025). La registrazione del convegno è disponibile su YouTube a link: https://www.youtube.com/live/ptvPwlMPNzw?si=0qGi35uOGA7QCpIy.

Archivi letterari digitali nativi[28], ha tra i suoi obiettivi principali l'avvio di una mappatura dei materiali nativi digitali d'autore sul territorio italiano. Questo processo inizierà con l'invio, nei prossimi mesi, di un questionario a biblioteche, archivi, musei e università, al fine di raccogliere dati che saranno successivamente integrati attraverso consultazioni e verifiche dirette (Allegrezza et al. 2025). In parallelo, l'iniziativa *Share. Pratiche di cultura al digitale*[29], promossa nell'ambito di Dicolab[30], il sistema formativo per la trasformazione digitale del patrimonio culturale del Ministero della Cultura (PNRR Cultura 4.0), sta conducendo una mappatura più ampia delle esperienze di digitalizzazione e delle iniziative digitali nel settore culturale italiano. Attraverso una survey online sempre aperta, *Share* raccoglie dati su collezioni digitalizzate, progetti di accessibilità e innovazione gestionale, alimentando una mappa in continuo aggiornamento del patrimonio culturale italiano nella sua dimensione digitale. I dati raccolti confluiranno nella Mappa del Patrimonio coordinata dall'Istituto centrale per la digitalizzazione del patrimonio culturale - Digital Library, favorendo un approccio integrato alla conoscenza e valorizzazione del panorama digitale italiano.

La mappatura fin qui proposta, pur essendo parziale, mostra chiaramente una netta predominanza di progetti europei e americani nella letteratura scientifica internazionale, evidenziando la mancanza o la sottorappresentazione di iniziative sviluppate in altri contesti geografici. È stato tuttavia possibile rintracciare alcune iniziative in Brasile, dove il Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC)[31] dal 2019 conserva il fondo ibrido della deputata federale Luiza Erundina de Sousa, ma non è stato possibile recuperare informazioni sulle metodologie messe in campo per la gestione del nativo digitale (Barrero Junior 2024). Un contributo metodologico originale proviene dall'Observatório da literatura digital brasileira (Ctrl+S)[32], che con Atlas da Literatura Digital Brasileira[33] affronta la preservazione di opere di letteratura elettronica attraverso un approccio di documentazione contestuale (metadatazione specialistica, registrazioni video, testimonianze autoriali, *webarchiving*) piuttosto che la preservazione integrale dell'oggetto digitale, sviluppando tassonomie specifiche per poetiche e tecnologie native digitali (Athayde e Rocha 2022).

Nei contesti asiatici emerge un orientamento che sembra tendere maggiormente verso il *Personal Digital Archiving* (PDA) e il *Personal Information Management* (PIM), ossia sulla ricerca delle modalità in cui

[28] https://aiucd.github.io/aldina/.
[29] https://www.fondazionescuolapatrimonio.it/share-pratiche-cultura-digitale/.
[30] http://dicolab.it/.
[31] https://cpdoc.fgv.br/.
[32] https://www.observatorioldigital.ufscar.br/.
[33] https://www.observatorioldigital.ufscar.br/#atlas.

le persone gestiscono o tengono traccia dei propri file digitali. Si tratta principalmente di studi statistici e *literature reviews* in cui si indagano le pratiche di gestione documentale personale, ma senza uno specifico taglio archivistico o letterario (Park e Oh 2018; Yasmeen et al. 2019; Zhao et al. 2019; Ali e Warraich 2020, 2021, 2022, 2023; Alon e Nachmias 2020; Minarso et al. 2023; Parmar 2025).

1.3 Il problema della descrizione

Nello specifico contesto della descrizione archivistica, Jean Dryden (1995) è tra le prime a segnalare l'assenza di strumenti adeguati alla descrizione di archivi elettronici, osservando come le prassi descrittive del tempo[34] non fossero in grado di rendere conto della varietà dei file digitali e dei loro contesti di creazione. Pochi anni dopo, Adrian Cunningham (1999) osserva come il settore degli archivi digitali di persona, in particolare, avesse fatto progressi molto lenti rispetto a quello degli archivi istituzionali.

Nel 1994 l'International Council on Archives (ICA) pubblica la prima edizione del General International Standard Archival Description ISAD(G))[35], ideato per essere un *framework* generale e flessibile per la descrizione archivistica, teoricamente applicabile indipendentemente dalla natura o dall'estensione degli archivi descritti o dal livello di descrizione. Tuttavia, sebbene ISAD(G) sia formalmente agnostico, è stato elaborato in un'epoca in cui gli archivi erano ancora prevalentemente analogici e sulla base di una comprensione relativamente limitata delle esigenze informative dei materiali nativi digitali (Bunn 2021, 2).

Un primo salto qualitativo si ha nel 2007 con il progetto PARADIGM: sebbene l'obiettivo principale fosse sviluppare una metodologia per la preservazione a lungo termine alle collezioni di materiali personali digitali, nell'output del progetto, *PARADIGM Workbook on Digital Private Papers,* un capitolo è dedicato anche alla descrizione del nativo digitale, in cui si propone l'integrazione di standard archivistici, come ISAD(G) e l'Encoded Archival Description (EAD)[36], con standard definiti "tecnici", quali Metadata Encoding and Transmission Standard (METS) [37] e PREservation Metadata:

[34] Gli strumenti analizzati da Dryden comprendevano RAD (Rules for Archival Description), MARC(AMC), AACR2 (Machine-readable data files) e APPM (Archives, Personal Papers and Manuscripts).

[35] https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition/.

[36] https://www.loc.gov/ead/.

[37] https://www.loc.gov/standards/mets/.

Implementation Strategies (PREMIS)[38], aprendo la strada a un approccio ibrido (Bodleian Libraries and Gardens e John Rylands Library 2007).

Il progetto *Born Digital Collections: An Inter-Institutional Model for Stewardship* (AIMS) (2012) consolida l'idea che i principi fondamentali della descrizione, concretizzati nel conservare il contesto e garantire l'accesso, restassero validi anche nel digitale, ma dovessero essere ricalibrati alla luce della fluidità del medium, in particolare nella prospettiva dell'ordinamento (Langdon 2016).

Parallelamente, la Society of American Archivists (SAA) contribuisce al dibattito con un volume collettaneo della collana *Trends in Archives Practice*[39] dedicato all'ordinamento e alla descrizione che includeva riferimenti specifici ai materiali nativi digitali (Prom e Frusciano 2013). Nel suo contributo, J. Gordon Daines (2013) sostiene l'applicabilità delle prassi consolidate ai materiali digitali, ma riconosce al contempo che standard come ISAD(G) e DACS[40], pur essendo *format-agnostic* per definizione, non sono in grado di catturare tutti i metadati necessari per gestire i materiali digitali, raccomandando l'uso di "companion standards" per gestire le informazioni di carattere più tecnico. Nello stesso volume, il saggio di Sibyl Schaefer e Janet Bunde (2013) sottolinea la distinzione concettuale destinata a influenzare la letteratura successiva, ossia quella tra elementi descrittivi tradizionali e metadati gestionali, e l'importanza della loro coesistenza. Su questa linea, Kat Timms (2013) sperimenta la descrizione di materiali nativi digitali con lo standard RAD[41] e ne sottolinea le limitazioni, mentre Cyndi Shein propone di combinare descrizioni EAD a livello di serie con metadati conformi a Dublin Core (DC)[42] e METS per descrivere i contenuti dei file e per documentare le relazioni gerarchiche e di versionamento. In sintesi, tutti i contributi riflettono una tendenza condivisa: integrare gli standard tradizionali con metadati e strumenti specifici per i materiali digitali, combinando informazioni descrittive e gestionali per rispondere alle esigenze del contesto digitale.

Nonostante la convergenza verso questo approccio, ancora nel 2016 John Langdon segnala la persistente assenza di linee guida condivise e strutturate. La letteratura, pur avendo individuato criticità e proposto adattamenti delle pratiche esistenti, non offriva infatti raccomandazioni concrete, lasciando spesso ai

[38] https://www.loc.gov/standards/premis/.

[39] https://www2.archivists.org/publications/book-publishing/trends-in-archives-practice.

[40] Describing Archives: A Content Standard (DACS): si tratta dell'implementazione statunitense degli standard internazionali ISAD(G) e ISAAR(CPF) per la descrizione dei materiali d'archivio e dei loro soggetti produttori. https://www2.archivists.org/groups/technical-subcommittee-on-describing-archives-a-content-standard-dacs/describing-archives-a-content-standard-dacs-second-.

[41] Rules for Archival Description (RAD), standard canadese per la descrizione archivistica, con ultimo aggiornamento risalente al 2008 (Bureau of Canadian Archivists, Planning Committee on Descriptive Standards 2008).

[42] https://www.dublincore.org/specifications/dublin-core/dces/.

singoli professionisti l'onere di sperimentare soluzioni, di cui riporta alcuni esempi (Langdon 2016, 43). Un primo nodo evidenziato da Langdon riguarda l'ordinamento: nel 2006 Catherine Stollar Peters articola la tensione tra ordinamento archivistico tradizionale e le possibilità offerte dai metadati a livello di singolo elemento per gli oggetti digitali (Stollar Peters 2006). Analizzando le relazioni tra materiali analogici e digitali nell'archivio di Michael Joyce, Peters individua una sincronicità del processo creativo nell'utilizzo dei due supporti, che la porta a sviluppare una strategia di descrizione che tratta entrambi in un unico sistema (DSpace[43]), utilizzando di un set di metadati di base. Tuttavia, alcuni metadati fondamentali, come i percorsi originali, non potevano essere registrati attraverso i campi di descrizione disponibili e sono stati documentati separatamente in fogli di calcolo resi disponibili come documentazione di progetto (Stollar Peters 2006, 29).

Una scelta analoga è stata compiuta da Simon Wilson, che nel trattare i documenti di Stephen Gallagher ha adottato un primo livello di ordinamento basato sulla tipologia dell'opera (romanzo, racconto, sceneggiatura, ecc.), indipendentemente dal formato dei materiali (Wilson 2012). Anche per Laura Carroll e i suoi colleghi il lavoro sui documenti di Salman Rushdie si è basato su «a commitment to approach the material as holistically as possible, to prioritize the integration of paper and digital, and to balance the needs of donors with those of researchers» (Carroll et al. 2011, 61). Un esempio di approccio alternativo, invece, è quello degli archivisti della Stanford University, che hanno scelto di organizzare i documenti digitali di Robert Creeley e Stephen Jay Gould in serie separate e di descriverli solo a livello di serie (AIMS 2012, 88-89).

Il saggio di Langdon si concludeva provocatoriamente con la stessa diagnosi formulata dalla Dryden venti anni prima: la descrizione archivistica del digitale, e degli archivi di persona in particolare, versava ancora in una «kind of paralysis», «with everyone waiting for someone else to take the initiative» (Dryden 1995, 106; Langdon 2016, 49).

In anni più recenti, diverse istituzioni universitarie statunitensi hanno sviluppato iniziative coordinate per standardizzare l'uso di DACS in ArchivesSpace[44] per le collezioni native digitali. Tra le esperienze più importanti si annoverano quella della Born-Digital Content Description Task Force di Harvard (2023) e, in particolare, quella della University of California (2017), a cui fanno riferimento le linee guida sviluppate da altre università, quali la Yale University (2020) e la University at Buffalo (s.d.). Queste iniziative hanno puntato a standardizzare un set di metadati fondamentali e a garantire un aggiornamento

[43] https://dspace.org/.

[44] https://archivesspace.org/.

periodico delle pratiche descrittive, concentrandosi prioritariamente sulla descrizione a livello di collezione piuttosto che su livelli di dettaglio.

In Italia, in linea con le tendenze internazionali e con l'urgenza percepita come prioritaria, il dibattito sul *born-digital* si è concentrato prevalentemente sul tema della preservazione. L'Archivio Massimo Vannucci costituisce il primo caso di descrizione archivistica di un archivio personale nativo digitale italiano, affrontata secondo lo standard ISAD(G), con particolare attenzione alla corrispondenza elettronica del deputato (Allegrezza e Gorgolini 2016). L'inventario, elaborato sulla piattaforma Memorie di Marca[45] (basata sul software AtoM[46]) non risulta, tuttavia, attualmente consultabile.

Più recente è l'esperienza di *Pavia Archivi Digitali* (PAD), che nel 2025 ha avviato un progetto di descrizione dei fondi applicando ISAD(G) agli archivi di Silvia Avallone e Paolo di Paolo tramite la piattaforma Archimista Web[47]. Tale approccio, motivato dall'esigenza di fornire un primo quadro descrittivo con gli strumenti già a disposizione, ha confermato l'utilità di ISAD(G) come prima chiave di accesso, ma al contempo ha ribadito la limitata granularità degli schemi tradizionali di fronte alla complessità dei materiali (Gorini e Giagnolini 2025).

Tanto per l'Archivio Vannucci quanto per i fondi Avallone e Di Paolo, la descrizione si è spinta, per la maggior parte delle serie, fino al livello del documento, grazie all'estrazione automatica della struttura delle cartelle e alla sua importazione in formato XML all'interno delle rispettive piattaforme. Ciononostante, in entrambi i casi la curatela manuale si è rivelata particolarmente onerosa (Giagnolini 2018; Gorini e Giagnolini 2025).

In questo scenario, la riflessione archivistica è chiamata a interrogarsi sulla possibilità di potenziare le forme di automazione per ottimizzare i processi descrittivi. Negli anni Novanta, David A. Wallace (1995) aveva già immaginato sistemi in grado di catturare automaticamente i metadati lungo tutto il ciclo di vita dei documenti, riducendo la necessità di un intervento postumo. Infatti, la natura stessa del digitale rende oggi questa prospettiva concretamente praticabile: metadati e informazioni contestuali, come approfondito nei capitoli 4 e 7, possono infatti essere estratte direttamente dai documenti. Numerosi tool sono stati sviluppati a tale scopo, e un elenco è consultabile nel *Community Owned Digital Preservation Tool Registry* (COPTR)[48].

[45] https://new.memoriedimarca.it/.

[46] https://www.accesstomemory.org/it/.

[47] https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioRedazionale/servizi-e-informazioni/Enti-e-Operatori/cultura/Biblioteche-ed-archivi/archimista/archimista.

[48] https://coptr.digipres.org/Main_Page.

Un esempio concreto di avanzamento verso processi automatizzati è fornito dall'esperienza della British Library, che nel 2017 imposta un workflow integrando nelle descrizioni dei fondi nativi digitali di persona anche le informazioni provenienti dall'estrazione dei metadati effettuata tramite DROID[49] (Pledge e Dickens 2018). DROID è uno strumento software *open source* sviluppato da The National Archives per eseguire l'identificazione automatica in *batch* dei formati di file, al fine di individuare i migliori criteri di gestione in termini di preservazione[50]. Nel workflow impostato alla British Library, i valori estratti da DROID, come denominazione, formato, codice hash, dimensione, e data di ultima modifica, venivano selezionati e copiati in un file XLSX per il caricamento nel sistema archivistico della biblioteca (Integrated Archives and Manuscripts System (IAMS)). Prima del caricamento, i metadati venivano organizzati manualmente in un file XLSX, seguendo i campi descrittivi obbligatori secondo ISAD(G) e lo stesso sistema IAMS; successivamente, venivano raggruppati per supporto d'origine, associati a un identificativo univoco e infine pubblicati nel catalogo per l'accesso degli utenti (Explore Archives and Manuscripts Catalogue) (Pledge e Dickens 2018, 64-65). Questo approccio, sebbene adeguato per collezioni di dimensioni ridotte – come quella di Wendy Cope[51], il caso di studio su cui si era basato il suo sviluppo – si rivela ben presto troppo dispendioso in termini di tempo e risorse per essere applicato a raccolte più estese, rendendo necessario lo sviluppo di soluzioni in grado di automatizzare almeno parte delle operazioni (McKean 2025, 108-9). Dal 2018 vengono quindi avviati esperimenti di automazione, inizialmente con Google OpenRefine[52] e successivamente con Python, che portano allo sviluppo dell'Automated Metadata Solution (AMS). L'obiettivo raggiunto da AMS è stato ridurre tempi ed errori nella fase più onerosa del workflow del 2017, ossia la conversione dei metadati prodotti da DROID in un formato compatibile con il sistema archivistico della biblioteca (IAMS). Il sistema AMS genera automaticamente template di catalogazione in formato CSV, che vengono poi integrati manualmente dall'archivista con descrizioni testuali di contenuto secondo lo standard ISAD(G), con controlli sulla sensibilità dei dati e con eventuali segnalazioni di criticità tecniche ai curatori.

[49] https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

[50] DROID utilizza delle firme specifiche per identificare e segnalare il formato e la versione dei file digitali. Queste firme sono memorizzate in un file XML generato dalle informazioni registrate nel registro tecnico PRONOM, una risorsa online gestita da The National Archives per reperire informazioni certe sui formati di file, sui prodotti software e su altri componenti tecnici necessari per garantire l'accesso a lungo termine ai documenti elettronici e ad altri oggetti digitali di valore culturale, storico o commerciale. Per ulteriori informazioni: https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

[51] Si tratta di 57 floppy disk IBM PC da 3,5 pollici (75,0 MB), 19 floppy disk Amstrad da 3,0 pollici (14,3 MB) e una chiavetta USB da 16 GB, contenente cartelle di posta elettronica di Microsoft (MS) Outlook in formato DBX, insieme ad ulteriori documenti MS Word (McKean 2025, 108) .

[52] https://openrefine.org/.

Secondo McKean, pur garantendo un modello di accesso funzionale, l'approccio (tuttora in uso, con progressivi perfezionamenti) presenta ancora alcuni limiti. In particolare, l'automazione implementata tende a semplificare eccessivamente i dati, trattando i file come unità isolate e trascurando le relazioni e le dipendenze proprie degli ambienti digitali, con il risultato di produrre una descrizione appiattita della realtà. Inoltre, le informazioni contestuali spesso veicolate da etichette fisiche o dai supporti non trovano al momento integrazione nei processi automatizzati (McKean 2025, 117).

In opposizione a questa dinamica di appiattimento e nella prospettiva di valorizzare i contesti digitali, Bunn individua nei grafi della conoscenza (*knowledge graphs*) una possibile soluzione tecnologica per superare i limiti degli attuali strumenti descrittivi (Bunn 2021, 2). Nei *knowledge graph* i dati assumono la forma di Linked Data, e, se pubblicati liberamente, di Linked Open Data (LOD): una modalità di diffusione di dati strutturati, leggibili e processabili dalle macchine, che includono collegamenti tra concetti rappresentativi di entità reali, identificati in modo univoco e persistente grazie agli strumenti del Web Semantico (Bizer et al. 2011). Tali dati, oltre a essere processabili automaticamente, sono anche accessibili all'uomo tramite linguaggi di interrogazione come SPARQL, che consentono accesso diretto all'informazione.

In quest'ottica, Langdon sottolinea come sia necessario individuare cosa rappresentare, ma anche considerare i nuovi usi che i ricercatori fanno oggi delle informazioni (2016, 46-47). Grazie agli strumenti digitali, attività di ricerca precedentemente impraticabili con archivi tradizionali diventano possibili, permettendo di combinare, analizzare e interrogare i dati con nuove metodologie. Questa condizione deve necessariamente avere una ricaduta sulla descrizione archivistica, che deve evolvere per sostenere la scoperta del materiale e, al contempo, il potenziale utilizzo dei suoi dati descrittivi (Langdon 2016, 47). In questo senso, presentare le descrizioni come LOD significa offrire modalità di interrogazione libere e non vincolate alle strutture predefinite dei sistemi informativi tradizionali, ampliando le possibilità di analisi e di valorizzazione della documentazione digitale.

Alla luce di quanto emerso, questa ricerca individua nella convergenza di quattro elementi chiave un approccio integrato per affrontare le sfide della descrizione dei materiali nativi digitali. In primo luogo, è fondamentale definire chiaramente le esigenze di rappresentazione, tenendo conto della complessità dei contesti digitali e delle modalità di utilizzo da parte dei ricercatori. In secondo luogo, la modellazione tramite LOD permette di garantire coerenza, interoperabilità e collegamenti tra le informazioni, superando i limiti dei sistemi tradizionali. Il terzo elemento riguarda l'automazione, che consente l'estrazione sistematica delle strutture e dei metadati nativi, riducendo tempi ed errori e liberando risorse per interventi a più alto valore aggiunto. Infine, è necessario definire modalità efficaci per la restituzione

e la fruizione dei dati descrittivi, in modo da valorizzare pienamente i materiali digitali e rendere le informazioni facilmente interrogabili, integrabili e utilizzabili in contesti di ricerca avanzata.

2. Il soggetto produttore nel contesto digitale: cinquanta interviste d'autore

La descrizione dell'archivio passa necessariamente per un'analisi del profilo dell'intellettuale e del suo rapporto con il digitale. Ben prima di costituire una sfida per le istituzioni culturali, il digitale rappresenta un'urgenza gestionale per gli stessi autori, la cui attività creativa e professionale dipende anche dalla capacità di mantenere accessibili nel tempo i propri materiali di lavoro. Dunque, la figura dell'autore come soggetto produttore d'archivio assume nella contemporaneità una dimensione più complessa e attiva: più che nel contesto analogico, lo scrittore diviene il primo responsabile della gestione, della selezione e della sopravvivenza del proprio patrimonio digitale. Dopo aver delineato il quadro teorico e lo stato dell'arte (cap. 1), per rispondere alla domanda su come si configura un archivio d'autore contemporaneo e quali pratiche adottino gli autori (RQ1), è necessario assumere il punto di vista dei soggetti produttori. Attraverso l'analisi di cinquanta interviste a finalisti dei premi Strega e Campiello (1985-2024), questo capitolo fornisce un quadro di come autrici e autori si rapportino alla gestione del proprio archivio, con particolare attenzione al passaggio dalla scrittura su supporti analogici a quella digitale e, più in generale, ai mutamenti nei processi di produzione, raccolta e conservazione dei materiali di lavoro[53].

2.1 Introduzione

Diversi sondaggi nel contesto internazionale hanno approfondito le pratiche di *Personal Information Management* (PIM) e *Personal Digital Archiving* (PDA) nell'ambito della produzione letteraria contemporanea. Becker e Nogues, attraverso un questionario rivolto a 118 scrittori di diversa esperienza (prevalentemente poeti), rilevano una proliferazione disorganizzata di file digitali in più luoghi di conservazione, con un elevato rischio di perdita dovuto all'inadeguatezza delle strategie archivistiche adottate (2012, 509). Lo studio di Micunovic, Marčetić e Krtalić, basato su interviste semi-strutturate a nove membri della Croatian Writers Society e della Croatian Writers' Association, sottolinea il ruolo centrale della tecnologia nel processo creativo. Sebbene il campione limitato non abbia permesso di trarre conclusioni generalizzabili, è emerso come gli scrittori abbiano assunto, più o meno consapevolmente, l'onere di organizzare e conservare il materiale che creano e raccolgono in ambiente digitale (2016, 12).

[53] Gli esiti della ricerca descritta in questo capitolo sono presentati nell'articolo Giagnolini L., *Riflessi digitali: indagine sugli archivi d'autore contemporanei*, «Annali d'Italianistica», 43 (2025).

La principale conclusione dello studio, tuttavia, è la scarsa consapevolezza degli autori contemporanei riguardo al valore del patrimonio digitale e alla necessità di preservare il contesto delle loro opere (2016, 14). La ricerca di Krtalić e Dinneen, condotta su diciotto noti scrittori e artisti neozelandesi di diversi ambiti creativi, presenta risultati in linea con le ricerche precedenti in merito ai fattori che influenzano la gestione delle informazioni personali, come le emozioni legate all'organizzazione e alla conservazione dei documenti e la valutazione del valore del proprio archivio (2024, 200). Su quest'ultimo aspetto, evidenziano la complessa interazione tra valore percepito e strategie di archiviazione adottate (2024, 201) e il conflitto di opinioni su ciò che deve essere conservato e presentato alle generazioni future (Krtalić e Dinneen 2024, 199, 201).

Nonostante le differenze metodologiche e di campione, questi studi convergono sull'assenza di approcci sistematici alla gestione degli archivi personali e sulla necessità di una maggiore standardizzazione e consapevolezza delle pratiche di conservazione (Becker e Nogues 2012, 503; Krtalić e Dinneen 2024, 199), in linea con le indagini PIM in contesti più ampi (Sinn et al. 2017). Viene evidenziata, a livello internazionale, l'urgenza di sviluppare strumenti e servizi mirati, promuovendo una collaborazione più efficace tra produttori d'archivio e istituzioni culturali, al fine di supportare la preservazione e l'accessibilità del patrimonio digitale nel lungo periodo (Krtalić e Dinneen 2024, 200; Becker e Nogues 2012, 509-10).

Nel panorama italiano, l'inizio di questo filone di ricerca nel contesto letterario può essere fissato al 1983 con la pubblicazione per "Tuttolibri" della celebre indagine *Scusi, lei lo scriverebbe un romanzo con il computer?* di Luciano Curino. L'autore riportava i pareri di Alberto Moravia, Mario Soldati, Piero Chiara, Italo Calvino, Andrea Zanzotto, Masolino D'Amico e Maria Corti in merito alla possibilità di usare il computer per scrivere le proprie opere (Curino 1983). Dall'inchiesta emerse un giudizio sostanzialmente negativo e di grande diffidenza (con l'eccezione di D'Amico). Carmen Ragusa, che recupera l'indagine di Curino, ha tratteggiato una storia della videoscrittura in Italia in cui si rintracciano ulteriori e significative testimonianze sulla percezione del passaggio dalla macchina da scrivere al computer negli anni Ottanta, tra iniziale scetticismo e progressiva sperimentazione (2021, 93-143). A partire dagli anni Novanta, gli studi di Domenico Fiormonte sul rapporto tra scrittori e *word processor* (1996; 1997) offrono una ricca analisi del panorama di fine secolo, basata su interviste dirette e fonti edite, che permette di tracciare un quadro delle modalità con cui gli autori hanno gradualmente accolto il digitale nella scrittura e nella revisione testuale. Recentemente, Carbé ha ripreso questi studi confrontandoli con testimonianze italiane ed estere, mettendo in luce come la diffidenza iniziale abbia lasciato spazio a una piena integrazione degli oggetti tecnologici nei processi creativi, oggi elementi

indispensabili sui loro tavoli di lavoro (Carbé 2023). Oggetti con cui, secondo Carbé, «dobbiamo fare i conti per comprendere dove si svolge il mestiere della scrittura degli ultimi decenni, un mestiere che dalla pagina bianca, demone e promessa di ogni scrittore, si trasferisce su un altrettanto temibile e promettente file vuoto» (Carbé 2023).

Per quanto riguarda la contemporaneità italiana, le pubblicazioni e progetti più recenti offrono frammenti di testimonianze, spesso inserite in progetti dal focus più ampio. Ad esempio, il progetto *A Carte Scoperte*, curato da Paola Italia e dal Master in Editoria cartacea e digitale dell'Università di Bologna, raccoglie una serie di interviste ad autori collaboratori del Master, analizzando i loro tempi, metodi, spazi e strumenti di scrittura (Italia 2021). In modo analogo, nel volume *Dove si scrive, come si scrive* (Sanzogni 2023), ventotto scrittori hanno aperto le porte dei loro luoghi di scrittura, intesi sia come spazi fisici che come rifugi mentali, descrivendo come lo spazio influenzi il loro processo creativo. Una riflessione sul futuro degli archivi letterari è emersa anche dal convegno *Il futuro della memoria. Dove, come, cosa salvare* (Fondazione Arnoldo e Alberto Mondadori, 5 novembre 2024), dove quattro scrittori italiani (Marco Balzano, Donatella Di Pietrantonio, Antonio Franchini e Helena Janeczek) si sono confrontati su cosa, della loro opera, trasmettere alle generazioni future e cosa disperdere. Il dibattito, arricchito dagli interventi video di otto autori internazionali tra cui Jennifer Egan, Geoff Dyer e Fernando Aramburu, ha esplorato i confini del testo letterario nell'epoca digitale estendendo la riflessione anche oltre il contesto italiano (Giagnolini e Giglio 2024).

Preziose informazioni si possono chiaramente rintracciare nell'ambito degli stessi fondi nativi digitali d'autore, ma i casi di studio italiani, come evidenziato nel capitolo 1.2, sono ancora troppo pochi per delineare *pattern* e pratiche condivise. Inoltre, questi archivi, così come molte testimonianze raccolte nel tempo, appartengono già a finestre temporali digitalmente distanti, impedendo un quadro esaustivo del panorama di produzione documentaria odierno. Per documentare le pratiche di produzione e gestione dei materiali digitali e fornire strumenti utili a chi, nei prossimi decenni, sarà chiamato a recuperarli, conservarli e studiarli, nell'ambito di questa ricerca è stato elaborato un sondaggio focalizzato sulla prospettiva dei soggetti produttori nella contemporaneità letteraria italiana, analizzata attraverso cinquanta interviste condotte con finalisti dei premi Strega e Campiello degli ultimi trentacinque anni.

La ricerca, strutturata attorno a sedici domande, mira a identificare le pratiche gestionali adottate dagli scrittori, dall'utilizzo di specifici dispositivi e software alla gestione delle stesure, dai sistemi di backup alla loro "volontà d'archivio", una *nuance* della volontà d'autore, intesa come desiderio di conservare i propri materiali anche per i posteri (Italia e Zanardo 2023, 14-16). Analizzando al contempo il grado di consapevolezza dei soggetti produttori rispetto al digitale e l'influenza della tecnologia sul loro modus

operandi, l'obiettivo è contribuire alla comprensione delle nuove dinamiche di produzione e conservazione degli archivi d'autore attraverso le testimonianze dirette degli autori stessi.

2.2 Metodologia

Stante l'obiettivo di indagare il comportamento degli autori come soggetti produttori d'archivio nell'era digitale, sono state individuate quattro principali aree di indagine:

1. Contesti di produzione documentale: uso di dispositivi, uso del cartaceo, trasferimento delle informazioni tra supporti cartacei e digitali e tra dispositivi digitali diversi.
2. Modalità di gestione e archiviazione dei materiali: rapporto fra analogico e digitale; criteri di denominazione dei file e organizzazione delle cartelle, strategie di *versioning* per il monitoraggio delle stesure successive.
3. Strategie di conservazione e trasmissione: valutazione della consapevolezza autoriale rispetto alla necessità di garantire la sopravvivenza del proprio archivio, sia nel breve termine (backup e sicurezza dei dati) che nella prospettiva di una trasmissione ai posteri.
4. La presenza e l'influenza di nuove forme d'archivio, nuovi spazi di scrittura e nuovi strumenti (social media, siti web, blog, utilizzo del web e di software particolari).

Per raccogliere dati utili alla comprensione di questi aspetti, è stata elaborata un'intervista articolata nelle seguenti domande:

1. Di quale genere si occupa?
2. Scrive solo in digitale o utilizza ancora carta e penna? Se sì, per qualche fase di scrittura?
3. Quali dispositivi digitali utilizza (appunti e annotazioni incluse)? (pc, cellulare, tablet etc.)
4. Se utilizza più di un dispositivo, come si sviluppa il suo lavoro fra di essi? Come gestisce i file da un dispositivo all'altro? (ad es.: scrive su un cloud collegato ai vari dispositivi per avere il testo sempre aggiornato? Invia via email/messaggistica i file dall'uno all'altro dispositivo?)
5. Con ogni probabilità il suo archivio è ibrido, ossia composto da una parte di documenti cartacei e una parte di documenti digitali. Come dialogano le due partizioni?
6. Stampa i testi che scrive in digitale? Se sì, per cosa utilizza la stampa? (Ad es.: conservazione, correzione bozze, lettura etc.)
7. Fa attenzione ai formati (.docx, .jpg, .pdf, etc.) con cui salva i suoi documenti?
8. Utilizza criteri specifici per la denominazione dei suoi file?
9. Come tiene traccia delle varie versioni di uno scritto?

10. Organizza le sue cartelle secondo un criterio specifico? (ad es., attività, tema, opera, tipologia documentaria etc.)
11. Effettua dei backup? Se sì, come e con quale frequenza?
12. Ha un sito web o un blog personale? Se sì, lo considera parte del suo archivio? Ha mai pensato al *web archiving*?
13. Utilizza i social? Se sì, li considera parte del suo archivio? Ha mai salvato una copia dei suoi dati?
14. Ha mai pensato di conferire una copia del suo archivio digitale ad una istituzione culturale?
15. Senza credenziali (come nome utente e password) in futuro sarà sempre più complicato accedere ad account e supporti. Ha mai pensato di lasciare le sue credenziali ad una persona di fiducia o di redigere un testamento per il digitale?
16. Ci sono ulteriori aspetti del suo processo creativo digitale che ha piacere di condividere? (utilizzo di internet, di particolari software, di email etc.)

Definita l'intervista come metodo di indagine, il passo successivo è stato l'individuazione del campione di scrittrici e scrittori a cui sottoporla. La scelta è ricaduta sui finalisti dei premi Strega e Campiello dell'ultimo trentennio, le cui opere possono essere analizzate come punti di convergenza tra riconoscimento critico e successo di pubblico, investite da una consacrazione che è al contempo autoriale e commerciale (Simonetti 2023). Pur non esenti da differenze, entrambi i premi sono caratterizzati dalla presenza di un doppio filtro: prima specialistico, poi di un pubblico prevalentemente «non specialistico e centrifugo» che «fa delle scelte culturali dello Strega (e per altri versi del Premio Campiello) un banco di prova sociologico relativamente attendibile di un gusto plasmato da una élite di letterari e sottoposto al collaudo di lettori colti» (Simonetti 2023, 14).

Tuttavia, va riconosciuto che tale scelta introduce un bias nei dati. Come evidenziato da Luigi Matt, i premi letterari offrono «una mappa molto lacunosa e perciò di fatto distorta» del campo letterario italiano, privilegiando sistematicamente autori pubblicati dai grandi editori a discapito di «collane commercialmente più deboli, ma in grado di proporre autori interessanti» (Matt 2009, 251-52). Pur essendo consapevoli della limitazione del campione in termini di rappresentatività rispetto alla pluralità di voci e forme di scrittura, la scelta del corpus è stata dettata dalla necessità di analizzare un insieme di autori con caratteristiche raffrontabili, permettendo, al contempo, di ampliare il numero di autori rispetto alle analisi qualitative e quantitative precedenti (Krtalić e Dinneen 2024; Micunovic et al. 2016).

Infatti, ai vincitori del Campiello viene riconosciuto un *fil rouge* che passa per la presenza di tematiche familiari (Gambaro 2004, 108-9) e di un "altrove" storico-geografico nella costruzione e nello svolgimento della trama (Verona 2019, 80). Raffaele Crovi riscontra nei romanzi del Campiello le

costanti di un buon livello di scrittura e una struttura romanzesca forte (Gambaro 2004, 108-9); affinità che Lorenzo Tomasin individua, invece, in un impianto linguistico simile, privo di sperimentazioni, volto a soddisfare il gusto popolare (Verona 2019), e dunque, tendenzialmente, il mercato.

Per i testi premiati dallo Strega, Simonetti riconosce «una solida e piana tenuta narrativa», capace di interessare anche il consumatore occasionale, essendo al tempo stesso anche

> testi artisticamente ambiziosi e, a qualche titolo, "di qualità" (dove la qualità letteraria non è una sostanza oggettivamente misurabile: più che altro un polline comunicativo, una vernice di aspirazione, insomma un abito su misura che un marketing attento può far indossare a qualsiasi oggetto artistico, indipendentemente dal fatto che sia più o meno riuscito) (Simonetti 2023).

Dunque, lungi dal voler individuare un canone, il tentativo è stato quello di individuare un campione di autori che intercettasse, per motivi diversi, «una tendenza o una sensibilità del proprio tempo: il gusto della giuria, degli editori, del pubblico *in quel momento storico*» (Simonetti 2023, 45).

Gli autori sono stati individuati fra i finalisti nell'arco cronologico 1985-2024 per raccogliere testimonianze che riflettano l'evoluzione dell'integrazione del digitale nelle pratiche di scrittura, dalle prime sperimentazioni fino al suo ruolo pienamente consolidato nel processo creativo. I contatti degli autori sono stati reperiti mediante ricerche online, privilegiando indirizzi email e profili social, ai quali è stata inviata una richiesta di partecipazione alla ricerca[54]. Il campione finale è costituito dagli autori che hanno risposto all'appello, per un totale di cinquanta adesioni[55].

Per garantire flessibilità nella modalità di partecipazione, agli autori è stata data la possibilità di rispondere sia oralmente (tramite chiamata, videochiamata o messaggi vocali) sia in forma scritta. Questa scelta ha permesso di raccogliere un numero maggiore di testimonianze, sebbene abbia comportato leggere discrasie nell'ampiezza e nell'articolazione delle risposte, tendenzialmente più dettagliate nelle

[54] Alcuni recapiti, non reperibili online, sono stati cortesemente forniti dalle prof.sse Emmanuela Carbé, Paola Italia e dal prof. Giuseppe Antonelli.

[55] Hanno aderito allo studio: Eraldo Affinati, Ippolita Avalli, Marco Balzano, Mario Biondi, Maria Grazia Calandrone, Giulia Caminito, Andrea Canobbio, Paola Capriolo, Ennio Cavalli, Antonella Cilento, Giuseppe Conte, Cesare De Marchi, Paolo Di Stefano, Sandro Frizziero, Veronica Galletta, Fausta Garavini, Vittorio Giacopini, Tommaso Giartosio, Pietro Grossi, Helena Janeczek, Massimo Lugli, Giuseppe Lupo, Maurizio Maggiani, Valerio Magrelli, Paolo Malaguti, Franco Matteucci, Piero Meldini, Giovanni Montanaro, Giuseppe Montesano, Marta Morazzoni, Giulio Mozzi, Davide Orecchio, Valeria Parrella, Francesco Pecoraro, Enrico Pellegrini, Claudio Piersanti, Tommaso Pincio, Laura Pugno, Daniela Ranieri, Elisabetta Rasy, Raffaella Romagnolo, Alessandra Sarchi, Pietro Spirito, Fabio Stassi, Andrea Tarabbia, Nadia Terranova, Filippo Tuena, Mariolina Venezia, Alessandro Zaccuri, Ade Zeno.

interviste orali. Le risposte sono state raccolte e organizzate sia in formato tabulare (XLSX), per facilitare l'analisi comparativa, sia in un file di testo, per una lettura più scorrevole[56].

2.3 Risultati

L'analisi dei dati si è svolta in tre fasi principali:

1. *Codifica tematica.* Adottando un approccio basato sulla Grounded Theory (Charmaz 2006, 42-71) alle risposte degli autori sono state associate delle parole chiave al fine di renderle confrontabili e idonee a un'analisi quantitativa. La codifica è stata eseguita manualmente per garantire una maggiore sensibilità all'interpretazione del testo. I dati codificati sono stati organizzati in un file tabulare (XLSX), che mantiene il riferimento all'autore intervistato e include una o più colonne per ciascuna domanda, in funzione della natura delle informazioni da estrarre[57]. In alcuni casi, la codifica è stata diretta e immediata, in altri ha comportato una sintesi o un'astrazione del contenuto testuale. Per questo motivo si è ritenuto necessario effettuare anche una ricontestualizzazione finale del dato, recuperando i dettagli e il contesto delle risposte integrali (punto 3).
2. *Analisi quantitativa e visualizzazione.* I dati codificati sono stati importati in un ambiente Python[58] utilizzando la libreria pandas[59] che consente di gestire file XLSX e organizzare le informazioni in strutture dati (*DataFrame*), facilitandone l'elaborazione e l'analisi. I dati sono stati elaborati per fornire visualizzazioni grafiche create tramite le librerie matplotlib[60] e seaborn[61].
3. *Interpretazione dei dati emersi*. L'ultima fase di analisi si è concentrata sull'interpretazione integrata dei dati, con l'obiettivo di individuare pattern, correlazioni tra variabili e tendenze emergenti, contestualizzando la categorizzazione tematica all'interno del quadro complessivo delle risposte. L'analisi ha messo in relazione dati quantitativi e qualitativi, permettendo di cogliere sia pratiche generalizzabili sia le narrazioni soggettive degli autori. Per facilitare

[56] I file che riportano le interviste in forma integrale sono disponibili nel *repository* Zenodo dedicato al progetto (Giagnolini 2025b) disponibile al link: https://zenodo.org/records/15063775.
[57] Il file è disponibile nel *repository* dedicato al progetto (Giagnolini 2025b) consultabile al link: https://zenodo.org/records/15063775.
[58] Il Jupyter Notebook contenente il codice sviluppato per l'analisi e la visualizzazione dei dati è disponibile al seguente link: https://colab.research.google.com/drive/1etqm0vLh1b7n6ffGDdjbjqE6VfDjHP1S?usp=sharing.
[59] https://pandas.pydata.org/.
[60] https://matplotlib.org/.
[61] https://seaborn.pydata.org/.

l'analisi, lo studio delle risposte è stato suddiviso in quattro aree di indagine individuate al capitolo 2.1:

- Contesti di produzione documentale nell'era digitale (analisi domande nn. 1-4).
- Modalità di gestione e archiviazione dei documenti digitali e analogici (analisi domande nn. 5-6, 8-10).
- Strategie di conservazione e trasmissione del patrimonio documentario (analisi domande nn. 7, 11, 14-15).
- Nuove forme d'archivio e scrittura (analisi domande nn. 12-13, 16).

2.3.1 Contesti di produzione documentale nell'era digitale

Sul totale dei cinquanta intervistati, il 98% delle autrici e degli autori dichiara di occuparsi di narrativa[62], dato che chiaramente non sorprende vista la natura del campione. La domanda è stata condotta per semplici finalità di rilevamento statistico, con l'obiettivo di tracciare una panoramica della varietà testuale e documentaria presente negli archivi d'autore ed eventualmente verificare, attraverso l'analisi delle risposte successive, analogie e differenze nelle pratiche di gestione documentale anche in relazione alle diverse tipologie di documento. Questo approccio, lungi dal voler alimentare il dibattito sulla validità epistemologica delle classificazioni per generi, ha permesso di evidenziare che il 68% degli autori intervistati si cimenta anche in altre forme di scrittura. Il 19% si occupa anche di articoli e contributi per quotidiani o periodici; l0,6% di testi per il teatro; il 31% scrive saggi o pamphlet; il 15,5% poesie; il 5,2% traduzioni, il 3,4% testi autobiografici e il 17,2% testi di altra e varia natura (Figura 2.1).

Basandosi esclusivamente sulle dichiarazioni spontanee degli autori, tuttavia, i dati raccolti risentono di una certa disomogeneità nelle risposte: non tutti gli intervistati hanno segnalato con lo stesso grado di completezza le diverse tipologie testuali praticate, e alcuni hanno probabilmente omesso attività secondarie o ritenute meno rilevanti rispetto alla propria identità autoriale primaria. Una rilevazione più accurata avrebbe richiesto la consultazione sistematica della bibliografia completa di ciascun autore, attingendo a cataloghi, database e repertori bibliografici. Ciononostante, ai fini della presente ricerca i dati auto-dichiarati si sono rivelati sufficienti per fornire una mappatura indicativa della varietà documentale presente negli archivi e per orientare l'indagine sulle pratiche conservative.

[62] L'eccezione è Tommaso Giartosio, che si definisce "poeta, autobiografo, saggista, e autore e conduttore radiofonico" (Giagnolini 2025b), finalista del premio Strega nel 2024 con *Autobiogrammatica* (Minimum Fax, 2024).

Interrogati poi sulla natura, analogica o digitale, del loro processo di scrittura, il 22% ha dichiarato di adottare un approccio interamente digitale sin dalle fasi iniziali. La maggioranza, pari al 38%, predilige un processo prevalentemente digitale, ricorrendo a carta e penna solo occasionalmente. Il 26% opta invece per una metodologia ibrida, impiegando strumenti analogici nelle fasi iniziali o per apportare correzioni in fasi intermedie del lavoro. Infine, il 14% ha un processo ibrido che rimane ancora fortemente legato alla scrittura su carta (Figura 2.2). Analizzando queste percentuali in relazione all'età anagrafica, emerge che l'adesione a processi prevalentemente digitali non sembra dipendere da fattori generazionali (Figure 2.3 e 2.4).

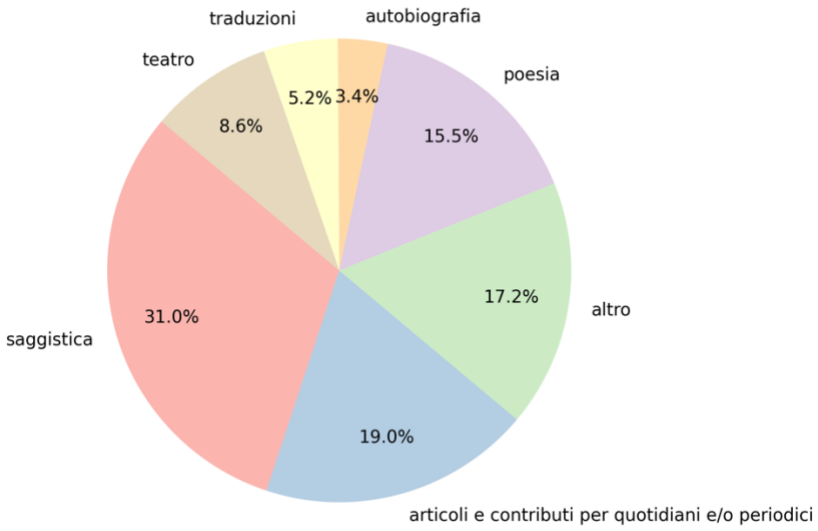


Figura 2.1 Distribuzione dei generi e delle tipologie testuali prodotte dagli autori oltre alla narrativa.

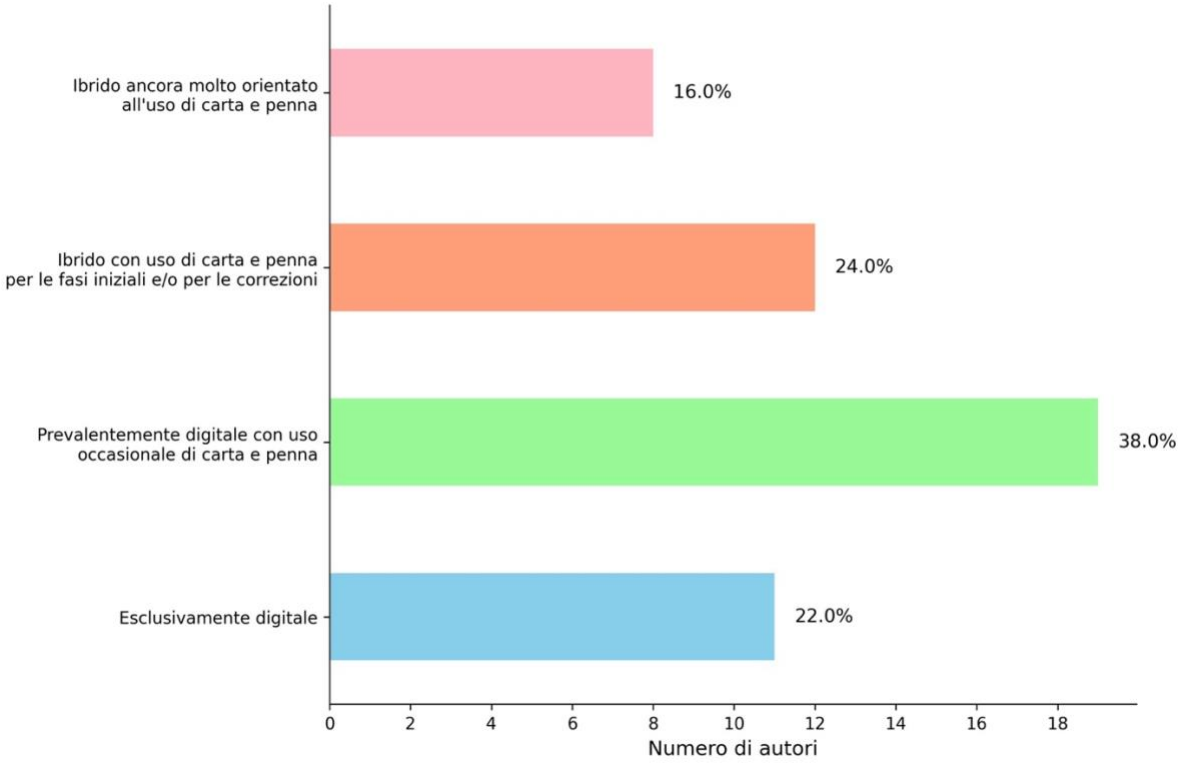


Figura 2.2. Distribuzione delle tipologie di processo di scrittura.

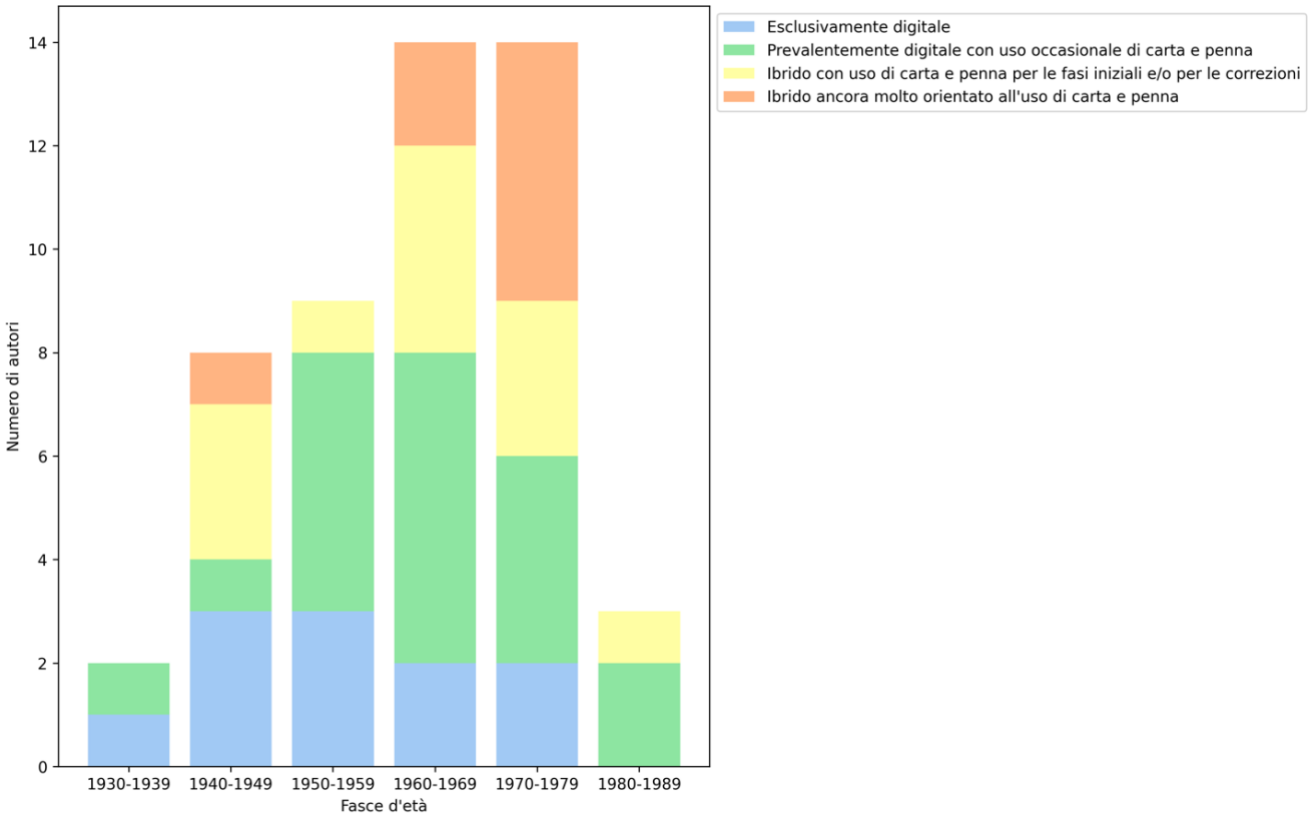


Figura 2.3. Distribuzione delle tipologie di processi di scrittura per fasce d'età.

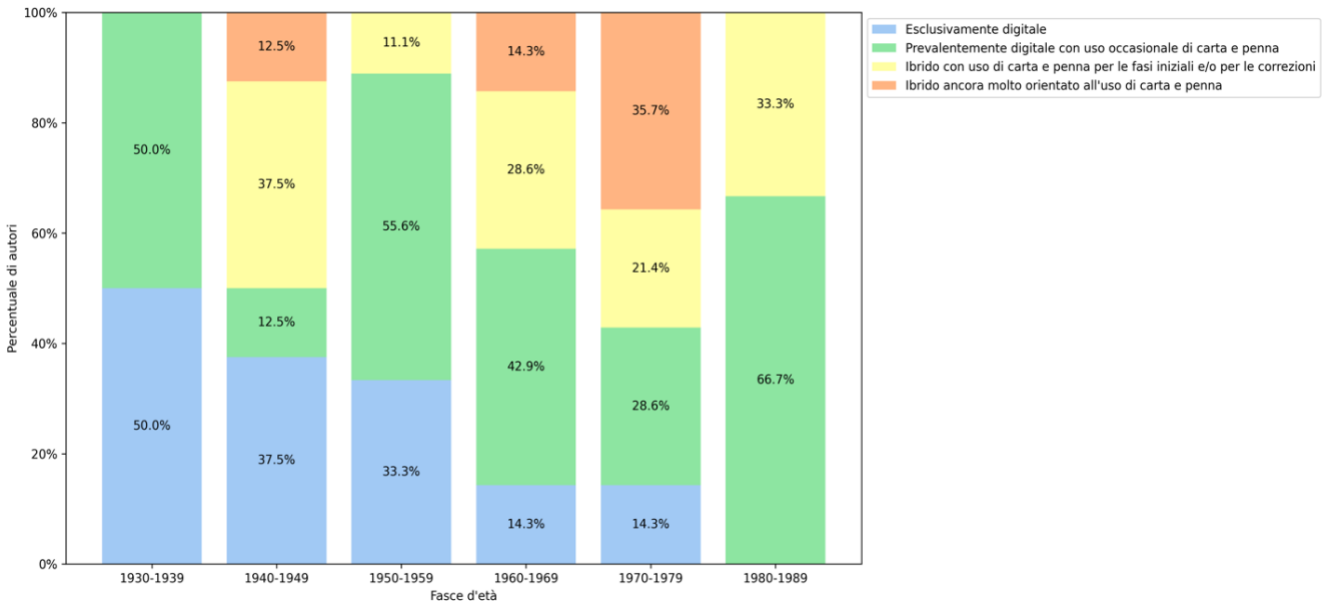


Figura 2.4. Distribuzione delle tipologie di processi di scrittura per fasce d'età (espressi in percentuale).

Al contrario, il processo interamente digitale ha i suoi picchi più alti nei nati tra il 1930 e il 1959. Gli intervistati appartenenti alla fascia 1930-1939 sono due: Biondi scrive «soltanto digitale, dalla fine degli anni Ottanta» e Garavini afferma che a mano non scrive quasi più nulla, perché ormai ben «più abituata a usare il computer» (Giagnolini 2025b)[63]. Nella fascia 1950-1959, ben l'88,9% degli intervistati segue processi totalmente o prevalentemente digitali. Nelle prime tre fasce figurano quelli che Domenico Fiormonte, nell'individuare alcune categorie per identificare le diverse attitudini degli scrittori nell'uso del computer, definirebbe «*aficionados* della prima ora» (2003, 71). Come Biondi, infatti, anche Meldini: «dal 1987, anno in cui ho acquistato il mio primo pc (un Amstrad PCW con sistema operativo CPM Plus), scrivo esclusivamente al computer; la penna mi serve solo per rapidi appunti e brevi annotazioni che in genere, una volta utilizzati, cestino». Fra i pionieri anche Maggiani e Pecoraro, accomunati dall'aver iniziato la carriera di scrittori in età adulta utilizzando i primissimi modelli in circolazione. Maggiani inizia a scrivere con un Macintosh nel 1985 e ricorda come fossero «tempi in cui i computer MS-DOS avevano sullo schermo dei pixelini verdi, che tra l'altro era complicato gestire; invece, Apple mise in vendita questo prodotto, per nulla complicato, ed era un modo completamente diverso di lavorare, un modo nuovo». Ed è proprio con i comandi DOS che inizia a scrivere Pecoraro, grazie al computer dell'allora compagna. Con una metafora materiale, afferma che scrivere al computer per lui era come

[63] Al fine di garantire una lettura più scorrevole e ridurre la ridondanza di riferimenti bibliografici associati alle citazioni dei testi delle interviste oggetto di questo studio, tutte le citazioni presenti in questo capitolo, quando non accompagnate da un riferimento bibliografico esplicito, si riferiscono al dataset delle interviste disponibile al link: https://zenodo.org/records/15063775.

«scrivere sull'acqua», rispetto alla macchina da scrivere, che equivaleva a «scrivere sulla pietra». Ma fra le risposte figura anche chi, come Lugli, pur essendo migrato ad un processo prevalentemente digitale, ha vissuto questo passaggio come un trauma:

> credo di appartenere a quella generazione che è stata traumatizzata, destrutturata dall'arrivo della digitalizzazione. [...] Con l'arrivo del computer ho iniziato a scrivere in digitale ma stampavo tutto: non riuscivo a concepire una correzione che non fosse cartacea. Poi è cambiato tutto e adesso non concepisco nemmeno l'idea di stampare: tutto quello che ho sono file.

Magrelli scrive (e detta) in digitale anche per "economia":

> Per me è stato fondamentale scoprire anni fa l'app *Dragon Dictation*, che per la prima volta mi ha dato la possibilità di dettare un testo e di ritrovarlo per iscritto. È stato veramente un passaggio fondamentale: [...] ricordo di aver scritto interi articoli dettando a questa sorta di segretario che ormai è alla portata di tutti noi. Ricordo ancora un vecchio critico che mi diceva: "ma perché non fai come Henry James, che dettava al suo segretario?" e la mia risposta era: "perché io guadagno meno del segretario di Henry James".

Tuttavia, in questa analisi, così come per lo studio delle risposte successive, è essenziale considerare la distribuzione del campione: oltre la metà degli intervistati (28 su 50) è nato tra il 1960 e il 1979. Questo dato, quindi, non permette di generalizzare le osservazioni sulla distribuzione anagrafica di determinati fenomeni. Tuttavia, insieme ad alcune risposte fornite nelle domande successive, contribuisce a sfatare il mito secondo cui l'età avanzata implichi necessariamente una nostalgica preferenza per la carta. Al contrario, è proprio fra gli scrittori più giovani, nati fra gli anni Settanta e Ottanta, che troviamo le affermazioni più forti legate alla scrittura su carta. Per Montanaro, ad esempio, la matita «pare una forma di possesso del mondo ancora fortissima». Cilento e Romagnolo affermano di *pensare* grazie al movimento della mano sulla carta. Cilento passa al digitale solo «[q]uando un racconto o un romanzo iniziano a consistere di molto materiale scritto su carta (il che per i romanzi può voler dire anche accumuli di due o tre anni di scritture esplorative)»: il digitale è innanzi tutto operazione di trascrizione ragionata. Per Romagnolo il passaggio analogico è fondamentale soprattutto per l'inizio della stesura: «qualsiasi cosa rappresenti un cominciamento ha bisogno del passaggio su carta. [...] È come se mi servisse per attivare il cervello". Per Galletta la scelta fra scrittura analogica o digitale dipende principalmente dalle condizioni materiali in cui si trova. Afferma, per esempio, di aver scritto l'ultimo libro usando solo quaderni americani, perché si trovava a casa dei genitori in Sicilia, senza computer. Per Galletta «scrivere a mano e scrivere al computer sono azioni che hanno una connessione diversa con la testa [...]. Scrivere

molto lentamente e con fatica per alcuni tipi di cose che scrivo mi è più congeniale, è più funzionale». Simile considerazione ricorre nelle parole di Grossi (che con Galletta condivide anche la passione per i quaderni americani):

> [c]on carta e penna mi abbandono ad una specie di scrittura automatica. Questo per nessun tipo di romanticismo, voglio precisare, ma semplicemente perché mi viene meglio così. Sarebbe molto più comodo poter scrivere direttamente in digitale, mi farebbe risparmiare molto tempo, ma quando scrivo la prima stesura ho bisogno di staccare la parte critica, la parte "da lettore" del mio cervello, e seguire le mie immagini. E questo mi viene molto più naturale scrivendo a mano.

In contrapposizione, Montanaro fa emergere un tema ricorrente e trasversale: «[i]l digitale mi permette di seguire rapidamente i miei pensieri» perché «sostanzialmente il pensiero si traduce immediatamente in "inchiostro"».

Nonostante alcune persistenze del cartaceo, è evidente che a distanza di quarant'anni dall'indagine di Curino la scrittura è diventata impensabile senza l'utilizzo del computer. Non solo: il processo creativo ha sconfinato anche verso altri dispositivi, che questa ricerca ha cercato di intercettare per sondare l'estensione dell'archivio d'autore su supporti differenti. Ne è emerso che il 60% degli autori utilizza altri dispositivi oltre al computer, nella fattispecie il cellulare (utilizzato dal 46% degli scrittori) e il tablet (dal 24%). Il 10% dispone di tutti e tre i dispositivi[64] (Figura 2.5).

Per quanto riguarda le finalità d'uso, quattordici autori impiegano tablet e smartphone esclusivamente per appunti rapidi, mentre altrettanti li utilizzano anche per processi di scrittura più strutturati; soltanto due autori adottano entrambi gli approcci. È interessante notare come, anche in contesti di scrittura più articolati, emerga una preferenza per l'uso di tablet e smartphone nella realizzazione di testi caratterizzati da rapidità e ridotta estensione, tipicamente in situazioni emergenziali o laddove la praticità del mezzo risulti superiore all'utilizzo del computer. Curiosamente, però, c'è chi usa lo smartphone specificamente per scrivere poesie, come Giartosio e Pugno. Anche se non emerge direttamente dall'intervista, anche per Magrelli deve essere un'abitudine, data la presenza del file "poesie_cell" all'interno dei materiali che ha conferito al Centro Manoscritti (Milone 2025, 70). Conte, invece, usa carta e penna solamente per le poesie.

[64] Dall'integrazione con i dati emersi da altre risposte, risulta anche un 10% di autori che utilizza due computer.

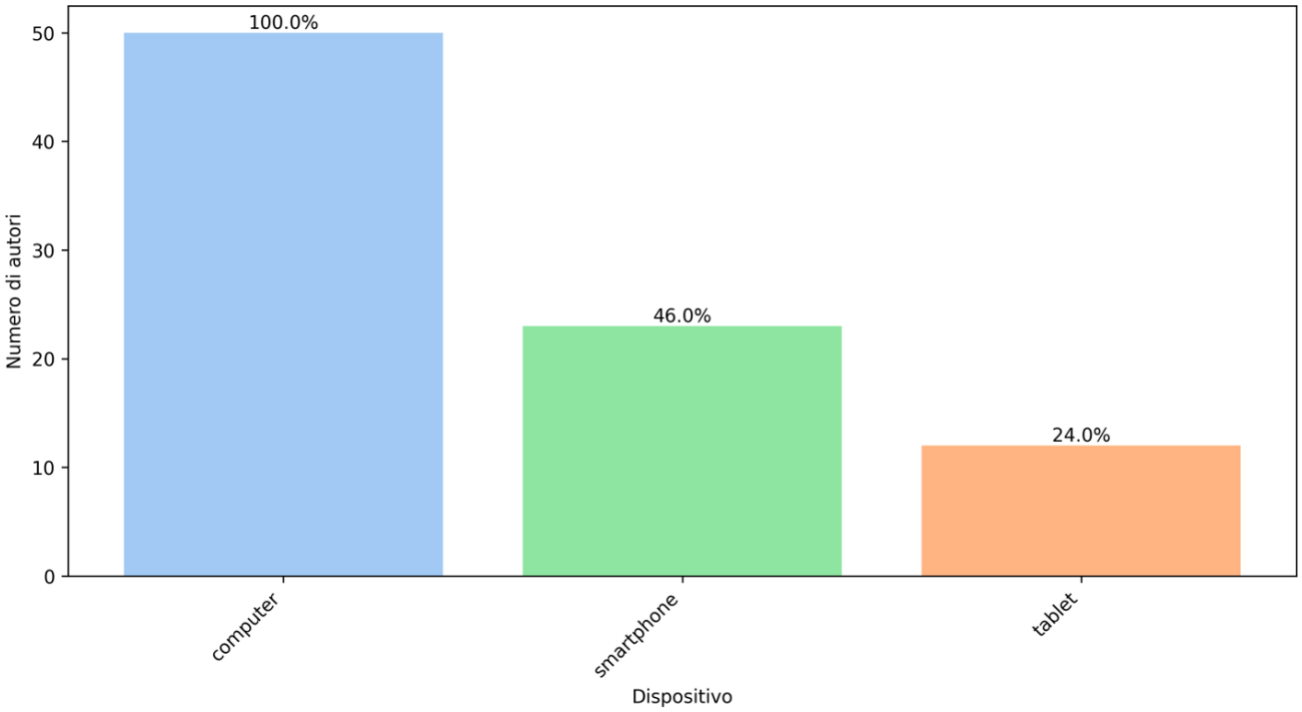


Figura 2.5. Distribuzione delle tipologie di dispositivi utilizzati per la scrittura.

Non sembra esistere un nesso evidente fra la tipologia di dispositivi utilizzati e il dato anagrafico (Figura 2.6), ma un ostacolo concreto nell'utilizzo degli smartphone da parte delle generazioni più mature potrebbe essere di natura fisiologica, come premette Maggiani: «per me il cellulare non è utilizzabile per la scrittura perché non vedo cosa c'è scritto». Non emerge alcuna correlazione significativa nemmeno fra il numero di dispositivi utilizzati e l'età anagrafica (Figura 2.7), ma un corollario importante del dell'uso combinato di dispositivi, dal punto di vista filologico e archivistico, è il dialogo di bitstream che si instaura fra di essi. Se il processo creativo si sviluppa su più supporti, quali strategie di trasferimento di file vengono adottate? E, dunque, quali e quante tracce digitali ne rimangono? La tecnica più diffusa è, senza dubbio, l'auto-invio di email, praticata dal 42% degli autori (Figura 2.8). L'uso dell'email è seguito, con un certo distacco, dall'affidamento a sistemi cloud, adottati dal 20% dei rispondenti. Questo strumento viene preferito da quegli autori che prediligono l'accesso a una copia sempre aggiornata del materiale, a prescindere dal dispositivo utilizzato. Il tradizionale utilizzo della chiavetta USB riguarda il 18% degli intervistati, seguito dalle operazioni di vera e propria trascrizione da un dispositivo all'altro (8%), talvolta effettuate tramite dettatura (2%). Cavalli fa riferimento al "copia-e-incolla", probabilmente alludendo a sistemi di sincronizzazione che permettono di copiare e incollare contenuti tra dispositivi collegati. Allo stesso modo, Lupo menziona lo scambio di file tramite Bluetooth, probabilmente in relazione alla funzione AirDrop di Apple.

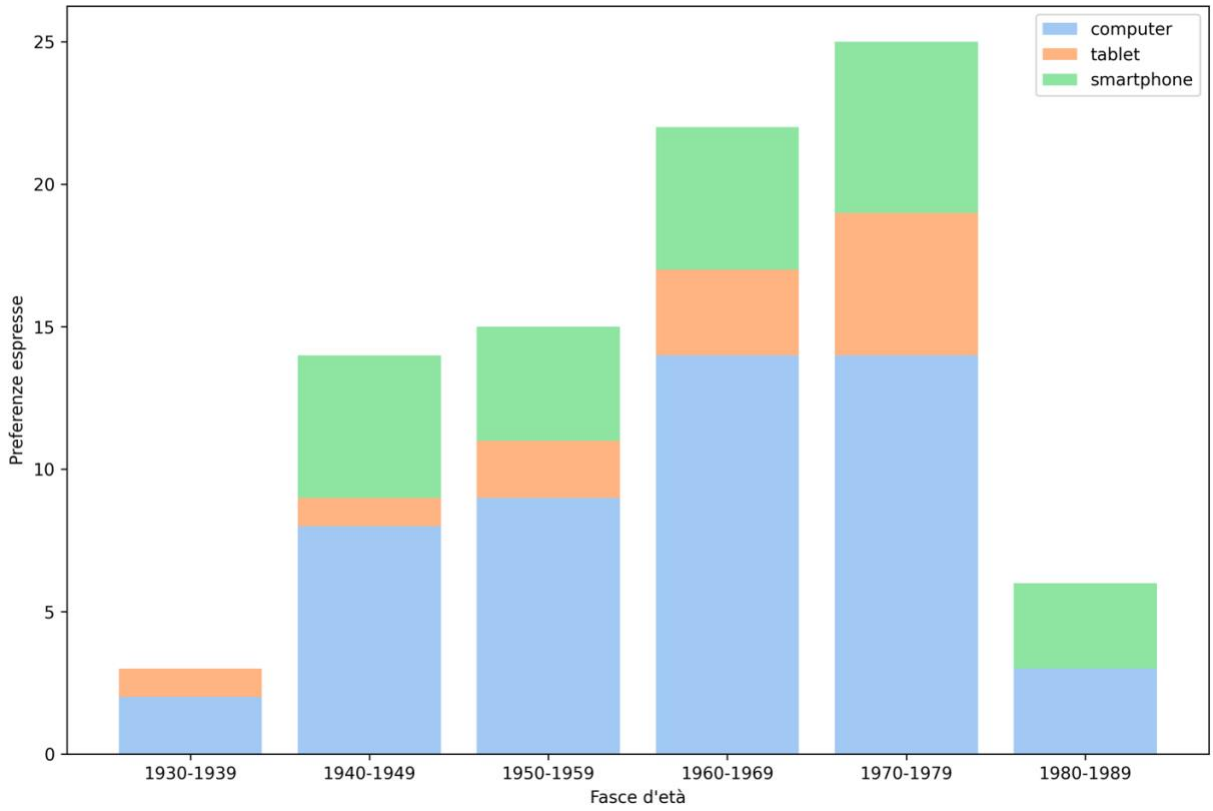


Figura 2.6. Distribuzione dei dispositivi utilizzati per la scrittura per fasce d'età.

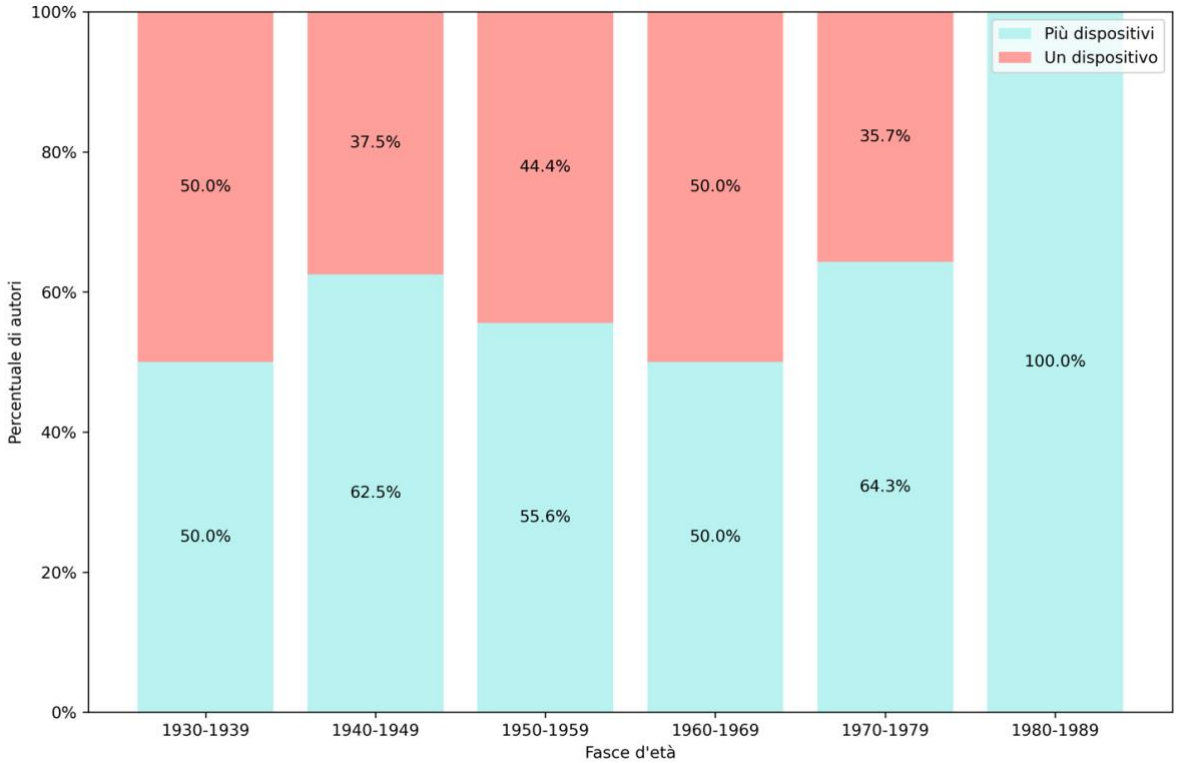


Figura 2.7. Distribuzione del numero di dispositivi utilizzati per fasce d'età espresso in percentuale.

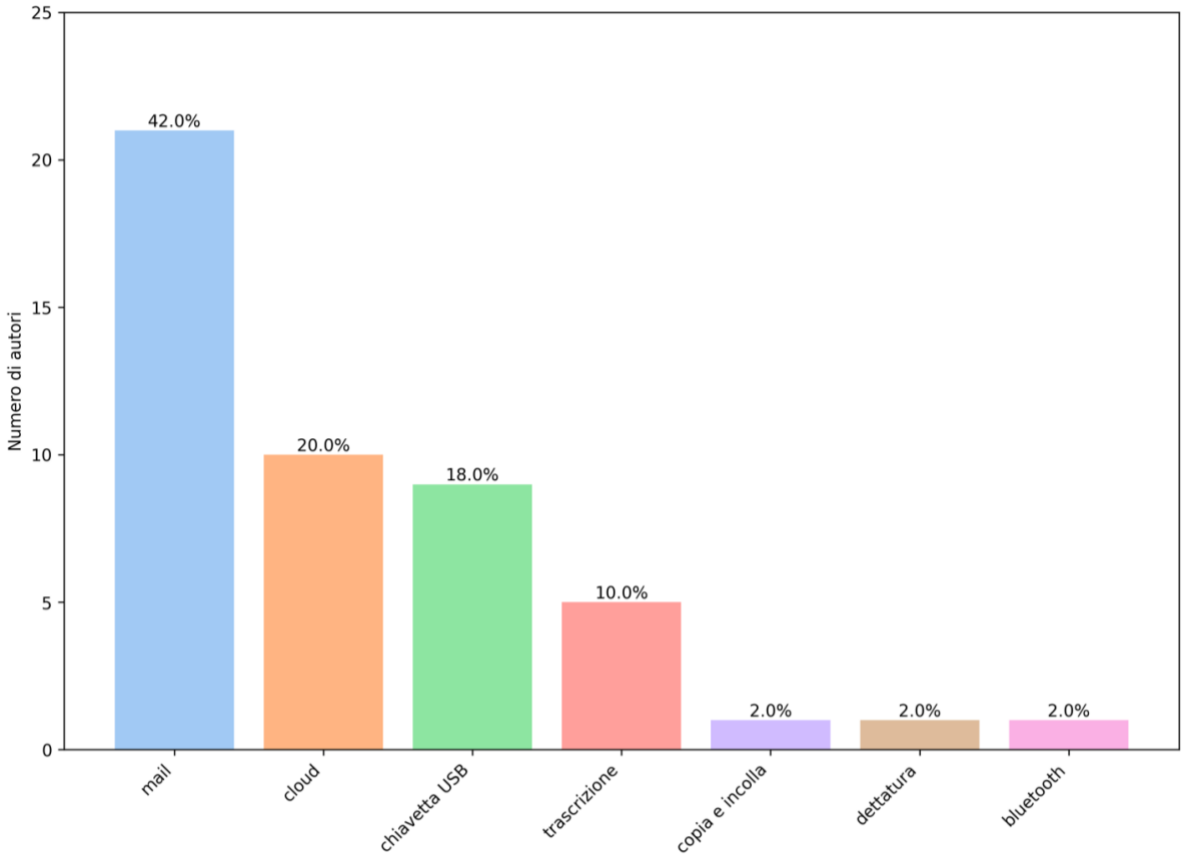


Figura 2.8. Distribuzione delle strategie utilizzate per il trasferimento di file fra dispositivi.

2.3.2 Modalità di gestione e archiviazione documentale

La domanda relativa alla natura del proprio archivio e del rapporto che intercorre fra analogico e digitale ha fornito risposte difficili da categorizzare per il loro carattere idiosincratico. Il dato complessivo indica che la maggioranza degli autori possiede un archivio ibrido, sebbene consistenza, composizione ed estensione temporale dei materiali varino notevolmente. Un gruppo di autori [65], rintracciabili principalmente fra coloro che hanno ancora un processo pienamente ibrido, considera l'analogico e il digitale in stretto dialogo, attribuendogli la stessa importanza nel processo creativo. Spirito, ad esempio, non riesce nemmeno a concepire «una produzione di carattere letterario o saggistico senza avere l'appoggio di un archivio [analogico] a portata di mano». Altri autori ricorrono al materiale cartaceo solo in una fase embrionale del processo, per attività di autodocumentazione o per recuperare idee[66]. Fra questi Frizziero, ad esempio, riserva più importanza al digitale perché è la partizione d'archivio che contiene materiali già rielaborati e vagliati. Similmente, anche Tarabbia: «l'archivio è ibrido, ma ogni cosa prima

[65] Si segnalano, in particolare, le risposte di Cilento, Giacopini, Grossi, Lupo e Spirito.
[66] Si vedano a proposito le risposte di Balzano, Cavalli, Frizziero, Tarabbia e Galletta.

o poi finisce in un file». Per Galletta, invece, la centralità spetta all'analogico, perché contiene la documentazione più significativa dal punto di vista dell'atto creativo. Il valore della carta emerge anche in Montanaro, ma per l'ultimo anello del processo: ha una copia cartacea di tutto quello che ha pubblicato. Della sua intera produzione mancano all'appello solo due pezzi; sa quali sono e perché non sono nel suo archivio: «Il fatto che manchi un pezzo lo rende ancora più umano, più bello, più significativo. [...] Non è la stessa cosa vederlo online, anche perché la carta esiste ed occupa spazio. [...] Istintivamente quello che per me ha un valore, deve comunque diventare carta. Altrimenti per me non esiste».

Avalli, nel confrontare il cartaceo e il digitale, formula un'affermazione di notevole interesse: «li considero, e sono, un testo unico». Questa dichiarazione, incisiva e sintetica, offre un'importante prospettiva su come occorrerebbe orientare l'approccio alla curatela e allo studio di tali archivi. In contrasto, Affinati comunica che nel suo archivio «di fatto si è creata una cesura fra il mondo cartaceo e quello digitale». Affinati, infatti, appartiene a un altro raggruppamento di autori che, pur riconoscendo il portato del proprio archivio cartaceo, lo lega principalmente ad una dimensione temporale passata. Nel suo caso, l'ibridismo non è parte del processo creativo attivo, ma deriva dalla sedimentazione dell'informazione su supporti diversi in seguito al mutamento del modus operandi[67].

Otto autori affermano di avere un archivio esclusivamente digitale perché hanno una produzione effettivamente esclusivamente nativa digitale o perché, pur avendo una minima produzione cartacea, se ne disfano o non la considerano parte del proprio archivio. In quest'ultima casistica rientra Calandrone: «tutto quello che scrivo, a prescindere dal supporto, una volta che è stato rielaborato e portato compimento lo elimino, cancello le prove»[68].

È poi interessante osservare come la biblioteca d'autore (Braida e Cadioli 2011) – intesa qui come raccolta di edizioni di opere proprie e come insieme dei volumi consultati nel processo creativo – venga inclusa pienamente nel concetto di archivio da diversi scrittori. A riguardo, le risposte di Caminito, Conte, Parrella e Terranova mettono in luce una dimensione dinamica d'archivio, in continuo dialogo ed evoluzione con le fonti librarie.

Zaccuri sottolinea come la graduale "informatizzazione" del suo archivio sia avvenuta anche attraverso il passaggio da plichi di fotocopie ai PDF, amplificando quello che definisce l'"effetto Umberto Eco":

[67] Quadri simili appaiono anche nelle risposte di Biondi (che considera il suo archivio cartaceo "di natura quasi preistorica") Capriolo, Maggiani, Magrelli, Matteucci, Morazzoni, Montesano e Stassi.
[68] Simile procedura si riscontra anche nelle risposte di Lugli, Matteucci, Montesano e Venezia.

accumulare un numero eccessivo di materiali «seguendo la filosofia del *just in case*, ovvero con la convinzione che, prima o poi, potrebbero tornare utili», ma che raramente verranno ripresi in mano.

Nelle risposte di Affinati, De Marchi, Garavini, Orecchio e Stassi si individuano operazioni di digitalizzazione del cartaceo finalizzate a migliorarne l'accessibilità, sia per i documenti di ricerca che per quelli analogici di propria produzione. Romagnolo attua invece il processo inverso, stampando anche le fonti che trova nativamente in digitale.

Il ricorso alla stampa, come evidenziano diversi autori, è un importante anello di congiunzione fra analogico e digitale e rimane una forte esigenza: l'80% dichiara di ricorrere alla stampa, contro il 20% che ormai ne fa a meno. Stampare viene percepito come un processo di oggettivazione e oggettificazione necessario finalizzato soprattutto alla lettura (tendenzialmente affiancata alla correzione) (Figura 2.9). Del resto, numerosi sono gli studi che ne testimoniano i benefici, rispetto ad una lettura digitale (ad esempio: Lenhard, Schroeders, Lenhard; Mangen, Walgermo, Brønnick; Singer, Alexander). Lo schermo, per usare le parole di Montesano «è troppo benevolo e fantasmatico».

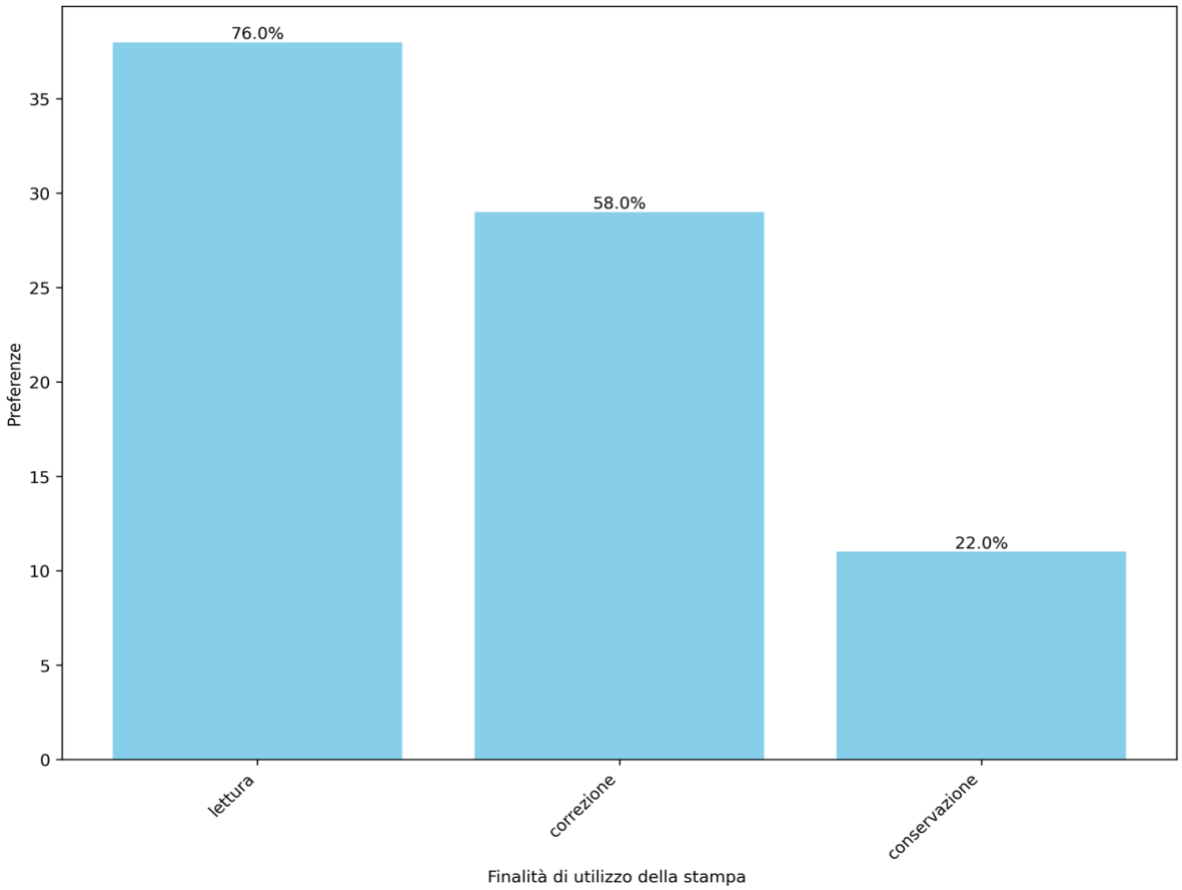


Figura 2.9. Distribuzione delle finalità di stampa.

Il 22% degli autori generalmente conserva le stampate, mentre il 14% stampa ma cestina una volta terminate lettura e correzioni[69]. A questo proposito, Lupo si riaggancia al valore dell'analogico:

> Devo dire che l'idea di non avere il cartaceo, di gettare le stesure, le carte, le diverse redazioni di un libro, l'idea di liberarsene, mi mette in inquietudine. Mi sentirei come un albero senza radici. E le radici sono proprio gli appunti, le stesure, le bozze: non concepisco un libro senza radici. Poi, magari, sono delle radici che nessuno mai andrà a vedere o che io stesso non prenderò più in mano.

Da una lettura trasversale delle risposte fornite a questa domanda e alla precedente, emerge un gruppo di autori in linea con le parole di Lupo, che considera il rapporto con la carta come un vero e proprio legame affettivo[70], talvolta senza attribuirle altro ruolo. Ironicamente, altrettanti autori fanno riferimento alla conservazione in questa chiave come a una sorta di feticismo[71]. Fra gli autori che, invece, hanno completamente smesso di stampare, spicca la considerazione di Affinati:

> Una volta li stampavo, adesso non più. Anche le bozze le correggo sullo schermo. Come se ci fosse stata in me una mutazione nel rapporto con la scrittura. E quindi anche con l'esperienza della realtà. Credo si tratti di un cambiamento epocale di grande portata, le cui conseguenze sono ancora imprevedibili. Il mio atteggiamento a tale riguardo è positivo, senza posizioni ideologiche precostituite.

Maggiani e Pugno correlano questo fenomeno anche al completamento del processo di digitalizzazione della filiera editoriale. La tendenza alla stampa, comunque, anche laddove permane, appare generalmente diminuita rispetto al passato e posticipata ad una fase avanzata del lavoro[72].

Per quanto riguarda la denominazione dei file, il titolo del romanzo (definitivo o più spesso provvisorio) si è rivelato, prevedibilmente, un criterio molto comune. La strategia più utilizzata, attestatasi al 36%, è la combinazione del titolo con una numerazione progressiva a indicare il susseguirsi delle versioni. Abbastanza comune anche l'unione di titolo e data di stesura (18%)[73]. Talvolta vengono anche utilizzate

[69] Nelle risposte a questa e alla precedente domanda, Lugli e Spirito riferiscono di ricorrere allo scarto anche per il ben noto problema dello spazio.
[70] Si vedano, per esempio, le considerazioni di Canobbio, Magrelli, Mozzi, Montanaro, Zeno.
[71] Si vedano le considerazioni di Grossi, Montesano, Romagnolo, Rasy (alla domanda n. 9) e Zeno.
[72] Si vedano, a riguardo, le testimonianze di Balzano, Calandrone, Garavini, Giartosio, Matteucci, Orecchio, Parrella, Pellegrini, Pugno, Ranieri, Tarabbia e Venezia.
[73] La data di stesura o di ultima modifica viene ricercata anche nei metadati oppure nel contenuto (appuntata in calce al documento) come metodo per rintracciare precedenti stesure.

parole chiave che aiutino, come afferma Rasy, «il teatro della memoria» (16%). Seguono poi concatenazioni molto soggettive, come sigle, suffissi e prefissi, oppure il nome del revisore a cui il file è stato dato in lettura. Il 16% degli autori dichiara di non avere un criterio specifico per la denominazione dei file, che spesso si traduce con l'utilizzo di termini piuttosto randomici o criteri misti, talvolta con il risultato che all'autore stesso risulta difficile ritrovare i suoi stessi file[74].

La denominazione dei file è la principale modalità utilizzata per gestire il *versioning* del proprio lavoro, ossia la distribuzione delle varie stesure (62%). Non mancano processi apertamente ibridi: il 12% degli autori afferma che ricostruisce o potrebbe ricostruire i vari momenti di lavoro grazie all'alternanza di file e stampate. Solo tre autori menzionano esplicitamente la stratificazione delle versioni nelle caselle email, un dato sorprendente considerando le numerose risposte che attribuiscono allo scambio di email un ruolo centrale nel processo, soprattutto come forma di backup[75]. Il 10% è in grado di recuperare versioni precedenti in considerazione della specifica della cartella in cui si trovano. Chi salva più versioni, comunque, tende a crearne una nuova solo in caso di modifiche rilevanti. Al contrario, il 30% degli scrittori non ne tiene affatto traccia: il 22% lavora sempre sullo stesso file, mentre l'8% elimina le versioni precedenti. Tra questi, tre autori conservano però le stampate intermedie[76].

Nella maggior parte dei casi, i file vengono organizzati attorno al concetto di libro, inteso come opera e/o progetto (58%). Tuttavia, il livello di profondità in cui questo avviene nella struttura del *file system* varia: alcuni scrittori, ad esempio, adottano una suddivisione preliminare per tipologia documentale, come Meldini: «i files di Word (doc e docx) sono ordinati nelle seguenti cartelle: Articoli, Lettere, Racconti, Romanzi, Studi, ecc. La cartella Romanzi è a sua volta ordinata in sottocartelle intitolate ai singoli romanzi, e così via». Chi scrive anche per la stampa tende a creare cartelle intitolate "articoli" o salvate con il nome della testata, talvolta suddivise per annate in caso di collaborazioni durature.

All'interno delle cartelle dedicate ai romanzi, molti autori adottano ulteriori suddivisioni per distinguere, ad esempio, materiali di lavoro, vecchie stesure, prove di copertina, materiali promozionali.

Sette autori affermano di non adottare alcuna particolare organizzazione per i propri file, la cui gestione, secondo le parole di Tuena «[d]ipende un po' dal caso, dalla fretta, dalla confusione che regna sulla mia homepage».

[74] Si vedano, per esempio, le risposte di Maggiani e Venezia.
[75] Si veda, a questo proposito, il capitolo 3.3.
[76] Dato che emerge in De Marchi, Lupo e Spirito.

In generale, emergono due approcci: da un lato, una curatela assidua o almeno periodica; dall'altro, una stratificazione cronologica di materiali (più o meno organizzati a seconda dei casi). Janeczek, ad esempio, lascia che i file di lavoro si accumulino nella cartella di sistema *Documenti*, ma crea cartelle dedicate ai materiali di supporto. Piersanti, invece, adotta un metodo più strutturato, riorganizzando periodicamente i file per periodi cronologici e scartando quelli irrilevanti. Operazioni, tuttavia, non sempre soddisfacenti o sufficienti: lo stesso Piersanti definisce il proprio archivio come «un labirinto irrisolvibile».

2.3.3 Strategie di conservazione e trasmissione

Il primo aspetto indagato nell'ambito delle strategie di conservazione e trasmissione digitale riguarda l'attenzione alla scelta del formato di file, assicurata dal 92% degli autori. Il dato è certamente mediato dal fatto che la maggioranza degli autori scrive utilizzando l'applicativo Microsoft Word e si affida al formato di salvataggio standard (.docx).

Interrogati sulle abitudini di backup, il 94% ha dimostrato di effettuare backup completi o parziali[77], mentre solo il 6% ha dichiarato di non curarsene. Il 44% si affida a backup su hard disk esterno, il 30% su cloud. Una sola autrice, Pugno, utilizza anche un sistema NAS (Network Attached Storage). Spiccano, poi, due percentuali significative che indicano come molti autori effettuino il backup non dell'intero contenuto, ma di singoli file di lavoro[78]: il 28% dichiara di effettuare backup via email (inviando a se stessi o ad altri il documento in allegato[79]), il 22% su chiavetta USB (Figura 2.10).

[77] Sono stati considerati nel computo anche gli autori che adottano misure minimali di precauzione e salvataggio dati; dunque, anche gli autori che alla domanda generale hanno risposto "no", ma che con un'integrazione successiva hanno riportato, ad esempio, l'abitudine di auto-inviarsi mail o salvare file in chiavetta.

[78] Da alcune risposte, anche i sistemi cloud risultano utilizzati, talvolta, in modo simile. Balzano, per esempio: "salvo anche su cloud le cose che mi interessano, ma non l'intero contenuto del mio computer".

[79] Particolarmente significative, a riguardo, le risposte di Frizziero, Grossi e Parrella.

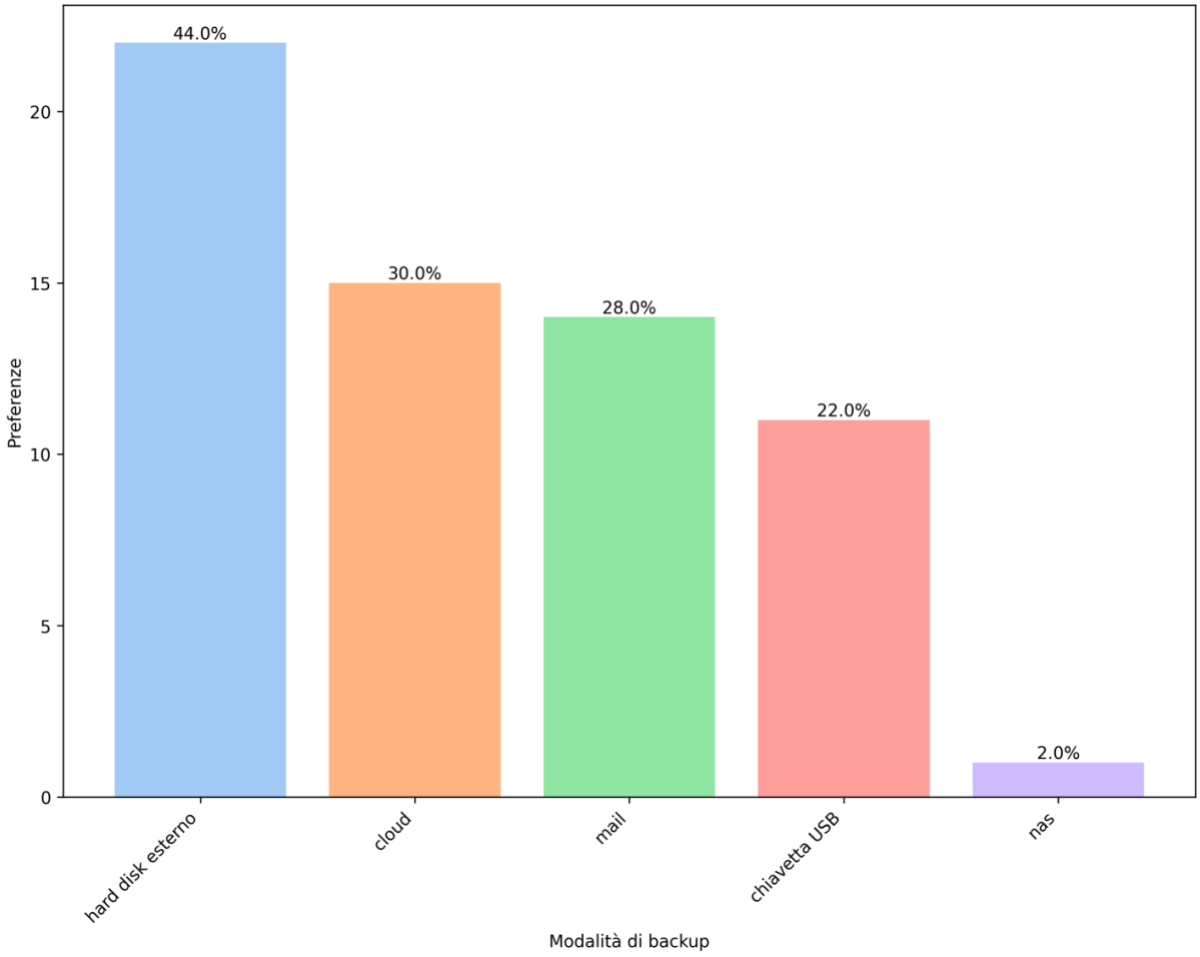


Figura 2.10. Distribuzione delle modalità di backup.

Relazionando la distribuzione delle tipologie e del numero di dispositivi utilizzati con l'età dei partecipanti non sembrano esistere particolari correlazioni (Figura 2.11 e Figura 2.12). Piuttosto, si delineano tendenze trasversali, tra cui la messa a punto di procedure di backup in seguito a episodi di perdita di materiali, percepiti come eventi traumatici[80]. Per Piersanti è stato uno «shock psichico»; Grossi lo ricorda «con un tale senso di frustrazione – per usare un eufemismo» che da quel momento è diventato estremamente scrupoloso nel fare backup. Balzano, Calandrone e Piersanti, memori di un furto, appena hanno potuto si sono rivolti a backup su cloud. Magrelli, finito lo spazio cloud, ha optato per una certa metodicità: «ho comprato ben due hard disk e tutti i venerdì – la Passione di Cristo – copio sui due hard disk tutto quello che ho".

[80] Si segnalano, in particolare, le esperienze riportate da Balzano, Calandrone, Grossi e Piersanti.

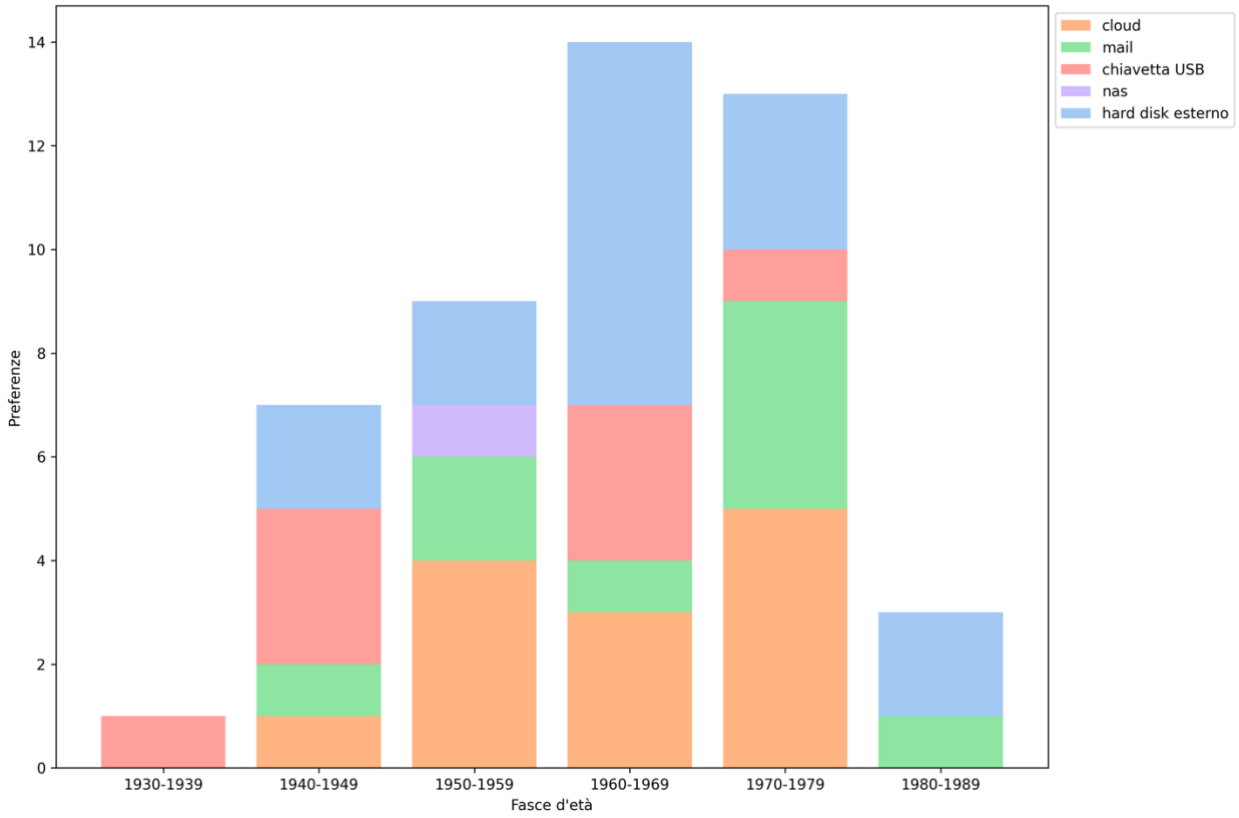


Figura 2.11. Distribuzione delle modalità di backup utilizzate suddivise per fasce d'età (gli autori hanno espresso una o più preferenze).

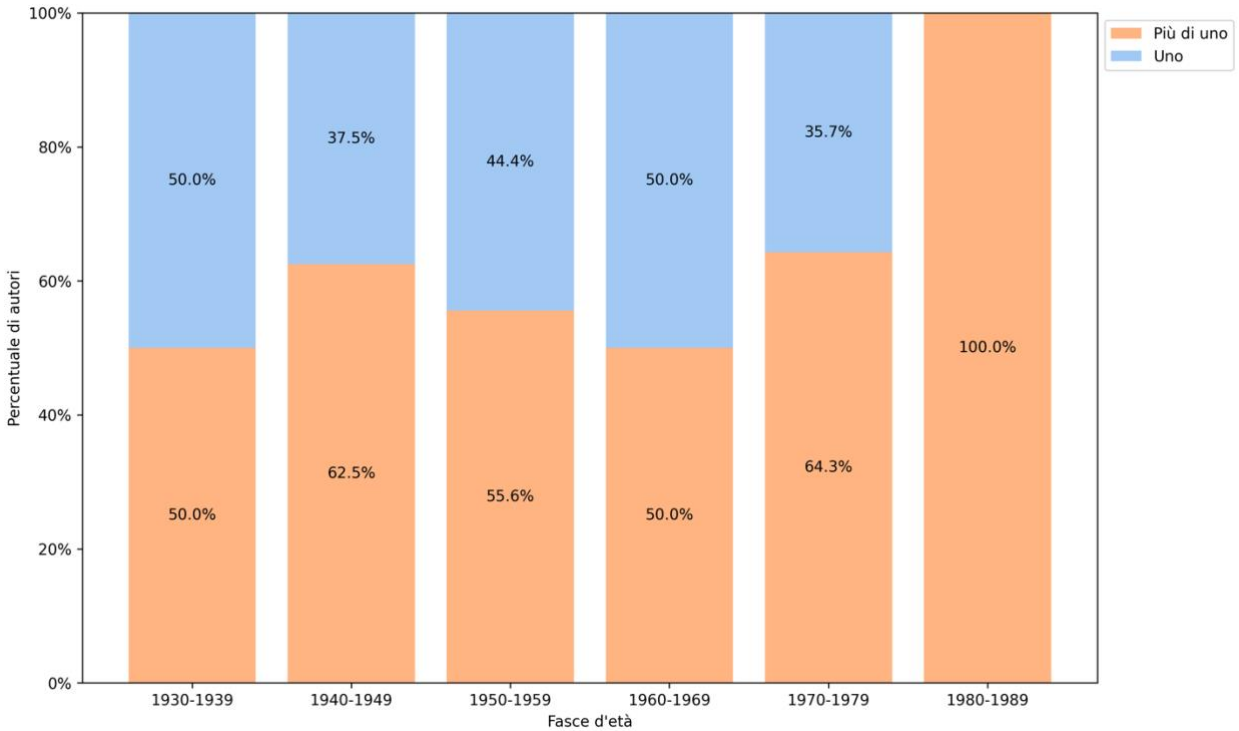


Figura 2.12. Distribuzione dell'utilizzo di uno o più dispositivi espressa in percentuale.

La consapevolezza dimostrata dagli autori in relazione all'importanza di effettuare il backup è piuttosto generalizzata, anche se presenta sfaccettature diverse. Il problema dell'obsolescenza, ad esempio, è particolarmente chiaro a coloro che hanno attraversato diversi cambiamenti tecnologici. Zaccuri e Spirito hanno iniziato delle operazioni di recupero di file dai loro vecchi floppy disk. Cavalli, che era solito fare copie su chiavetta USB (di tipo A), ha cambiato questa abitudine perché sempre meno computer hanno l'ingresso compatibile. Piersanti ha la sensazione che il digitale sia «di una fragilità mostruosa» e che un giorno ripiangeremo la scrittura manoscritta, perché con gli attuali ritmi di sviluppo «documenti non possano durare. Proprio come è successo con i CD: tra due anni cambieranno di nuovo tutti i formati».

Galletta e Lupo dimostrano una certa consapevolezza sia tecnologica che etica, problematizzando il fatto che i cloud sono sistemi proprietari. Galletta li utilizza comunque, ma fa fronte a questo tema procedendo con periodici backup su dischi rigidi; Lupo, invece, li rifiuta per via dell'insofferenza «nei confronti della tecnologia che assuma qualsiasi genere di controllo».

La diffusione della pratica di backup nella gestione documentale degli autori, seppur motivata da esigenze personali, potrebbe riflettere una forma di curatela consapevole riconducibile, implicitamente, al concetto di volontà d'archivio (Italia e Zanardo 2023; Giagnolini 2023, 93). Nella presente indagine, si è scelto di esplorare il tema in maniera esplicita, chiedendo agli autori se avessero mai considerato il conferimento del proprio archivio ad una istituzione culturale, in particolare della partizione digitale. Il 42% aveva già preso in considerazione questa opzione, mentre il 58% non ci aveva mai riflettuto. Lo farebbe, effettivamente, il 44% degli autori; il 24% rimane in dubbio, il 18% si oppone all'idea[81] (Figura 2.13). Dai dettagli delle risposte, emergono una pluralità di approcci e sensibilità. Da un lato, sei dei nove autori che escludono la possibilità di un conferimento giustificano la loro posizione sostenendo che solo ciò che viene effettivamente pubblicato meriti di essere reso pubblico e di perdurare nel tempo, e che ciò debba avvenire esclusivamente nella forma di libro[82]. In questa prospettiva, il resto del materiale viene considerato come privato o di scarso valore, in una posizione anti-filologica che trova espressione nelle parole di Pincio: «Sono per il testo che resta. [...] Il testo deve vivere in sé per sé, non perché ha alle spalle una serie di pezze di appoggio». Parrella approfondisce ulteriormente la sua idea richiamando il tema delle pubblicazioni postume e reclama il diritto all'oblio di ciò che l'autore decide di non pubblicare.

[81] Dal 14% non è stato possibile individuare nettamente una preferenza.

[82] Si vedano le risposte fornite da Calandrone, Conte, Matteucci, Parrella, Pincio e Zeno. Montesano esprime lo stesso concetto in risposta alle pratiche di backup.

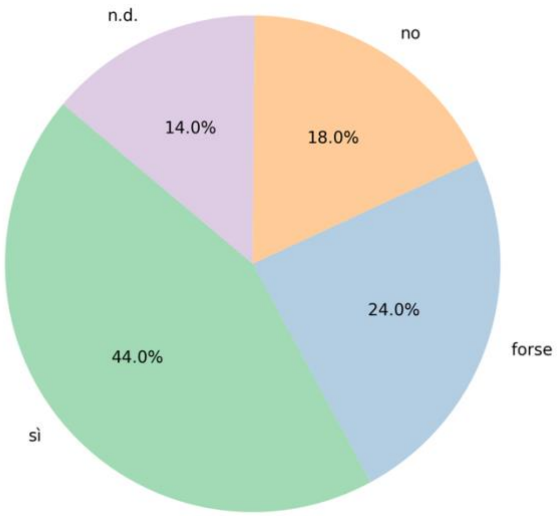


Figura 2.13. Distribuzione delle risposte riguardo la disponibilità effettiva a procedere ad un conferimento archivistico.

In contrasto con chi esclude la possibilità di un conferimento archivistico, Pecoraro adotta un approccio propositivo con il quale giustifica anche una sua certa tendenza all'accumulo: per lui tutto il materiale raccolto ha un valore, anche se non è immediatamente evidente. Ritrova, rilegge e riutilizza i suoi scritti, trovando spesso piacere nel rincontrare alcune pagine dopo anni: «[è] come leggere qualcosa scritto da un'altra persona, è interessante. È tutta memoria, che ci aiuta a comprendere la realtà». Non a caso, Pecoraro ha già effettuato un conferimento al progetto PAD del Centro Manoscritti e dimostra anche una spiccata sensibilità archivistica, riconoscendo l'importanza di integrarlo con la produzione più recente.

Anche altri autori hanno già intrapreso la strada di un conferimento istituzionale. Tra coloro che hanno donato il materiale cartaceo troviamo Capriolo e Piersanti, che hanno donato materiali al Centro Manoscritti, e Garavini, che ha affidato tutto il suo archivio cartaceo al Gabinetto Vieusseux. Sul fronte digitale, oltre che da Pecoraro, l'esempio è offerto da Pugno e Magrelli, che sempre al Centro Manoscritti hanno conferito copie di documentazione *born-digital* (Giagnolini e Baldini 2025; Milone 2025). Meldini, pur esprimendosi positivamente, lo fa solo "in astratto", esprimendo dubbi sulle capacità tecniche delle istituzioni nel gestire archivi digitali. Altri ancora stanno programmando o esplorando più concretamente ipotesi di donazioni, come nel caso di Di Stefano, che ha un accordo "non scritto" con la Fondazione Corriere della Sera, Spirito, che si rapporta con l'Archivio di Stato di Trieste, e Tarabbia, in relazione alla Fondazione Mondadori.

Un dato che spicca è che diversi autori, pur non escludendo la possibilità di un conferimento, considerano il proprio archivio come privo di interesse per altri se non loro stessi[83]. Per alcuni, l'idea stessa di conferire il proprio archivio a un'istituzione culturale è percepita come una manifestazione di superbia o, comunque, di una alta considerazione di sé che non hanno. Molti vedono in questa scelta un riconoscimento che non si sentono di assumere autonomamente[84]. Inoltre, per gli autori più giovani, la domanda stessa risulta prematura: Caminito, ad esempio, ritiene di essere semplicemente agli inizi della propria carriera di scrittura.

Le risposte di Balzano e Pellegrini evidenziano il tema della commistione tra materiali privati e lavorativi, un aspetto che da sempre caratterizza gli archivi di persona. Balzano si dice disposto a cedere il proprio archivio a un'istituzione, ma solo se in grado di separare adeguatamente ciò che ha scritto dalla sua vita privata, sottolineando che il confine tra pubblico e privato sia più facilmente oltrepassabile nel digitale. Pellegrini, avvocato, evidenzia in questo senso il contrasto contenutistico delle sue email, osservando che, pur essendo coperte dal diritto di riservatezza, alcune di esse contengono elementi letterari che meriterebbero di essere conservati.

Piersanti, in chiusura, riflette ironicamente: «So che ora a Pavia c'è anche una sezione informatica e credo sia ancora un ottimo centro. Ma se dovessi mandare oggi a Maria Corti le cose che ho all'interno del mio computer, non saprei neanche da dove iniziare, tale è grave e imperdonabile il disordine».

L'indagine sulla volontà d'archivio digitale non può prescindere, poi, dalla riflessione sulle prospettive testamentarie e sulle modalità di trasmissione dell'eredità digitale. I dati raccolti delineano un quadro ancora molto incerto: solo il 28% degli intervistati ha affidato le proprie credenziali a persone di fiducia, mentre il 72% non ha adottato alcuna misura in tal senso. Sulla possibilità di definire una qualche forma di testamento digitale il 16% afferma che lo farebbe, l'8% resta in dubbio, mentre il 26% si attesta più nettamente sul "no"[85] (Figura 2.14). In questo ultimo gruppo sono presenti gli autori che affermano di non avere password a protezione dei propri dispositivi, per cui non ritengono necessarie ulteriori disposizioni. Ma, soprattutto, è interessante notare l'emersione dell'idea che la perdita di materiale per i posteri non sia necessariamente un male[86]. Nelle parole di Spirito, per esempio, questo pensiero viene giustificato dalla percezione del medium: «[s]e nella carta la memoria si fa materia, nella dimensione

[83] Si vedano in merito, le considerazioni di Balzano, Caminito, Galletta, Giacopini, Matteucci, Terranova e Tuena.
[84] Si vedano, a vario titolo, le riflessioni di Balzano, Frizziero, Montanaro, Pecoraro, Romagnolo e Zaccuri.
[85] Per la metà delle risposte non è stato possibile rintracciare una posizione chiara a riguardo.
[86] Questo aspetto ricorre nelle risposte fornite da Montanaro, Montesano, Piersanti, Spirito e Tuena. Pellegrini fa una simile considerazione nella risposta dedicata alle strategie di backup.

telematica la memoria è più fluttuante, più fantasmatica ed è forse meno affascinante, anche se non meno importante».

Ma ci sono anche autori più sensibili al tema, che ne riconoscono la centralità e che sollecitano un intervento delle istituzioni per una regolamentazione chiara[87]. Con grande sensibilità archivistica, Spirito passerebbe per un'iniziativa testamentaria vincolata anche al cartaceo. Malaguti, nel delineare la complessità della questione, sottolinea come sia a maggior ragione importante avere la consapevolezza di «cosa lasciare a chi», sia per sé che per l'eventuale interesse che il materiale potrebbe un giorno avere.

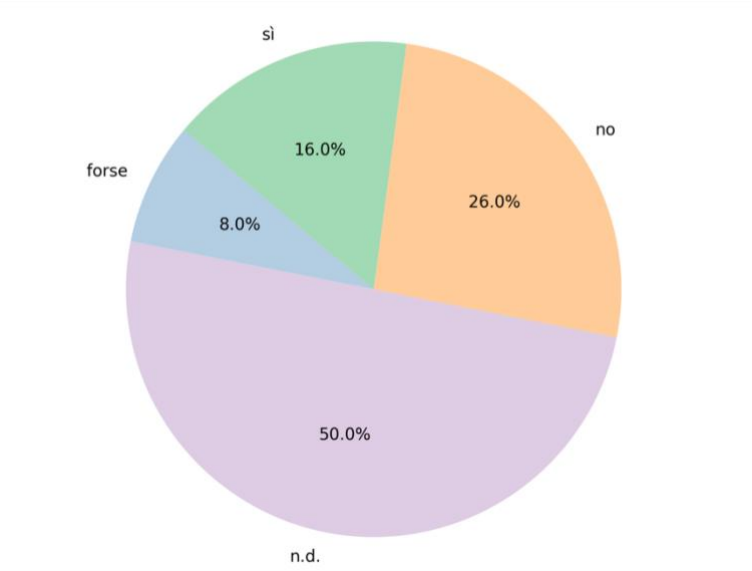


Figura 2.14. Distribuzione delle risposte sulla possibilità di adozione di un testamento digitale.

2.3.4 Nuove forme d'archivio e di scrittura

Il web ha facilitato la creazione e l'utilizzo di spazi di creatività alternativi rispetto a quelli tradizionali dell'epoca otto-novecentesca, dando origine a nuove pratiche di scrittura, lettura e interazione. Questo fenomeno è stato seguito con attenzione anche dagli editori, che hanno selezionato utenti/scrittori con un potenziale commerciale, ma hanno anche rintracciato in rete scritture con una connotazione letteraria distintiva (Piazza 2020). Questa evoluzione ha avuto origine dalle mailing list, è passata attraverso i blog e i siti web, fino a raggiungere, infine, i social media, determinando inediti sconfinamenti nella produzione documentale d'autore.

[87] L'appello emerge, in particolare, da Giacopini, Malaguti e Lupo.

L'indagine condotta evidenzia come il 40% degli autori possieda un sito web o un blog, il 16% lo abbia avuto in passato e il 44% non lo abbia mai utilizzato[88] (Figura 2.15). Analizzando i risultati per distribuzione anagrafica, emerge che nelle fasce centrali del campione si registra sia il maggiore utilizzo di questi strumenti sia il loro progressivo abbandono (Figura 2.16 e Figura 2.17).

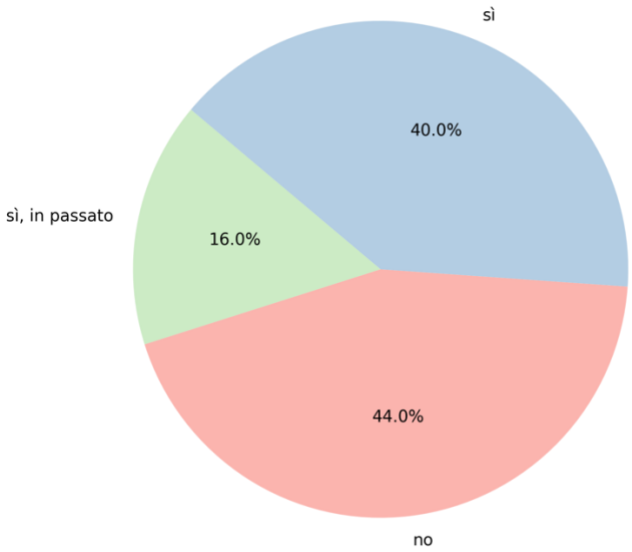


Figura 2.15. Distribuzione della presenza di blog o sito web d'autore.

[88] Tra i primi anni Duemila e l'inizio degli anni Dieci, blog e riviste online hanno costituito in Italia uno spazio autonomo di produzione e discussione letteraria, caratterizzato da forte interazione tra scrittura creativa, critica militante e dibattito politico-culturale. Esperienze come Nazione Indiana (https://www.nazioneindiana.com/), Carmilla Online (https://www.carmillaonline.com/) e Vibrisse (https://vibrisse.wordpress.com/), blog fondato da Giulio Mozzi, hanno configurato un sistema policentrico capace di incidere sulle dinamiche di visibilità e legittimazione nel campo letterario contemporaneo (Bortolotti 2008; Casadei 2015). Diversi degli autori intervistati nel presente lavoro hanno partecipato a queste comunità digitali attraverso interventi, collaborazioni redazionali o pubblicazioni di testi.

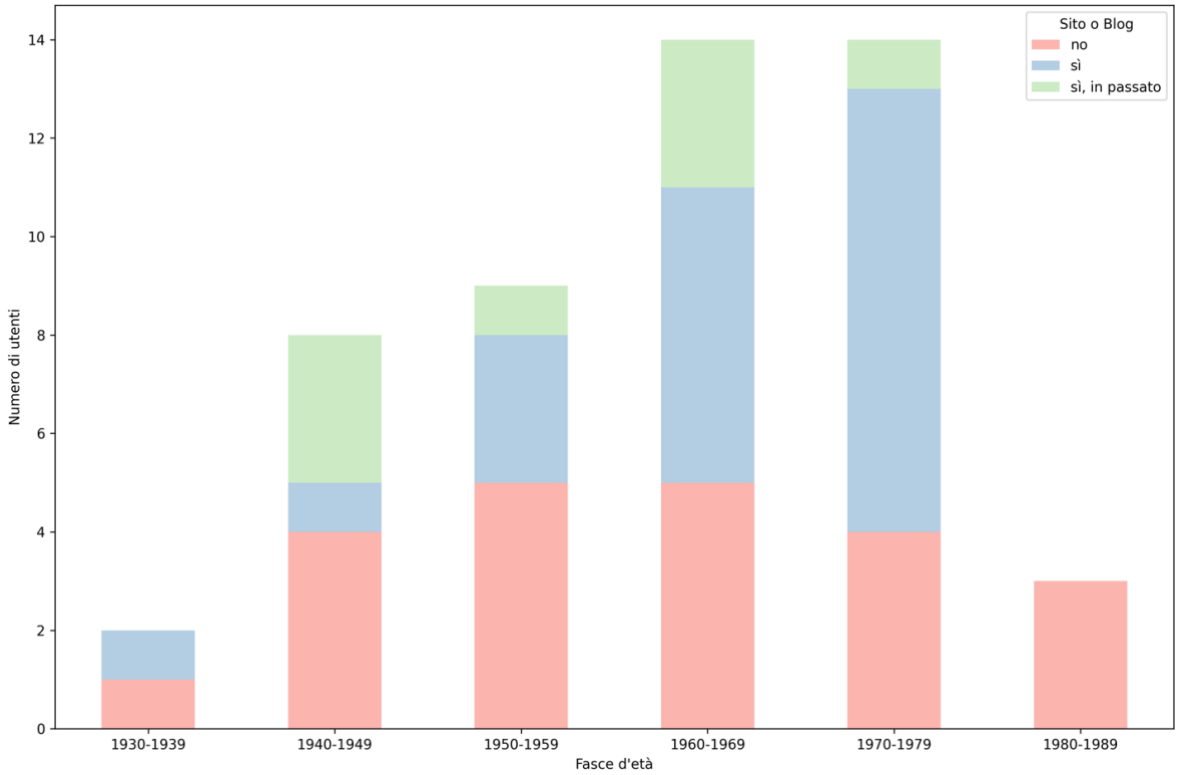


Figura 2.16. Distribuzione della presenza di blog o sito web d'autore in base alle fasce d'età.

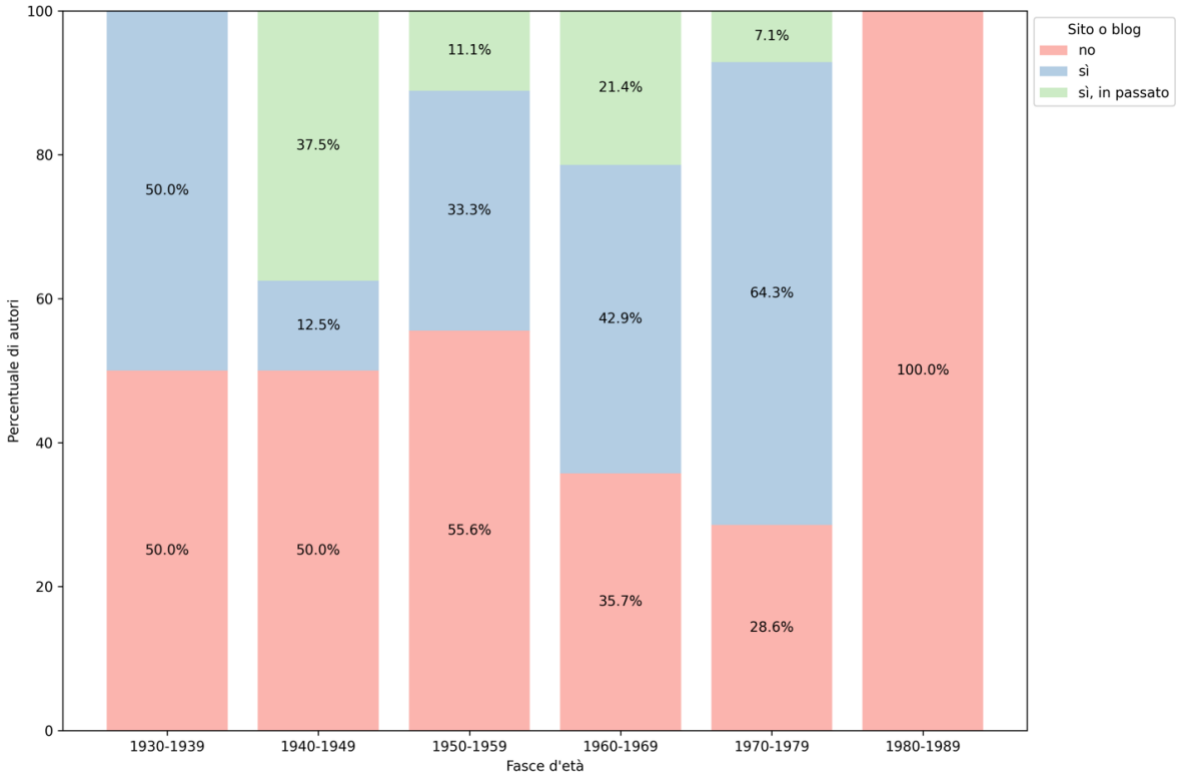


Figura 2.17. Distribuzione della presenza di blog o sito web d'autore in base alle fasce d'età, espresso in percentuale.

Tra coloro che possiedono un sito web, più della metà lo considera parte integrante del proprio archivio, mentre tre autori escludono questa declinazione. Alcuni precorritori, come Mario Biondi, ne sottolineano l'importanza documentale:

> L'ho creato personalmente nell'aprile del 1995 [...]. Costituisce una sorta di web archive di ciò che concerne la mia scrittura creativa dal punto di vista critico – recensioni, interviste, commenti – e conserva quasi tutta la mia attività giornalistica (centinaia di pezzi). Oltre alle foto. Personali e di viaggio.

Tuttavia, sette autori, pur dichiarando di avere un sito, non lo gestiscono direttamente: in questi casi, il sito esiste più per iniziativa dell'editore che per espressa volontà autoriale.

Per alcuni scrittori, il web e il contesto del blog hanno rappresentato l'origine stessa della loro attività letteraria, come per Pecoraro. La prima versione del suo blog, Tash-Blog, era ospitata sulla piattaforma *Splinder*, che nel novembre 2011 annunciò la chiusura con soli tre mesi di preavviso, avvenuta poi il 31 gennaio 2012[89]. Grazie all'intervento di Gino Roncaglia, fu possibile recuperare i contenuti con un'operazione di salvataggio durata quasi cento ore di download[90]. Analogamente, Mozzi ha salvato in extremis una copia del suo blog, inizialmente pubblicato su *Clarence*[91], che «è crollata all'improvviso, una notte».

In totale, solo sei autori dichiarano di avere delle copie di backup del proprio sito. Tra questi, figurano anche due degli otto scrittori che hanno deciso di chiuderlo. Fra le ragioni principali dell'abbandono, si riscontrano l'elevato impegno richiesto nella gestione di un blog e dell'emergere di nuovi strumenti di comunicazione. Zaccuri afferma: «Ad oggi, esiste ancora il dominio alessandrozaccuri.it, ma risulta "under construction" [...] In prospettiva non penso di riattivarlo. In ogni caso, se lo facessi, dovrei impostare un progetto per renderlo più sito e meno blog». Anche Romagnolo conferma questa tendenza: il suo blog è stato sostituito da un sito minimale, con collegamenti ai suoi profili social.

I risultati dell'indagine confermano che i blog e i siti web personali non sono scomparsi. Tuttavia, le testimonianze raccolte delineano un progressivo declino della loro popolarità e una trasformazione delle loro funzioni: da strumenti di scrittura e interazione a semplici vetrine di presentazione dell'autore. Questo cambiamento è strettamente legato all'ascesa dei social media, in particolare Facebook, che ha

[89] https://web.archive.org/web/20111124000003/http://www.splinder.com/.

[90] Almeno una copia di backup del blog è conservata nella partizione dell'archivio digitale di Pecoraro conferita al progetto PAD nel 2015, oggi parte integrante delle collezioni del Centro Manoscritti dell'Università di Pavia. Pecoraro ha successivamente trasferito il blog su WordPress. Il blog è ancora online ma è stato progressivamente abbandonato dall'autore.

[91] https://web.archive.org/web/19970616104801/http://www.clarence.com/whatis.shtml.

incorporato molte delle caratteristiche delle piattaforme precedenti, divenendo un ponte verso nuove modalità di scrittura e interazione[92]. Per questo motivo si è voluto indagare il loro utilizzo da parte degli autori, che per il 76% afferma di essere iscritto ad almeno un social media. Il 4% ha utilizzato social media in passato, mentre il 20% non ha mai avuto alcun profilo (Figura 2.18). Nella distribuzione dei risultati per fasce d'età, emerge che l'assenza dai social è più marcata nei nati tra il 1930 e il 1959 (Figura 2.19 e Figura 2.20).

Il social media più diffuso, fra gli autori intervistati, è Facebook, a cui sono iscritti ventotto autori, seguito da Instagram con ventidue iscritti, X con tredici, TikTok e LinkedIn con due utenti ciascuno (Figura 2.21).

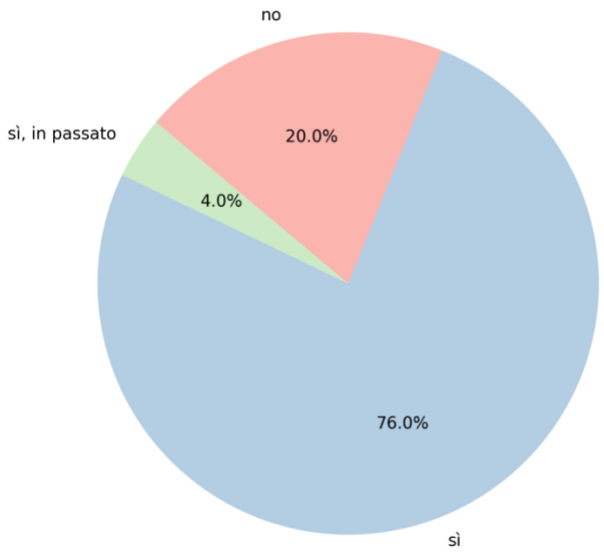


Figura 2.18. Distribuzione della presenza degli autori sui social media.

[92] Si segnala, a questo proposito, l'inchiesta di Andrea Lombardi sul rapporto tra gli scrittori e Facebook, condotta attraverso una serie di interviste pubblicate sul sito *Le Parole e le cose*. L'indagine è stata inaugurata da un'intervista a Francesco Pecoraro (A. Lombardi, *Scrittori e Facebook/1. Francesco Pecoraro*).

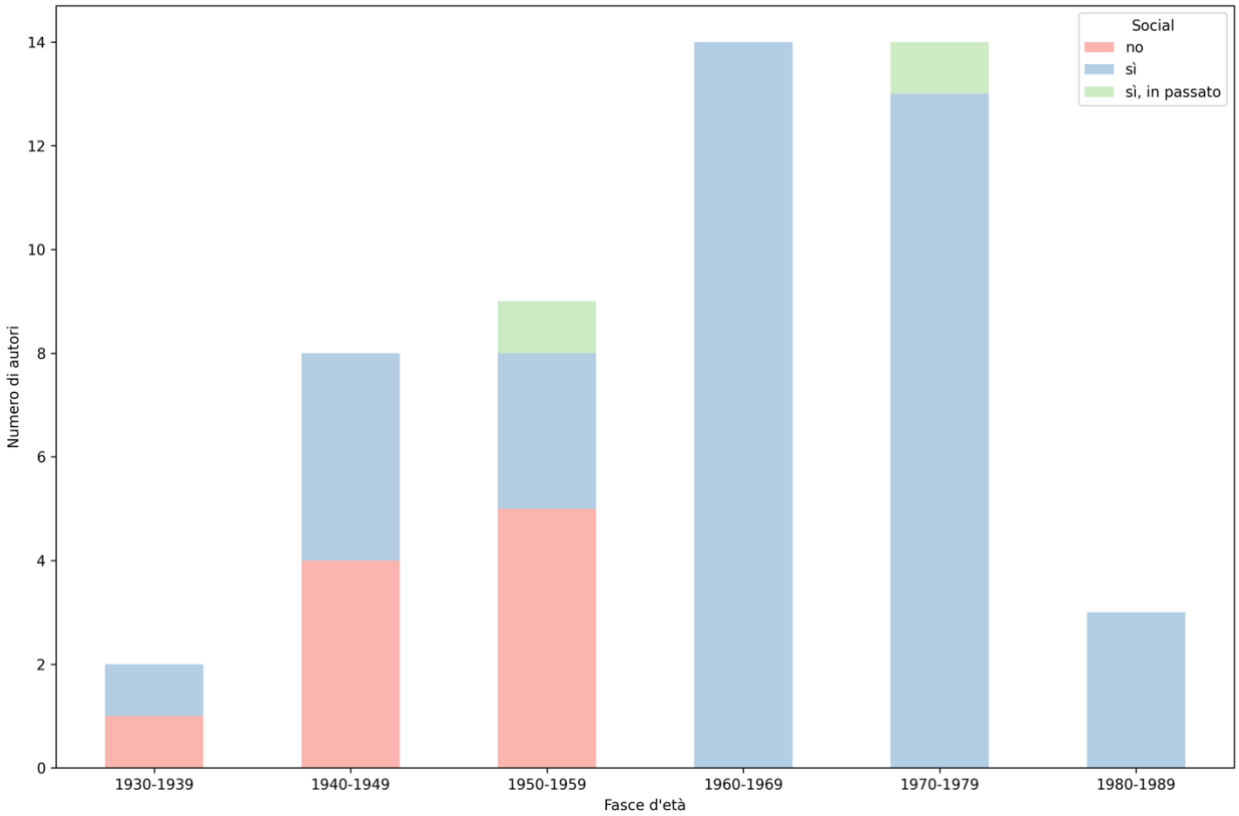


Figura 2.19. Distribuzione della presenza degli autori sui social media in base alle fasce d'età.

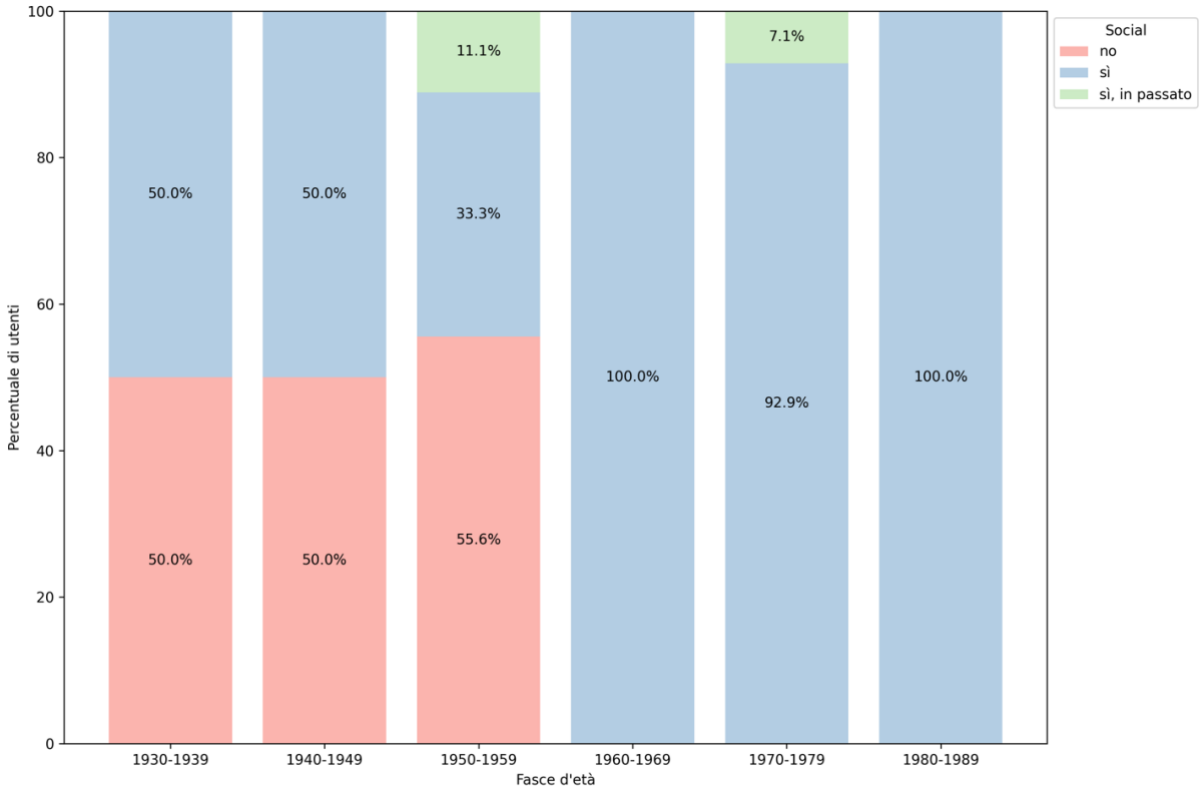


Figura 2.20. Distribuzione della presenza degli autori sui social media in base alle fasce d'età (espresse in percentuale).

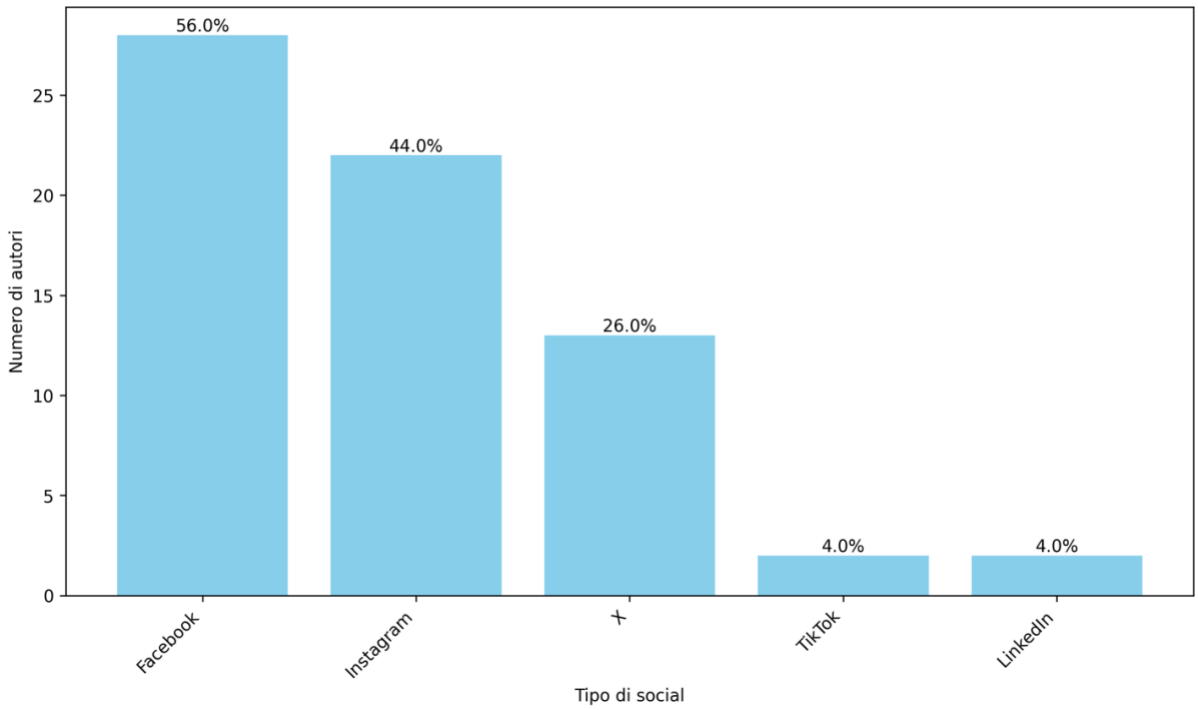


Figura 2.21. Distribuzione delle piattaforme social adottate dagli autori.

Come siti e blog, i social media possono essere considerati parte integrante degli archivi di persona, se intesi come spazi digitali in cui vengono creati, accumulati e conservati contenuti relativi alle attività personali e professionali (Cannelli e Musso 2022, 260). Dalle risposte degli autori emerge come dodici di loro non li ritengano parte del loro archivio. Altrettanti, invece, considerano i social come spazi di espressione capaci di conservare tracce della loro attività creativa. Pincio, ad esempio, riconosce un ruolo archivistico, quasi diaristico, ai social, ma al tempo stesso pratica un'eliminazione selettiva dei contenuti, come Mozzi e Tarabbia, che sfruttano la funzione "ricordi" di Facebook per rimuovere post passati. Al contrario, Galletta, pur utilizzando ampiamente i social, afferma di non considerarli parte del proprio archivio. Tuttavia, è tra i pochi autori ad aver intrapreso un'attività sistematica di archiviazione del materiale pubblicato:

> Un paio di anni fa avevo deciso di raccogliere tutto ciò che avevo scritto su Facebook, che utilizzo intensamente dal 2008. Ogni giorno consultavo la funzione "ricordi", selezionavo i contenuti, li trascrivevo in un file e ne modificavo la privacy impostandola su "solo io", senza cancellare nulla ma rendendolo invisibile agli altri. Dopo circa sette mesi ho interrotto il progetto, avendo già raccolto un milione e trecentomila battute: una quantità incredibile. Non ho mai usato la funzione di Facebook per scaricare una copia dei miei dati, poiché desideravo organizzarli con un ordine preciso.

Anche Tarabbia, pur dichiarando di cancellare con "gioia" i propri post, riconosce che i contenuti ritenuti di valore abbiano sempre trovato una collocazione in altri luoghi. Emerge, quindi, una dinamica complessa, in cui la natura effimera dei social si intreccia con il desiderio di permanenza della scrittura, portando gli autori a sviluppare pratiche individuali di gestione e preservazione dei propri testi.

Nel riscontro di Pecoraro è peculiare la differenza di percezione (e funzione) tra il suo blog, accuratamente recuperato dal web e consegnato a PAD, e i contenuti social, di cui l'autore ignora serenamente le sorti, preferendo ascriverli al ruolo di comunicazione quotidiana.

Un sentimento di frustrazione comune emerge riguardo alla trasformazione degli spazi online, da nicchie di discussione approfondita a piattaforme di massa che hanno contribuito a rendere i dibattiti più superficiali e sterili. Questo fenomeno ha portato molti scrittori a ridurre o abbandonare l'uso delle piattaforme. Autrici come Ranieri e Terranova spiegano la loro scelta di evitare i social per motivi di "igiene personale", non desiderando essere coinvolte nelle polemiche che talvolta accompagnano i loro articoli[93]. Spirito, pur utilizzando i social per scopi professionali, li considera infatti dannosi per il lavoro creativo, ritenendo che la scrittura necessiti di una dimensione distinta, lontana dall'immediatezza e dalla visibilità dei social, che rischiano di banalizzare il lavoro dello scrittore. Tuttavia, autori come Zeno, Balzano e Lupo continuano a vedere nei social un'opportunità positiva per un'interazione diretta con i lettori, riconoscendo la possibilità di stabilire un legame immediato e personale con il pubblico.

L'ultima domanda è stata lasciata aperta, offrendo spunti per rintracciare ulteriori e specifiche influenze del digitale. Il 40% degli autori ha riconosciuto l'importanza del web per il proprio processo creativo e sottolinea come browser e motori di ricerca abbiano incrementato sia la velocità che la quantità di informazioni disponibili. Come evidenzia Milone, la rapidità non si limita ai pochi istanti necessari per ottenere i risultati di una ricerca, ma comprende anche la capacità di reperire in tempi brevi informazioni specifiche all'interno delle risorse digitali (2025, 176), rendendole prêt-à-porter. Questo ha rappresentato una rivoluzione per gli scrittori che erano abituati a trascorrere ore in archivio e in biblioteca a cercare informazioni che oggi sono accessibili con un click.

Sembra emergere, in realtà, il Web non abbia trasformato radicalmente i processi, limitandosi a velocizzarli. Occorre però tenere presente un ulteriore aspetto che non è stato considerato dagli autori: l'enorme volume di risultati viene presentato con un ordine definito dagli algoritmi di *ranking*, che propongono le informazioni agli autori in un ordine di cui la logica rimane opaca rispetto ad una ricerca analogica (Milone 2025, 177). Per quanto ne riconoscano l'importanza, poi, per le generazioni più

[93] Si veda, a riguardo, anche la risposta di Pecoraro all'ultima domanda.

giovani l'impatto del Web non è stato così radicale. Come osserva Grossi: «Internet è interessante, ma mentre me lo domandava mi rendevo conto che anche questo non ha cambiato molto me, semplicemente perché io appartengo a una generazione che quella forma di digitalità l'ha sempre vissuta».

Un altro aspetto intercettato da più risposte riguarda l'impatto dell'editing digitale sulla scrittura e sulla produzione documentale. Questo fenomeno è specialmente apprezzato dalle generazioni che hanno vissuto la laboriosità della scrittura e della revisione manuale. D'altronde, la possibilità di modificare un testo in modo agile ed efficiente fu il fattore determinante nel persuadere molti autori ad adottare il computer come strumento di scrittura, come testimoniano le parole di Masolino d'Amico nell'intervista di Curino. In alcune esperienze, questo aspetto è andato oltre la mera facilitazione incidendo sullo stile, come per Calandrone: «Nel digitale basta un gesto e non c'è più niente: questo aiuta molto l'essenzialità. In questo senso la mia scrittura è cambiata, è molto meno barocca, più chiara. Forse c'entra anche l'invecchiamento, ma sicuramente c'entra molto anche il mezzo».

La facilità di editing implica altresì la possibilità di una revisione infinita, un aspetto percepito in modi differenti dagli autori. Alcuni vedono in questa caratteristica al contempo un potenziale e un vincolo, che può intrappolare la scrittura in un ciclo interminabile di perfezionamenti [94]. Per altri rimane un'opportunità straordinaria, cristallizzabile con le parole di Umberto Eco: «Il word processor, paradossalmente, potrebbe diventare uno strumento flaubertiano, che invoglia a cercare instancabilmente il *mot juste*» (Ferretti 1985, 50).

2.4 Discussione

L'analisi mette in luce la centralità del digitale negli archivi d'autore contemporanei, senza tuttavia sancire la definitiva scomparsa del cartaceo. La digitalizzazione del processo creativo emerge come tendenza predominante, ma il ruolo della carta persiste ancora in alcune nicchie, per alcuni autori in termini funzionali, per altri semplicemente emotivi. La persistenza dell'analogico non sembra essere il riflesso di un'abitudine generazionale, ma rispondere piuttosto a esigenze materiali e cognitive che il digitale non ha completamente soppiantato.

Dal punto di vista archivistico, il panorama appare estremamente eterogeneo. La gestione dei file segue criteri diversificati: mentre la maggior parte degli autori adotta sistemi di denominazione e archiviazione strutturati, una quota importante si affida ancora a pratiche più occasionali, con un conseguente aumento del rischio di dispersione documentaria. Fenomeni ampiamente diffusi, come l'auto-invio di email o

[94] Si veda, ad esempio, la risposta fornita da Pellegrini.

l'uso simultaneo di più dispositivi, contribuiscono a creare una ridondanza informativa e una stratificazione documentale particolarmente complessa. In questo contesto, gioca un ruolo cruciale anche la presenza online degli autori: la transizione dai siti web e blog ai social network, seguita ora da un primo segnale di disaffezione verso questi ultimi, rappresenta un'ulteriore testimonianza della rapidità e della fluidità con cui evolvono le forme documentali.

La consapevolezza della fragilità del digitale emerge nettamente, manifestandosi in molteplici forme e in risposta a diverse sollecitazioni. Questo dato si riflette soprattutto nelle pratiche di backup, che appaiono ormai consolidate, per quanto altamente variabili, in linea con quanto riportato da Becker e Nouges (2012). Si conferma anche la dimensione emozionale come fattore determinante nelle strategie archivistiche (Copeland 2011; Krtalić e Dinneen 2024), spaziando dal valore affettivo attribuito alla carta, alla sensazione di "liberazione" offerta dalla scrittura digitale, fino alla paura della perdita dei dati.

Uno dei nodi più problematici emersi è la trasmissione del patrimonio documentale digitale. Le risposte degli autori mostrano una varietà di posizioni che rende il concetto di "volontà d'archivio" molto frammentato. La maggior parte appare perplessa di fronte alla possibilità di predisporre accortezze per garantire l'accessibilità dei propri materiali, come il semplice lasciare le credenziali a una persona cara, soprattutto in assenza di chiare regolamentazioni in materia. Un quadro più roseo affiora in merito alla possibilità di conferire il proprio archivio a istituzioni culturali, anche se traspare una mancata percezione del valore documentario del proprio archivio, in linea con il quadro fornito da Micunovic, Marčetić e Krtalić (2016, 13). Questo dato potrebbe essere letto alla luce del processo di "desacralizzazione" dello scrittore (Simonetti 2023, 30) o, nel caso più specifico di questo dataset, dei meccanismi di selezione dei premi, che intercettano autori con carriere non sempre destinate a un effettivo riconoscimento. Concentrarsi sui finalisti di Strega e Campiello, infatti, esclude una parte importante della produzione letteraria contemporanea, in particolare quella legata a generi e forme narrative meno canoniche, come la narrativa di genere o la letteratura elettronica. Quest'ultima, fra l'altro, è un ambito di grande interesse per il contesto *born-digital*[95] (Bolter e Joyce 1987; Bolter 2001; Grigar e O'Sullivan 2021; Iadevaia 2022). Occorre anche segnalare che nella composizione anagrafica del dataset, prevalentemente superiore ai quarant'anni anni, sono sottorappresentate le fasce d'età più avanzate e quelle più giovani. Per future ricerche, dunque sarebbe auspicabile un ampliamento del campione per includere una maggiore diversità, sia in termini di numero che di profilo degli autori. Ciò consentirebbe di esplorare

[95] Sulla letteratura elettronica e le problematiche di conservazione si vedano le iniziative dell'Electronic Literature Organization (ELO), in particolare il programma PAD (https://eliterature.org/programs/pad/), e la ELO Collection, (https://collection.eliterature.org).

una gamma più ampia di approcci e strategie creative, riflettendo in modo più completo le complessità del panorama autoriale contemporaneo.

Le pratiche archivistiche degli autori contemporanei offrono preziose indicazioni per le istituzioni chiamate a preservare il patrimonio letterario. Comprendere i processi di creazione e le strategie adottate dagli autori consente di orientare le politiche di conservazione verso la ricezione e il recupero del digitale d'autore. In un panorama ancora poco documentato, gettare luce sul *modus operandi* degli autori e sulle tendenze coeve permette di formulare ipotesi più strutturate sulla presenza o sull'assenza di specifici documenti e metadati, recuperando informazioni di contesto e anticipando le sfide della preservazione.

Il ruolo degli autori si dimostra centrale per la futura conservazione dei loro archivi, soprattutto alla luce degli ostacoli informatici e legali che caratterizzano il digitale. Se i margini di intervento diretto delle istituzioni culturali risultano limitati in questa prospettiva, conservano comunque uno spazio d'azione cruciale: incentivare la collaborazione tra autori, eredi e archivi, promuovendo una sensibilizzazione diffusa sull'importanza del digitale d'autore e sulle sue specificità documentarie. Ma il compito delle istituzioni non si esaurisce in questo: è necessaria una conoscenza multidisciplinare che permetta di preservare, ma anche di rendere accessibili e intelligibili questi documenti. Accesso che non guardi solo al contenuto ma alla comprensione dei contesti digitali *tout court*. Solo attraverso un approccio di questo tipo sarà possibile garantire la trasmissione e la fruizione del digitale d'autore, lasciando agli autori il diritto «di scrivere sulle nuvole» (Pozzoli 1986, 23), anche quando la nuvola è un cloud.

3. L'Archivio Valerio Evangelisti

Date le considerazioni teoriche e il quadro fenomenologico emerso dalle interviste, questo capitolo completa la risposta alla RQ1 attraverso l'analisi di un caso di studio concreto: l'Archivio Valerio Evangelisti (1952-2022), un archivio ibrido con una componente nativa digitale particolarmente corposa. Il capitolo presenta il soggetto produttore nel suo contesto biografico e intellettuale (cap. 3.1), descrive il fondo nelle sue componenti analogica e digitale (cap. 3.2) e ne ricostruisce la storia archivistica (cap. 3.3). Infine, l'analisi del rapporto di Evangelisti con la tecnologia (cap. 3.4) offre un punto di osservazione privilegiato sui materiali: la sua consapevolezza delle dinamiche tecnologiche e documentarie permette di comprendere con maggiore profondità le pratiche di produzione, gestione e conservazione che hanno dato forma all'archivio.

3.1 Nota biografica

Valerio Evangelisti (Bologna, 20 giugno 1952 - Bologna, 18 aprile 2022) è stato scrittore, storico e militante politico. Nato a Bologna il 20 giugno 1952, frequenta il liceo classico Malpighi e si iscrive successivamente al Corso di Laurea in Scienze Politiche, indirizzo storico-politico, presso l'Università di Bologna, conseguendo la laurea nel 1976 (Università di Bologna 1976). Politicamente attivo sin dagli anni del liceo, durante l'Università fa parte della militanza studentesca di sinistra ed è fra i fondatori del Circolo Carlos Fonseca.

Poco dopo la laurea risulta vincitore di un concorso pubblico e accede alla Scuola Superiore di Pubblica Amministrazione; dal 1981 viene nominato funzionario direttivo del Ministero delle Finanze, operando inizialmente presso l'Intendenza di Finanza di Bologna e successivamente nella Direzione Compartimentale per Emilia-Romagna e Marche. Parallelamente, svolge attività didattica e ricerca accademica in ambito storico, contribuendo con articoli e saggi a riviste quali "Il Mulino", "Rivista di storia contemporanea", "Quaderni Emiliani", "Quaderni sardi di storia", e pubblicando cinque volumi: *Storia del partito socialista rivoluzionario*, *1881-1893* (con Emanuela Zucchini), pubblicato nel 1981 da Cappelli e frutto delle tesi di laurea degli autori (ristampato da Odoya nel 2013); *Il galletto rosso. Precariato e conflitto di classe in Emilia-Romagna, 1880-1980* (con Salvatore Sechi), Marsilio, 1982 (ristampato da Odoya, con la sola parte di Evangelisti, nel 2015, col titolo *Il gallo rosso*); *Sinistre eretiche. Dalla Banda Bonnot al sandinismo*, *1905-1984*, SugarCo, 1985; *Gallerie nel presente. Punks, Snuffs, Contras: tre studi di storia simultanea*, Lacaita, 1988; *Gli sbirri alla lanterna. La plebe giacobina*

bolognese dall'anno I all'anno V (1792-1797), Bold Machine, 1991, ristampato da Derive Approdi nel 2005.

Fra la fine degli anni Ottanta e l'inizio degli anni Novanta inizia a sperimentare più assiduamente con la narrativa. Nel 1993 vince il Premio Urania con il romanzo *Nicolas Eymerich, inquisitore*, pubblicato poi nel 1994 nella collana "Urania" di Mondadori, dando inizio a uno dei suoi cicli più noti dedicati all'inquisitore medievale catalano Nicolas Eymerich (1320-1399), che prosegue con altri undici titoli fino al 2018.

Nel 1995, *Il mistero dell'inquisitore Eymerich* compare in dieci puntate su *Il Venerdì di Repubblica*, per poi essere raccolto in volume. A cavallo fra gli anni Novanta e i Duemila, Evangelisti ricava tre sceneggiati radiofonici di trenta puntate ciascuno a partire dai suoi romanzi, andati in onda su Radio Rai 2 tra il 1999 e il 2001: *La scala per l'inferno*, *Il castello di Eymerich* (vincitore del Prix Italia per la migliore sceneggiatura) e La *furia di Eymerich*.

Nel 1999 pubblica in tre volumi *Magus. Il romanzo di Nostradamus*, che supera le 100.000 copie vendute, consentendogli di dedicarsi esclusivamente alla scrittura. Tra le sue opere successive si distinguono i romanzi del "Ciclo del Metallo" (*Metallo urlante*, 1998; *Black Flag*, 2002), omaggio alla musica heavy metal e a *Métal Hurlant*, storica rivista francese dedicata al fumetto e alla fantascienza. Seguiranno, alternando le pubblicazioni, il "Ciclo Americano", ispirato alle lotte sindacali statunitensi (*Antracite*, 2003; *Noi saremo tutto*, 2004; *One Big Union*, 2009), il "Ciclo Messicano", dedicato alla rivoluzione messicana (*Il collare di fuoco*, 2005; *Il collare spezzato*, 2006), la "Trilogia dei Pirati", sui filibustieri nei Caraibi della fine del Seicento (*Tortuga*, 2008; *Veracruz*, 2009; *Cartagena*, 2012) e la trilogia de *Il Sole dell'Avvenire* (2013-2016) in cui si narra il passaggio epocale dal mondo rurale all'agricoltura di tipo industriale.

Con Antonio Moresco firma *Controinsurrezioni* (2008). Tra i saggi critici, pubblica *Alla periferia di Alphaville* (2001), *Sotto gli occhi di tutti* (2004), *Distruggere Alphaville* (2006), e l'autobiografico *Day Hospital* (2013), in seguito a una malattia che lo colpisce nel 2009.

I suoi romanzi sono stati tradotti in vari Paesi, tra cui Francia, Spagna, Germania, Portogallo, Russia, Israele, Brasile, Repubblica Ceca, Slovacchia, Olanda (Chianese 2005, 142). Tra i riconoscimenti ricevuti: Premio Urania (1993), Grand Prix de l'Imaginaire (1998), Prix Tour Eiffel (1999), Prix Italia (2000), Premio Italia (2007), Premio Europa (Eurocon Praga 2002).

Negli anni, Evangelisti ha preso parte a progetti di critica letteraria e di controinformazione, assumendo ruoli di primo piano nella direzione editoriale di diverse riviste. Per dieci anni è stato direttore della

rivista *Progetto Memoria* e, successivamente, di *Carmilla*: nata in versione cartacea nel 1995, con la pubblicazione di quattro numeri, la rivista si è trasformata nel 2003 in una testata online, divenendo un importante punto di riferimento per il movimento del *New Italian Epic*.

Nel 2004 organizza una raccolta di firme in solidarietà a Cesare Battisti, raccogliendo oltre 1.500 adesioni (carmillaonline 2004). Si candida come indipendente alle elezioni europee del 2009 nella *Lista Anticapitalista* (coalizione della Federazione della Sinistra, Rifondazione Comunista-PdCI) (Corriere della Sera 2009). Nel 2011 si presenta alle elezioni comunali di Bologna con la lista *Sinistra per Bologna - Federazione della Sinistra*, nel 2018 dichiara sostegno a *Potere al Popolo!*, mentre nel 2021 ne è capolista alle amministrative di Bologna.

Muore a Bologna il 18 aprile 2022 a 69 anni (carmillaonline 2022).

3.2 L'archivio ibrido

L'Archivio Valerio Evangelisti offre una testimonianza completa e articolata della vita intellettuale, politica e creativa dell'autore, documentando oltre quattro decenni di attività attraverso una ricca stratificazione di materiali cartacei e digitali, configurandosi come un "archivio ibrido" (Allegrezza e Gorgolini 2016).

La partizione cartacea dell'archivio copre prevalentemente i primi anni di attività di Evangelisti, non solo come scrittore, ma anche in qualità di ricercatore di storia moderna e contemporanea e come attivista e militante politico. L'analogico comprende cinque buste conservate dall'Archivio Marco Pezzi presso l'Istituto Parri di Bologna che contengono principalmente saggi e articoli scritti da Evangelisti durante il suo percorso accademico, ma anche rassegna stampa, volantini e materiale propagandistico della sinistra studentesca bolognese degli anni Settanta e Ottanta e della sottocultura punk, per un totale di 49 unità archivistiche. Completano la sezione cartacea sedici quaderni con schemi, appunti e bozze dei suoi primi romanzi e 37 album fotografici. Complessivamente, la partizione cartacea copre un arco cronologico che va dal 1969 al 2010 e presenta una notevole varietà tipologica che riflette la poliedrica attività intellettuale del produttore: da militante politico, a storico specializzato in storia sociale dell'Emilia-Romagna, fino alla sua affermazione come scrittore.

L'archivio cartaceo documenta la formazione politica e intellettuale dell'autore durante i primi anni di carriera accademica, attraverso una stratificazione di materiali che ne tracciano l'evoluzione da studente a giovane ricercatore presso l'Università di Bologna. La documentazione del periodo 1987-1990 include curriculum vitae, elenchi delle pubblicazioni, tessere associative e biglietti da visita, insieme alla

corrispondenza con editori (Marsilio, Teti, Fanucci, Tunuè) e istituzioni culturali. Particolare interesse rivestono gli scritti giovanili del 1969, con componimenti poetici, appunti di lettura e annotazioni critiche.

A questa documentazione si affianca la produzione accademica più matura, costituita dai materiali preparatori, dalle bozze e dai dattiloscritti dei suoi principali saggi storici, nonché dagli estratti delle sue pubblicazioni su riviste specializzate quali "Il Mulino", "Rivista di storia contemporanea" e "Quaderni Emiliani", che attestano il riconoscimento scientifico raggiunto negli studi di storia sociale dell'Emilia-Romagna.

Di particolare interesse storiografico sono i materiali relativi all'attivismo politico della sinistra bolognese e ai movimenti di solidarietà internazionale degli anni Settanta e Ottanta. In merito, il fondo conserva: volantini e manifesti della mobilitazione universitaria bolognese (1975-1976); documenti del Circolo Carlos Fonseca (1985-1986); materiali sulla rivoluzione sandinista e i conflitti centroamericani (1982-1997); periodici di movimento: "Quotidiano dei lavoratori" (1974-1977), "Tricontinental" (1984), "Democratic Palestine" (1988-1991).

L'interesse di Evangelisti per i fenomeni di cultura underground e la musica punk (Girolami 2011) è documentato da una raccolta di tredici fanzines punk, sia italiane che internazionali, datate 1981-1982, tra cui "Milano Punk. Situazione musicale", "Sabotage. Nihilist fanzine", "Rising Free Fanzine" e "ROCKGARAGE", che documentano la nascita e lo sviluppo del movimento punk in Italia e i suoi collegamenti con le scene europee. Completano questa documentazione alcuni dattiloscritti dello stesso Evangelisti, tra cui il saggio "Punk Rockers" che analizza gli aspetti politici del movimento, e materiali relativi al cinema indipendente, inclusi documenti sul film "Contras" scritto e realizzato dall'autore, che attestano il suo coinvolgimento diretto nella produzione culturale alternativa.

Il fondo conserva anche un'articolata attività di autodocumentazione che copre oltre due decenni (1980-2002) e riflette tanto gli interessi intellettuali quanto la ricezione critica dell'opera di Evangelisti. Si tratta, in particolare, di ritagli di stampa, organizzati per tematiche politiche, letterarie e culturali, che costituiscono una preziosa fonte per ricostruire il dibattito pubblico sui temi cari all'autore, dalla storia sociale alle questioni geopolitiche internazionali. Particolarmente rilevante è la rassegna stampa dedicata alla ricezione critica dei suoi primi libri, che documenta l'evoluzione della sua carriera letteraria in ambito nazionale e internazionale, includendo interviste, recensioni e articoli che testimoniano il progressivo riconoscimento del suo contributo alla letteratura italiana e la diffusione delle sue opere all'estero.

Sempre sul piano analogico, è da segnalarsi anche un consistente fondo librario, composto di circa 6.000 volumi, acquisito nel 2023 dal Dipartimento di Filologia Classica e Italianistica dell'Università di Bologna. La catalogazione del patrimonio librario è attualmente in fase di avvio, ma una prima ricognizione evidenzia diverse sezioni tematiche in stretto dialogo con il materiale documentario. In particolare, il fondo include le opere complete di Evangelisti nelle varie edizioni, una ricca raccolta di narrativa fantascientifica (con una consistente presenza della collana Urania), volumi di storia specialistica (soprattutto sulla storia dell'Inquisizione e sui periodi storici trattati nei suoi romanzi), testi di psicologia e una raccolta di riviste informatiche degli anni Novanta-Duemila. Completano la partizione analogica tre bobine Super 8 contenenti delle sperimentazioni di cinematografia amatoriale di fine anni Settanta inizio anni Ottanta di Evangelisti e suoi compagni dei tempi dell'università.

La componente più consistente dell'archivio, tuttavia, è rappresentata dai materiali nativi digitali, che documentano oltre tre decenni di attività di Evangelisti. Pioniere nell'uso del computer già dagli anni Ottanta, l'autore ha sperimentato costantemente con le nuove tecnologie, integrandole nel suo processo creativo e nella sua pratica di scrittura (Maugeri 2011). Grazie al lavoro dell'Associazione Valerio Evangelisti - Il Sol dell'Avvenire è stato possibile recuperare una parte importante delle sue tracce digitali: 389 floppy disk e tre hard disk (del computer fisso, del computer portatile e una memoria esterna) per un totale di circa 2,1 TB di materiali digitali (Figura 3.1)[96].

La distribuzione dei dati digitali evidenzia le diverse capacità e utilizzi dei supporti utilizzati da Evangelisti. La copia dell'hard disk del computer fisso, identificabile come la principale postazione di lavoro, conserva 65.448 elementi per un totale di 674,36 GB. La cronologia dei documenti conservati copre un arco temporale che si estende dal 1995 fino al 2022[97], a testimonianza di un utilizzo continuativo e stratificato del supporto.

[96] Per rappresentare la collezione di floppy disk nella Figura 3.1 sono stati presi in considerazione i 93 supporti di cui è stato effettuato il riversamento (4,85 GB). Si ipotizza tuttavia che il peso complessivo della collezione completa di floppy disk potrebbe raggiungere circa il doppio di quello attualmente riversato.

[97] L'ultimo documento riferibile all'attività di Evangelisti è datato, in termini di creazione, al 16 aprile 2022. Esistono però sei documenti successivi (datati fino a marzo 2024) verosimilmente dovuti all'utilizzo del computer da parte degli eredi o dell'Associazione Evangelisti per la gestione dell'archivio.

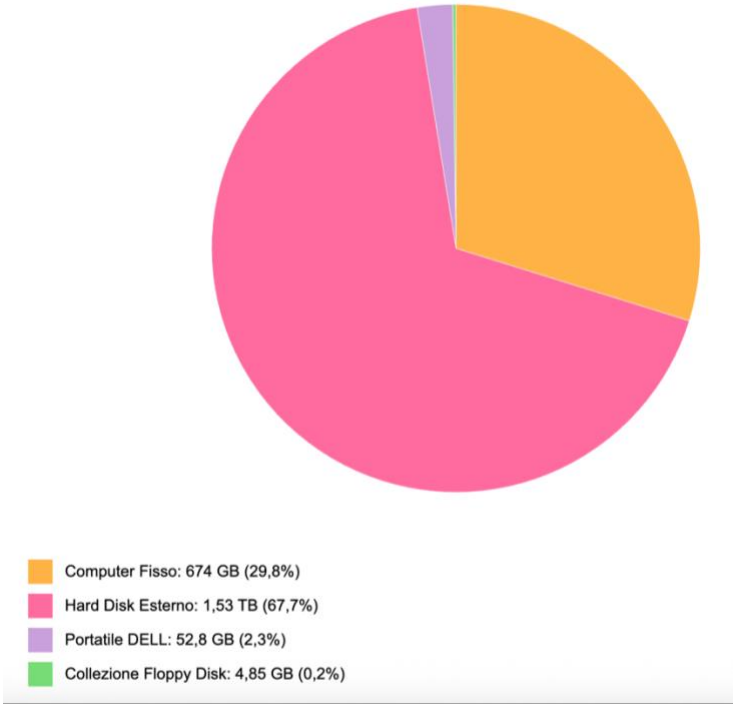


Figura 3.1. Distribuzione dei materiali digitali dell'Archivio Valerio Evangelisti per supporto di memorizzazione.

Dal punto di vista contenutistico, la copia dell'hard disk del computer principale documenta sia l'attività di scrittore sia la gestione personale e quotidiana dei materiali. Sono infatti presenti:

- cartelle di lavoro dedicate alla scrittura narrativa, organizzate generalmente per romanzi e ulteriormente suddivise in capitoli o sezioni;
- raccolte miscellanee di testi di Evangelisti ed altrui;
- raccolte di copertine delle edizioni nazionali e internazionali delle opere di Evangelisti;
- materiali di studio principalmente raccolti in diverse cartelle denominate *Biblioteca*, ma presenti anche fra le cartelle *Documenti* e *Download* e nelle cartelle di programmi *e-book reader*. Particolarmente corposa anche la sezione dedicata alle digitalizzazioni ad alta definizione di periodici socialisti a cavallo tra XIX e XX secolo, fornite da varie biblioteche dell'Emilia-Romagna e utilizzate, in particolare, per la redazione del ciclo "Il Sol dell'Avvenire";
- collezioni musicali, video, testi e fotografie personali, che restituiscono un quadro degli interessi culturali e privati;
- backup di sistema e degli account di posta elettronica;

- un insieme consistente di applicativi e file di sistema, necessari per il corretto funzionamento di computer e applicazioni, che costituiscono una parte rilevante della massa complessiva dei dati.

Nel suo insieme, questo segmento del fondo digitale si configura come la testimonianza più ricca e complessa del lavoro di Evangelisti, sia sul versante creativo e documentario, sia su quello gestionale e tecnico.

L'hard disk esterno rappresenta il supporto con la maggiore capacità complessiva: 1,53 TB distribuiti in 19.817 elementi. Il supporto è identificato come Western Digital WDBACW0020HBK-01, My Book 2TB USB 3.0 Series External Hard Drive. A differenza dell'hard drive del computer fisso, la struttura delle cartelle risulta più lineare e meno ramificata. Le caratteristiche complessive suggeriscono che questo dispositivo fosse impiegato prevalentemente come unità di backup del computer principale, con funzione di copia selettiva dei dati. Risultano infatti assenti la gran parte dei file generati dal sistema operativo e dalle applicazioni (abbondantemente documentati nell'hard disk del computer fisso), mentre sono conservati soprattutto documenti e raccolte probabilmente ritenute di maggiore rilevanza. Tale selezione rende l'insieme dei dati più compatto e mirato, evidenziando un uso intenzionale del supporto come archivio di sicurezza.

I materiali riversati dai floppy disk costituiscono la porzione più antica dell'archivio digitale con 2.214 elementi per un totale di 4,85 GB. Si tratta principalmente di materiali provenienti da 93 floppy disk 3,5" relativi ai primi romanzi di Evangelisti[98], ma anche alla redazione dei primi numeri di Carmilla, backup di messaggi BBS (Bulletin Board System)[99], appunti e materiali personali. Oltre ai floppy riversati su

[98] Le opere di cui si conservano materiali e stesure nei floppy disk sono: *Nicolas Eymerich, inquisitore* (Mondadori, 1994), *Le catene di Eymerich* (Mondadori, 1995), *Il corpo e il sangue di Eymerich* (Mondadori, 1996), *Il mistero dell'inquisitore Eymerich* (Mondadori, 1996), *Cherudek* (Mondadori, 1997), *Picatrix. La scala per l'inferno* (Mondadori, 1998), *Il castello di Eymerich* (Mondadori, 2001), *La furia di Eymerich* (Mondadori, 2003), *Magus. Il romanzo di Nostradamus. Vol. 1: Il presagio* (Mondadori, 1999), *Magus. Il romanzo di Nostradamus. Vol. 2: L'inganno* (Mondadori, 1999), *Magus. Il romanzo di Nostradamus. Vol. 3: L'abisso* (Mondadori, 1999), *Picatrix. La scala per l'inferno* (Mondadori, 1998).

[99] L'origine di questo mezzo di comunicazione risale al 1978, quando Ward Christensen e Randy Suess realizzarono a Chicago il primo BBS, noto come CBBS (Computer Bulletin Board System): gli utenti potevano collegarsi utilizzando un modem su linea telefonica per visualizzare i messaggi inviati da altri o pubblicare i propri secondo la logica di una bacheca di avvisi analogica (Huang et al. 2025; Calvo et al. 1996, 24–25). Tipicamente ogni BBS offriva le seguenti funzionalità: un forum di discussione in cui leggere e inviare post, un canale di comunicazione con l'amministratore di sistema (il cosiddetto *sysop*), archivi di file condivisi liberamente e una rubrica degli utenti registrati. Nelle versioni più avanzate furono aggiunti anche servizi di chat e giochi telematici (University of Washington, Paul G. Allen School of Computer Science & Engineering 2019). Poiché sfruttavano le linee telefoniche ordinarie, generalmente le BBS erano organizzate in reti notturne per non gravare sulle tariffe telefoniche in orari diurni (Di Corinto 2017). In Italia, la rete di BBS più famosa fu FidoNet, nata inizialmente negli Stati Uniti nel 1984 ad opera di Tom Jennings e approdata in Italia nel 1986 grazie all'attivazione del nodo *Fido Potenza* di Giorgio Rutigliano (Calvo et al. 1996, 24–25; Sgherza 2016; Di Corinto 2017). Al momento del suo massimo

hard disk, specificamente contenenti materiali testuali, l'Associazione conserva anche altri 272 floppy (software, videogiochi, gestione tipografica alle utility di sistema) che non sono stati oggetto di riversamento, a differenza del materiale di lavoro. Questo materiale è riconducibile, in parte, a supporti originali (ossia materiali editi o legati a riviste informatiche o confezioni commerciali) e in parte a supporti con materiale autoprodotto, con etichette manoscritte che testimoniano pratiche di copia, riuso e distribuzione informale. L'insieme appare composito e copre un ampio arco di interessi di Evangelisti.

Per quanto riguarda i software applicativi, sono presenti *MS Publisher 2.0*, *Aldus PageMaker 5.0*, *Microsoft Excel 5.0*, *Microsoft Word 6.0 (WinWord)*, *Micrografx Designer 3.1*, *Top Draw 2*, *Corel Draw 5*, *Lotus Organizer per Windows*, *Works*, *QEdit 3.0* e *Paint Shop Pro* (nelle versioni 3.12 e per Windows 95)[100].

Si segnala anche una sezione relativa a sistemi operativi, *driver* e *utility* che comprende, tra gli altri, *Microsoft Windows 2000*, *MS-DOS 5.0*, *Red Hat Linux 6.2*, *Norton Utilities 5.0*, *Norton Commander 3.0*, *QEMM 6.2* (per la gestione della memoria), *Stacker 3.0* per Windows/DOS (compressione dei dati), *Copy II PC 8.1*, *CopyWrite 8.0*, *WinTune 97*, il driver *Samsung CD/DVD-ROM Device Driver for E-IDE* e nuovi driver per schede video *ET4000*. Accanto a questi compaiono programmi di grafica e multimedia, come *Autodesk Animator*, *OCR Calera Recognition* (per il riconoscimento ottico dei caratteri), *SkyGlobe 3.5* (planetario interattivo).

sviluppo, FidoNet Italia raggruppava circa 300 nodi gestiti da *sysop* volontari. Le BBS erano perlopiù gestite in forma amatoriale, senza scopo di lucro, portate avanti da singoli, centri sociali e enti locali (Di Corinto 2017). Nel 1991 FidoNet contava oltre 10.000 nodi con circa 100.000 utenti in tutto il mondo, in un momento in cui Internet era ancora confinato in ambienti elitari, prevalentemente nelle Università (Di Corinto e Tozzi 2002).

[100] La presenza di floppy disk contenenti software applicativi, videogiochi e utility di sistema nel fondo Evangelisti solleva la questione dei software applicativi come oggetto di conservazione (o *software preservation*), un ambito specifico della preservazione digitale che ha acquisito crescente rilevanza negli ultimi due decenni (Lowood 2023). La conservazione del software richiede strategie conservative specifiche e più complesse rispetto ad altri oggetti digitali: non è sufficiente preservarne i file binari, ma è necessario garantire la ricostruzione, l'emulazione o la documentazione dell'ambiente tecnologico originario (hardware, sistema operativo, dipendenze e configurazioni) al fine di assicurarne l'effettiva eseguibilità e comprensione nel tempo (Acker 2021). Nel contesto degli archivi letterari, per esempio, la conservazione degli strumenti software utilizzati dagli autori permette di comprendere le pratiche di scrittura e i vincoli tecnologici che hanno condizionato la produzione dei testi. I supporti conservati documentano inoltre le modalità di circolazione del software negli anni Novanta, caratterizzate da pratiche informali di copia e condivisione testimoniate dalle etichette manoscritte e dai dischetti autoprodotti. Organizzazioni come il Software Preservation Network (https://www.softwarepreservationnetwork.org/) (Meyerson et al. 2017) e iniziative quali l'Internet Archive's Software Library (https://archive.org/details/softwarelibrary) hanno sviluppato metodologie e infrastrutture per affrontare le tematiche legate alla *software preservation*, dall'emulazione alle questioni di proprietà intellettuale.

Particolarmente ricco è il nucleo legato ai font e alla tipografia, con strumenti dedicati alla gestione e conversione dei caratteri: *Adobe Type Manager*, collezioni di *Windows TrueType Fonts*, *Atech Fast Font*, *Copy II PC Fontasy* e *AllType Converter (Universal Typeface Converter)*.

La sezione più consistente riguarda tuttavia i videogiochi, a testimonianza di una pratica di uso del computer non solo produttiva ma anche ludica. Sono stati individuati titoli appartenenti a generi diversi: simulatori e giochi d'azione come *Star Wars: X-Wing*, *BlackHawk*, *Terminator 2*, *Stellar 7* e *Syndicate*; avventure grafiche e interattive come *Indiana Jones and the Fate of Atlantis* (etichettato come "Indiana Jones IV"), *Sam & Max Hit the Road*, *Leisure Suit Larry 5*, *Daughter of Serpents*, *Elvira II: The Jaws of Cerberus*, *Shadow of the Comet* e *The Legend of Kyrandia*; puzzle e logica come *Hexxagon*, *Mahjong 32 bit*, *Kyodai Mahjong* e *Scacchi*; giochi d'azione come *Heretic*, *Doom*, *Zone 66* e *Total Carnage*; produzioni meno note come *Magellan*, *More Mandy*, *Tristan*, *La Espada Sagrada*, *Redhook's Revenge* (allegato alla rivista *PC Answers*) e *Invasion of the Mutant Space Bats*. Sono presenti, inoltre, giochi distribuiti tramite riviste come *PC Format* e *PC Globe*, nonché produzioni di carattere più indipendente o underground, come *Tangentopoli*[101] (corredato da materiali relativi all'omonima inchiesta).

È stato rinvenuto, inoltre, un computer portatile DELL contenente 227.027 elementi distribuiti su 52,8 GB, equipaggiato con un hard disk Western Digital WD800BEVS-22RST0 (modello da 2,5 pollici, capacità di 80 GB, serie *WD Scorpio*). Al momento, il dispositivo sembra includere solo due file DOCX riconducibili con certezza alla produzione di Evangelisti; tuttavia, la notevole quantità di elementi presenti richiede indagini più approfondite. Il ritrovamento, avvenuto nella primavera del 2025, non ha permesso di includere il portatile nell'analisi qui presentata, ma il materiale verrà integrato nelle fasi successive della ricerca.

La partizione nativa digitale dell'Archivio Evangelisti rappresenta un caso unico nel panorama italiano, non solo per l'eccezionale estensione, con oltre 2 TB di materiali, ma anche per la varietà e l'estensione cronologica dei contenuti, costituendo finora l'esempio più imponente e articolato di archivio d'autore digitale in Italia accessibile alla ricerca.

3.3 Storia archivistica

Il materiale analogico conservato dall'Archivio Marco Pezzi deriva dalla congiunzione di precedenti versamenti fatti dallo stesso Evangelisti integrati nel 2024 con carte sciolte ritrovate presso il suo appartamento bolognese e tre bobine rinvenute in cantina dai membri dell'Associazione Valerio

[101] https://www.mobygames.com/game/39689/il-grande-gioco-di-tangentopoli/.

Evangelisti-Il Sol dell'Avvenire. Il materiale sciolto era in stato di totale disordine ed è stato fascicolato dal personale dell'Archivio. I filmati Super 8, oltre ad essere conservati nelle bobine originali dall'Archivio, sono stati riversati nel 2024 su hard disk da un fotografo specializzato su commissione dell'Associazione.

Album fotografici e quaderni sono stati recuperati in uno scompartimento della libreria dell'autore dotato di sportelli; attualmente si trovano nei locali dell'Associazione Valerio Evangelisti - Il Sol dell'Avvenire, ma si prevede un successivo versamento all'Archivio Marco Pezzi per unificare la partizione analogica.

Il fondo librario è composto di volumi e riviste rinvenuti all'interno dell'appartamento e in cantina ed è attualmente conservato presso i depositi del Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna.

La partizione digitale del fondo è costituita da materiale acquisito dall'Associazione Valerio Evangelisti-Il Sol dell'Avvenire nell'estate del 2024. Il corpus comprende:

- un hard disk Samsung Portable SSD T7 1 TB USB tipo-C 3.2 Gen 2 su cui è contenuta la copia dell'hard disk della postazione di lavoro principale;
- una memoria esterna Western Digital WDBACW0020HBK-01 My Book 2TB USB 3.0 Series External Hard Drive;
- un computer portatile Dell con hard disk Western Digital WD800BEVS-22RST0 da 80 GB;
- 389 floppy disk da 3.5".

L'Associazione non ha avuto modo di recuperare il computer fisso originale che costituiva la principale postazione di lavoro di Evangelisti, ma ha potuto ottenere una copia dei contenuti trasferiti sull'hard disk Samsung Portable SSD T7 1. L'assenza di accesso al sistema originale non ha reso possibile ricostruire le caratteristiche tecniche del *file system* nativo; è tuttavia verosimile che si trattasse di un *file system* NTFS, poiché l'unico elemento certo riguarda l'utilizzo di un computer con sistema operativo Microsoft Windows recente[102].

La memoria esterna e i floppy disk sono stati rinvenuti in alcuni scompartimenti della libreria di Evangelisti, questi ultimi conservati in custodie di svariate dimensioni (Figura 3.2), prive di aggregazioni o organizzazioni evidenti. Dall'intervista effettuata a Emiliano Mattioli, membro dell'Associazione responsabile dei riversamenti dei floppy disk, emerge che sono stati sottoposti a migrazione

[102] Sono stati rinvenuti alcuni supporti ottici riconducibili a versioni precedenti di sistemi operativi Microsoft utilizzati da Evangelisti: un CD di *Windows XP Home Edition*, Service Pack 2; un CD di installazione di *Windows 95*; un CD di installazione di *Windows 98*; e un CD di *Windows NT Workstation* versione 3.51.

esclusivamente quelli contenenti materiali di lavoro e produzione intellettuale, mentre rimangono conservati solo in formato originale i floppy disk relativi a installazioni software, programmi e videogiochi (si veda, a proposito, il Capitolo 3.2). I riversamenti sono avvenuti in ambiente Microsoft Windows tramite un lettore floppy esterno e, laddove necessario, utilizzando il software *open source* TestDisk[103] per il recupero di partizioni danneggiate.

Tutto il materiale è stato oggetto di migrazione conservativa su commissione dell'Associazione nell'autunno 2024, con riversamento completo su due hard disk esterni Samsung Portable SSD T7 1 TB USB tipo-C 3.2 Gen 2 successivamente consegnati al laboratorio ADLab del Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna per le operazioni di ricognizione e descrizione archivistica. A partire da queste copie sono state effettuate copie di lavoro attualmente conservate su server del Dipartimento con *file system* Ext4[104]. Su esplicita richiesta della famiglia e dell'Associazione, sono stati eliminati dalle copie di sicurezza i materiali di natura strettamente personale e privata anteriormente all'avvio delle operazioni di descrizione.

Figura 3.2. Contenitori 3M in cui sono stati rinvenuti parte dei floppy.

Oltre alla composizione materiale fin qui delineata, la ricostruzione degli ambienti tecnologici di produzione documentaria è in parte recuperabile attraverso l'analisi della corrispondenza elettronica

[103] https://www.cgsecurity.org/wiki/TestDisk_IT.

[104] Server Dell PowerEdge R7525, dotato di due CPU AMD EPYC 7313 16-Core Processor (versione 25.1.1, 1500 MHz, capacità massima 3729 MHz, architettura a 64 bit), 256 GB di RAM, un'unità SSD da 1 TB e un array RAID da 12 TB su hard disk rotativi. Il sistema operativo è Ubuntu 24.04.3 LTS ("noble") con shell Bash 5.2.21-2ubuntu4 (/bin/bash).

dell'autore con la mailing list eymerich@gilda.it, che ha consentito di tracciare, almeno in parte, la parabola evolutiva dei dispositivi utilizzati dall'autore nell'arco temporale 2009-2021[105].

Nel 2009 l'autore disponeva di una doppia configurazione: un portatile Flybook, caratterizzato da elevata portabilità ma limitate capacità computazionali, utilizzato prevalentemente durante i viaggi; un computer fisso con caratteristiche hardware modeste, impiegato come principale stazione di lavoro. Questo sistema manifesta criticità nel gennaio 2010, quando si verificano malfunzionamenti che comportano il rischio di perdita del primo capitolo di un romanzo in fase di redazione (probabilmente *One Big Union*, Mondadori, 2011).

Nel gennaio 2013 acquisisce un nuovo computer con specifiche tecniche notevolmente migliorate (alimentazione potenziata e scheda grafica GeForce GTX 660), che diviene il principale ambiente di produzione. L'autore comunica via mail l'avvio di complesse operazioni di migrazione di dati dal sistema precedente, processo descritto come particolarmente laborioso. Questo dispositivo mantiene il ruolo di stazione di lavoro primaria fino al maggio 2021, quando le comunicazioni email dello scrittore lo descrivono come un computer ormai obsoleto e soggetto a rallentamenti e rischi concreti di perdita dati. La situazione precipita nel luglio 2021 con un *crash* totale del sistema. Durante questa fase critica, l'autore riporta di utilizzare temporaneamente un vecchio portatile (probabilmente il portatile DELL successivamente rinvenuto dall'Associazione), impiegato principalmente per connessione internet e ripristino della posta elettronica, nonostante le prestazioni estremamente ridotte. Nello stesso mese viene acquisito un nuovo computer e un tecnico specializzato interviene per il recupero dati dal dispositivo danneggiato. Al 19 luglio 2021 risulta completato il recupero della quasi totalità del materiale preesistente.

L'utilizzo di sistemi cloud è documentato all'interno delle cartelle, ma l'assenza delle relative credenziali ha impedito l'accesso ai contenuti, determinando una lacuna documentaria nella ricostruzione dell'ambiente digitale complessivo. L'archivio comprende, tuttavia, diversi backup in formato PST della posta elettronica provenienti dai due account principali di Evangelisti: nicolas@eymerich.com e

[105] Per ulteriori dettagli sulla mailing list si veda infra, p. 75. L'analisi della corrispondenza di Evangelisti nell'ambito della mailing list eymerich@gilda.it è stata condotta attraverso l'account di un membro dell'Associazione Valerio Evangelisti. L'archivio, tuttavia, conserva anche alcuni file di posta personale dell'autore in formato PST (Personal Storage Table), formato proprietario di archiviazione utilizzato da Microsoft Outlook. Per verificarne l'integrità sono state create copie decompresse al di fuori della struttura archivistica originaria. Il trattamento e la descrizione di questi materiali non sono stati previsti nell'ambito del presente progetto, sia per la notevole mole dei dati contenuti, sia per l'estrema delicatezza della corrispondenza privata. La gestione futura di questi materiali dovrà tenere conto sia delle problematiche tecniche legate all'obsolescenza del formato proprietario, sia delle questioni etiche relative alla tutela della riservatezza dell'autore e dei suoi corrispondenti.

eymerich@tin.it, quest'ultimo non più attivo negli ultimi anni di vita dell'autore. L'analisi dei file di backup ha rivelato anche l'esistenza dell'account valerioevangelisti22@gmail.com, del quale tuttavia non sono conservate tracce locali. L'esame dei log del sistema (Event Viewer Registri di Windows Applicazione) del computer portatile ha inoltre evidenziato eventi associati all'account valerioevangelisti75@gmail.com, anch'esso privo di archiviazione locale.

3.4 "Finché c'è lotta informatica c'è speranza": Evangelisti fra tecnologia e letteratura

Questo capitolo analizza la figura di Valerio Evangelisti come esempio di intellettuale capace di attraversare criticamente la transizione digitale, facendo della tecnologia non solo uno strumento di lavoro ma un oggetto di riflessione e sperimentazione culturale. Dal punto di vista archivistico, la sua produzione documentaria testimonia le modalità con cui la consapevolezza tecnologica dell'autore ha inciso sulla forma e sull'organizzazione del suo archivio.

Evangelisti può essere considerato un intellettuale paradigmatico della transizione digitale: pioniere nell'adozione tecnologica, critico nelle valutazioni e visionario nelle prospettive future. Il percorso di Evangelisti esemplifica il modello evolutivo che Irene Cacopardi ha identificato negli intellettuali contemporanei alle prese con la rivoluzione digitale: un passaggio dalla prima fase, caratterizzata dall'entusiasmo verso le potenzialità democratiche e collaborative delle nuove tecnologie, a una seconda fase in cui «emerge come la nuova invenzione crei problemi inediti e determini un nuovo campo epistemico» (Cacopardi 2023, 107). Questa transizione comporta «una presa di distanza, un cambio di prospettiva, l'assunzione di uno sguardo più critico» (Cacopardi 2023, 107) anche da parte di autori estremamente integrati nelle dinamiche digitali.

Il rapporto di Evangelisti con la tecnologia non è mai stato passivo o meramente strumentale, ma si configura come un dialogo costante tra innovazione tecnica e riflessione critica, tra possibilità espressive e consapevolezza dei limiti. Questa prospettiva permea la sua intera produzione letteraria – dalla fantascienza al fantasy storico, dalla narrativa sui movimenti operai nordamericani alle saghe piratesche – dove la dimensione tecnologica assume un ruolo duplice: medium narrativo e territorio di indagine intellettuale. La saga di Nicolas Eymerich, che lo ha reso noto al grande pubblico vincendo il Premio Urania nel 1993, rappresenta un perfetto esempio di come la tecnologia diventi elemento strutturante della narrazione: le vicende dell'inquisitore medievale si proiettano sistematicamente in futuri distopici, dove la tecnologia amplifica le dinamiche di controllo e oppressione già presenti nel passato storico.

Le interviste rilasciate negli anni da Evangelisti e numerosi scambi mail (intercorsi principalmente con le sue mailing list di riferimento) offrono uno spaccato sul rapporto dell'autore con l'innovazione tecnologica, iniziato precocemente: «Comperai il mio primo computer nel 1979, mi collegai a Internet quando ancora la rete non era pubblica, bensì riservata, in Italia, a poche università. Io e il computer facciamo, in pratica, una cosa sola, da decenni» (Maugeri 2011). Il computer diventa parte integrante del processo creativo, che l'autore organizza in una dimensione temporale specifica: la notte. Le ore notturne sono per Evangelisti uno spazio privilegiato di scrittura, «quando tutti dormono e nessuno ha la brutta idea di telefonarmi» (Girolami 2011). Tuttavia, non si tratta soltanto di una questione di tranquillità ambientale, ma della costruzione di un cerimoniale di machiavelliana memoria che riferisce l'autore stesso:

> [e]siste un rituale che è un po' quello che descriveva Machiavelli: arrivata una certa ora, vestiva panni regali e curiali, lasciava l'osteria in cui viveva solitamente e si ritirava fra i suoi libri. Io faccio un po' la stessa cosa – anche se adesso frequento meno le osterie rispetto a una volta. Però arriva una certa ora, di solito molto tardi, verso mezzanotte, e lì mi metto davanti al computer con un atteggiamento di grande rispetto per il computer e per quello che sto per fare[106] (Girolami 2011).

Il suo processo creativo notturno, d'altronde, è largamente documentato anche dalle tracce digitali lasciate nel suo archivio, che testimoniano orari di creazione e modifiche dei file attestati in piena notte (v. capitolo 8.5 e 8.6).

Nell'intervista rilasciata per il blog "libroguerriero" nel dicembre del 2013 (Marilù Oliva 2013), Evangelisti parla di tempi lavorativi particolarmente serrati:

> Scrivo minimo due pagine tutti i giorni. In realtà gran parte del tempo è presa, più che dallo scrivere, dalle ricerche. Anche nei miei romanzi fantastici c'è uno sfondo storico che cerco di rendere il più preciso possibile, quindi ho bisogno di leggere molto. Internet ha reso la vita più facile, certo, poi cerco di informarmi sulla realtà attuale, che ha sempre un nesso con quello che scrivo, e dunque leggo giornali on line e vivo in simbiosi con il computer. Complessivamente il

[106] Il riferimento è alla celebre lettera di Niccolò Machiavelli a Francesco Vettori del 10 dicembre 1513, in cui Machiavelli racconta di giornate spese in faccende triviali e di come al tramonto torni a casa, sentendo il bisogno di vestire panni "reali e curiali", per poter figurativamente entrare nelle antiche corti, cioè per leggere e mettersi in una dimensione spirituale di confronto con gli antichi. In questo quadro, nella seconda metà del 1513, Machiavelli scrive di getto il *Principe*, la cui redazione è annunciata nella lettera stessa (Biasiori 2018).

> tempo lavorativo sarà di 7-8 ore al giorno, suddivise in una tranche pomeridiana e in una tranche notturna, il che significa arrivare fino alle 6-7 di mattina.

La metodicità della produzione quotidiana di Evangelisti nasce da una routine quasi monastica, scandita dai ritmi di una vita deliberatamente solitaria. Questa disciplina rappresenta una scelta metodologica quanto una necessità pratica:

> scrivo un romanzo all'anno per un motivo: mi consente di mantenere il mio tenore di vita. Per cui, finché me li prenderanno, continuerò a scriverne uno all'anno. […] Poi vedo che col tempo diventa più faticoso, anche se è sempre una cosa entusiasmante. Frustrante ed entusiasmante. Frustrante perché non sai mai se raggiungerai il risultato, entusiasmante perché vivi mille vite. Per cui io sono stato sulle navi pirata, sono stato nel Medioevo, sono stato in giro per il tempo e per lo spazio (Caimi 2013b).

La scrittura diventa, così, un momento di attraversamento di mondi diversi, facilitato dalla connessione che abolisce i confini temporali e geografici. Aspetto fondamentale, quest'ultimo, anche per la consultazione delle fonti per la stesura dei suoi romanzi. Storico di formazione, di fronte alla difficoltà di proseguire con la carriera accademica, decide infatti di riprendere la storia in altre forme, mettendola al centro delle sue narrazioni, intessute su quadri storici attentamente ricostruiti (Caimi 2013b). La ricerca storica diventa per l'autore uno strumento di demistificazione del presente: «storicizzare il presente e mostrarne l'origine, l'evoluzione e i fattori culturali che l'hanno determinato è offrire al lettore un percorso per vincere il suo analfabetismo, per ritrovare l'importanza della memoria storica» (Sebastiani 2018, 218).

L'ampio utilizzo di fonti storiche è largamente testimoniato dalla composizione del suo archivio digitale, particolarmente ricco di materiali bibliografici e documentali scaricati dal web o ottenuti da riproduzioni digitali di materiali d'archivio. Questa facilitazione tecnologica, tuttavia, non sostituisce mai la ricerca tradizionale, che rimane centrale nelle sue abitudini e che lo portano a consultare, per ogni romanzo, «non meno di un centinaio di volumi» (Caimi 2013b), dimostrando come l'innovazione digitale per Evangelisti amplifichi, piuttosto che sostituire, i metodi di documentazione tradizionali.

Ma il web permette anche di avere interlocutori nuovi, individuando una dimensione dialogica che supera i canali editoriali tradizionali, creando spazi di confronto diretto e immediato: «[d]a un lato il web si presta a dei pericoli. Uno dei pericoli è che, dato che ognuno può parlare, si arrivi al livello più basso. Dall'altro lato tantissima gente che non aveva voce ha cominciato a essere nota proprio a livello di web» (Caimi 2013a). In questo senso, interrogato per Doppiozero sul modo in cui la rete abbia cambiato il

modo di fare letteratura, Evangelisti risponde: «[h]a concesso a tutti di farne, nel bene o nel male. Oggi lo scrittore sublime o il peggiore cretino riescono a farsi leggere, in tutto il mondo. A me sembra un progresso, con qualche remora marginale (riferita al cretino)» (Forlani 2018).

Attorno a Evangelisti, dalla fine degli anni Novanta, si stringe effettivamente una comunità di lettori che partecipano attivamente alla costruzione di uno spazio virtuoso di stimoli reciproci, in cui l'autore alternava stesura, ricerca e confronto diretto attraverso mailing list molto frequentate, in particolare eymerich@gilda.it[107]. In una mail alla lista del 27 luglio 2001, Evangelisti ne definisce il perimetro in questi termini:

> non siamo affatto in una lista di fantascienza. Questa mailing list è consacrata a tutte le tematiche presenti nei miei romanzi e nei miei scritti, e mi sembra chiaro che temi come la mondializzazione, la costruzione di uno stato autoritario, le perversioni dell'economia, i poteri occulti, il fascismo e la violenza vi figurano abbondantemente (Evangelisti 2001b).

La mailing list nasce nel 1997 a seguito della partecipazione di Evangelisti al Simposio Internazionale "Wilhelm Reich. storia di un'evoluzione metodologica. Verso una nuova epistemologia", svoltosi a Napoli il 5 e il 6 aprile 1997. Evangelisti ricorda che, contro i molti interventi che criticavano Internet, lui sostenne invece come la rete potesse servire a creare una comunità: «la creazione della mailing list fu un poco la traduzione in pratica delle mie tesi. Penso che la permanenza in vita della lista, pur tra mille crisi, dimostri che avevo ragione» (Evangelisti 2008).

Questa prima fase viene raccontata da Giovanni Secondulfo, uno dei fondatori della lista, nel suo contributo al volume *Il mondo dei fan club* (Adnkronos libri, 2000). Secondulfo riporta un'intervista a Evangelisti in cui emerge pienamente il ruolo della comunità digitale anche nel suo processo creativo:

> fin dall'inizio ho privilegiato il rapporto con i miei lettori, attraverso un uso intensissimo dei mezzi informatici. Prima c'è stata la rete telematica Fidonet, poi è venuta Internet. Me ne sono servito per entrare in diretto contatto con chi mi leggeva, per dialogare con loro, raccoglierne pareri, sottoporre al loro giudizio i miei romanzi, chiedere consiglio. A Napoli è nato in questo modo il primo nucleo del fan club, totalmente affidato alla comunicazione via computer, ma con una sua consistenza effettiva e dei risvolti pratici tutt'altro che trascurabili (per dirne una, i miei

[107] La mailing list è ad oggi ancora attiva e ha continuato a portare avanti riflessioni e attività su temi politici, sociali e letterari, fra cui le iniziative in memoria di Evangelisti.

> corrispondenti mi informano in tempo reale sullo stato della distribuzione dei miei romanzi, e io passo le notizie all'editore). Piano piano mi sono trovato a vivere in simbiosi con i miei fan, trasformatisi in amici che conosco uno per uno. Passo diverse ore al giorno a rispondere alla loro posta e a dialogare, attraverso la mia mailing list sui temi più svariati. È faticoso, ma lo considero parte del mio lavoro (Secondulfo 2000, 81)

La solidità di questa comunità si è concretizzata, a partire dal 2000, attraverso l'organizzazione di raduni annuali, eventi autofinanziati e completamente informali, ospitati in varie località italiane del nord e del sud. Incontri alla pari tra l'autore e i partecipanti, che nel corso del tempo hanno rafforzato legami e relazioni capaci di andare oltre gli scambi di idee e dialoghi portati avanti nella mailing list virtuale (Mailing List, s.d.).

In archivio troviamo traccia di questi scambi sia in copie di backup locali delle sue caselle di posta (in formato PST) presenti sull'hard disk principale e su quello esterno, sia in otto floppy disk contenenti messaggi inviati e ricevuti fra il 1° dicembre 1994 e il 28 dicembre 1997.Questi ultimi, in particolare, sono principalmente copie di messaggi salvati da BBS (Bulletin Board System), una forma pionieristica di forum telematici, in uso già prima della diffusione di Internet e del World Wide Web (Coleman 2013, 30). Gli argomenti trattati nei BBS spaziavano da temi tecnici (software, programmazione, reti, gaming) a questioni di impegno civico e culturale (democrazia elettronica, tematiche politiche, sperimentazione artistica) (Di Corinto 2017). Le BBS di fantascienza, come *Sf.Ita* di FidoNet, rappresentavano, in particolare, laboratori culturali dove si sperimentavano forme innovative di critica letteraria, discussione collettiva e condivisione di materiali (Sosio 1996, 2004).

Le comunità italiane di BBS si distinguevano per il forte spirito volontario e l'attenzione agli aspetti sociali e politici. Non è un caso che alla diffusione dei primi circuiti abbiano contribuito anche realtà legate ai centri sociali occupati e autogestiti, luoghi di antagonismo e sperimentazione culturale in cui si criticava la comunicazione verticale dei media tradizionali e si sviluppavano pratiche di autoproduzione tecnologica, dal software all'editoria (Di Corinto 2017). Proprio in questo contesto nacquero esperienze come Cybernet, una rete parallela a FidoNet alimentata dal gruppo della rivista *Decoder* (Capussotti 2001), prodotta da un gruppo punk in alcuni centri sociali milanesi come il Cox18 e il Leoncavallo, e la Avvisi Ai Naviganti BBS (AvANa BBS)[108], attiva dal 1994 all'interno del centro sociale Forte Prenestino

[108] AvANa è ancora attivo come *hacklab* al Forte Prenestino. Fra le varie attività, i partecipanti curano la trasmissione "Le dita nella presa" su Radio onda rossa. Cfr. https://avana.forteprenestino.net/about/.

di Roma, che affiancava attività di alfabetizzazione digitale a progetti di democrazia elettronica (Di Corinto e Tozzi 2002; Di Corinto 2017).

In queste reti gli utenti attivi erano spesso programmatori e militanti, *hacktivists*, che interpretano l'*hacking* come pratica sociale e politica, oltre che tecnica. L'*hacktivism* nasce infatti dall'incontro tra l'etica *hacker*, basata su valori come la condivisione della conoscenza, la cooperazione, la libertà di accesso all'informazione e la creatività come atto collettivo, e le pratiche dei movimenti sociali, caratterizzate da attivismo dal basso, orizzontalità decisionale e critica delle forme di potere e controllo (Karagiannopoulos 2021; Romagna 2020; Żuchowska-Skiba 2024). In questa prospettiva, il computer e le reti non sono più soltanto strumenti di lavoro o di consumo, ma si trasformano in spazi di conflitto e di liberazione, luoghi dove sabotare i modelli comunicativi dominanti e allo stesso tempo creare nuove forme di partecipazione democratica.

L'importanza dell'*hacktivism* risiede dunque nella sua capacità di ridefinire il ruolo della tecnologia come bene comune e strumento di emancipazione: invece di alimentare dinamiche di profitto e controllo, utilizzata in funzione della costruzione di comunità, al rafforzamento dei diritti civili e alla creazione di alternative culturali e politiche. La sintesi di questi due momenti, ossia la pratica *hacker* e la militanza politica, ha visto la nascita di un nuovo modello di informazione collettiva fondato su relazioni cooperative e comunicazione orizzontale, che avrebbe trovato piena espressione nello sviluppo dei media indipendenti sul Web, come ad esempio *Indymedia*, nato nel 1999 nello spirito della controcultura digitale (Garcelon 2006). In questo, come osserva Carlo Gubitosa, il fenomeno delle BBS ha avuto un ruolo fondativo:

> è sui BBS e non su internet che hanno inizio i percorsi di riflessione culturale e tecnologica sulle conseguenze della rivoluzione digitale. I BBS diventano un laboratorio di sperimentazione collettiva, in cui la società civile, tagliata fuori da un'Internet ancora elitaria, comincia a discutere di privacy, crittografia, editoria elettronica, censura e controllo delle informazioni da parte dei governi, tecnocrazia, diritti telematici, copyright, libertà del software, cultura cyberpunk (1999, 20).

Ben 577 file conservati nell'Archivio Evangelisti documentano la sua partecipazione a questi ambienti telematici pionieristici, costituendo una fonte preziosa anche per la ricostruzione della storia delle BBS italiane, tanto più significativa se consideriamo la scarsità delle tracce native digitali di questa

esperienza[109]. La partecipazione di Evangelisti alle comunità BBS rappresenta perfettamente quella dimensione originaria della tecnologia come laboratorio di esperienze democratiche e spazio privilegiato di creazione artistica, dove «l'attività letteraria (ri)trova una dinamica aperta e collaborativa che spesso crea nuove economie narrative» (Cacopardi 2023, 107).

La presenza di Evangelisti in tali circuiti non fu episodica né marginale, e si avvicinò alle comunità digitali con una preparazione tecnica non trascurabile. Come lui stesso ha ricordato in un'intervista rilasciata a Wired nel 2011, le sue prime esperienze di *hacking* risalgono agli anni universitari, quando riuscì ad accedere al sistema informatico dell'Università di Bologna scoprendo la password di amministrazione, impostata su "alma" (Girolami 2011).

Il periodo documentato dai floppy disk coincide con la fase cruciale di transizione dai BBS volontaristici e gratuiti al Web commerciale, una trasformazione che Evangelisti ha osservato e vissuto direttamente. L'esperienza in questi ambienti telematici ha fornito all'autore un osservatorio privilegiato sulle dinamiche del potere digitale emergente, in cui la libertà digitale è costantemente minacciata dalla logica economica dominante.

In un'intervista del 2008, a seguito dell'uscita di *Tortuga*, Evangelisti propone un parallelo, implicito anche nell'opera, fra la pirateria storica e quella informatica a cavallo degli anni Duemila. A proposito dei pirati del Seicento, afferma:

> vivevano di predazione e di una sorta di forma repubblicana di autogoverno: eleggevano i capi, i loro erano comandanti revocabili e, se non reputati all'altezza, passibili di pena di morte; avevano una vita sessuale decisamente libertina. Ma la loro libertà era un'apparenza concessa: i pirati furono ampiamente manovrati dalle potenze, la comunità di Tortuga fu a servizio della Francia ma quando non servì più al Re fu rasa a zero, lo stesso accadde ai pirati della Giamaica a servizio dell'Inghilterra (Garofalo 2008).

Il ragionamento si estende quindi al presente: «I pirati vivevano in un'area, il loro impiego andava a beneficio di uno e a danno di altri, allo stesso modo nella pirateria informatica non si è garantiti di essere liberi, perché il sistema fa parte di un'economia. Da quando esiste l'Indice dei titoli informatici, non si possono fare troppe battaglie» (Garofalo 2008).

[109] Per il contesto americano, il progetto textfiles.com, curato da Jason Scott, si occupa di recuperare e pubblicare i files provenienti dai BBS statunitensi degli anni 1980-1995. Cfr. http://www.textfiles.com/.

La visione dell'autore rimanda così alle tematiche affrontate negli stessi anni dalla comunità *hacker* e da riviste underground come *Phrack*[110] ("written by hackers for hackers"), che ponevano in discussione l'effettiva libertà dell'*hacking* contemporaneo e il rischio di una sua crescente subordinazione, talvolta inconsapevole, agli interessi delle multinazionali (Garofalo 2008).

Contro la persistenza delle dinamiche di potere anche negli spazi apparentemente più liberi della rete, Evangelisti invita a diffondere nel Web una cultura dell'auto-regolamentazione attraverso la *netiquette*, ossia «l'insieme delle norme di comportamento, non scritte ma a volte imposte dai gestori, che regolano l'accesso dei singoli utenti alle reti telematiche» (Treccani Vocabolario Online, s.d.). Evangelisti si richiama all'etica delle prime comunità telematiche di cui ha fatto parte, nelle quali la *netiquette* non era un semplice codice di buona condotta, ma una condizione pratica necessaria per il funzionamento delle comunità volontarie[111] (Garofalo 2008). La proposta di Evangelisti colloca l'autorità non in strutture verticali ma in dinamiche orizzontali di responsabilità condivisa, in linea con le pratiche delle comunità digitali alternative (Di Corinto e Tozzi 2002; Capussotti 2001). Questa posizione si inscrive nella consapevolezza dei meccanismi di un sistema dominante che genera isolamento e competizione distruttiva: «il neoliberismo presuppone proprio il venir meno di quella nozione di comune sostanza umana che si concretizza in solidarietà, cioè nell'antitesi alla competizione» (Evangelisti 2001a, 37). Di fronte a questa realtà, il suo impegno è declinato in termini di continuità militante, di lotta contro il sistema: «Se non sono io a portare avanti la lotta ci sarà sempre qualcuno che lo fa e io sarò sempre dal loro lato. Finché c'è lotta informatica c'è speranza». Riecheggia così il principio che la resistenza digitale non sia solo un gesto tecnico, ma un atto collettivo e politico destinato a prolungarsi nel tempo, al di là dei singoli individui (Jordan e Taylor 2004).

L'adozione delle tecnologie emergenti da parte di Evangelisti non deriva quindi da un fascino acritico per il nuovo, ma dal desiderio di una comprensione profonda delle trasformazioni in atto. Un esempio paradigmatico, in questo senso, emerge anche dalle sue considerazioni sul valore e sul futuro degli e-book risalenti alla fine degli anni Duemila. Infatti, pur avendo concesso i diritti per la digitalizzazione di una quindicina dei suoi romanzi, esprime un netto dissenso verso il libro elettronico inteso come mera copia digitale del suo gemello cartaceo:

[110] https://phrack.org/.

[111] A questo proposito, si segnala il cosiddetto "Jargon File", un documento originariamente redatto nel 1975 da Raphael Finkel della Stanford University e attualmente gestito da Eric S. Raymond, un esponente della cultura *hacker*. Il file mappa il vocabolario del gergo usato dagli *hacker* e dai professionisti dell'IT, ma contiene anche definizioni e regole di buona educazione da rispettare in rete (la cosiddetta *netiquette*). Cfr. http://www.catb.org/jargon/html/.

> Il mio pessimismo sugli e-book ha motivazioni completamente diverse da un atteggiamento oscurantista. Quando si passò dal manoscritto alla stampa non cambiarono solo le forme, ma anche i contenuti. Idem per il passaggio dalla radio alla televisione. Se nuovo è il mezzo, nuovo dovrebbe anche essere il prodotto di cui fruire. Anni fa vennero di moda i CD, che avrebbero, secondo gli ottimisti, sostituito libri ed enciclopedie. Tutto quello che fu divulgato con quel mezzo giace ora in un angolo, tra la polvere. Le enciclopedie, in particolare, ebbero il loro momento di popolarità. Alla voce "Hitler" si poteva ascoltare anche qualche passaggio di un discorso di Hitler. Ridicolo. Altrettanto ridicolo è pensare che un testo possa essere letto come un libro su un costosissimo (per ora) supporto. Senza possibilità – che non sia complicata – di tenerlo in mano agevolmente, sfogliarne pagine a caso, leggerne e rileggerne i risvolti di copertina. I vari formati di e-book possono andare bene per i quotidiani (la lettura su schermo è rapida, si possono scorrere i titoli), per Wikipedia, per voci ipertestuali. Non va bene per i romanzi, a meno che non comprendano musiche e filmati. La frontiera dell'utente informatico si è spostata da decenni oltre la pura lettura. Amazon ha dichiarato che i testi venduti per il suo Kindle sono ormai maggioranza. "Le Monde", pochi giorni dopo, ha rivelato che le cifre erano fasulle. Personalmente sono per la multimedialità. I libri elettronici possono soddisfare un'esigua minoranza. Via, invece, ai filmati, ai videogiochi, alle musiche, alle animazioni. Che lo scrittore, approfittando dei nuovi media, torni a essere l'artista globale che era un tempo (Maugeri 2011).

La critica di Evangelisti, dunque, non nasce da atteggiamenti conservatori, ma da una comprensione storica delle dinamiche dell'innovazione mediale: ogni passaggio tecnologico di successo comporta trasformazioni sia formali, sia contenutistiche. Questa consapevolezza storica lo conduce a immaginare un paradigma alternativo di e-book, concepito come testo multimediale, concetto che negli stessi anni prendeva piede con il termine di "enhanced e-book" (James e De Kock 2013). In una mail inviata alla community l'8 ottobre 2010, Evangelisti affermava:

> avendo a disposizione un mezzo così duttile, è mai possibile limitarsi alla pura riproduzione su schermo di un libro cartaceo? […] Nel mio romanzo che sta per uscire sono citati dei brani musicali. Un e-book degno di questo nome, a mio parere, dovrebbe comprendere dei link a degli mp3 con i pezzi che cito. E quando nomino un luogo, altri link dovrebbero collegarsi a foto o filmati. Solo così, secondo me, l'e-book avrà un futuro. E sarà una sfida eccitante per lo scrittore, costretto ad ampliare la sua gamma di strumenti espressivi e a cimentarsi con le nuove tecnologie. Ne ho parlato più volte con professionisti dell'editoria, ma non sembrano capire di cosa parlo. Gratta gratta, hanno sempre in mente la macchina da scrivere.

La riflessione continua anche in una mail di qualche ora dopo:

> Non parlo degli e-book di oggi, ma di quelli, ipotetici, di domani. A me piacerebbe molto scrivere un romanzo pienamente multimediale. Per il momento, la saga di Eymerich – terminata con il prossimo romanzo – continuerà nel videogioco concepito da Ivan Venturi e dal suo team[112], visto con ammirazione a Sansepolcro. Personalmente, non riesco a cogliere – specie pensando alle prossime evoluzioni – separazioni qualitative nette tra libro, film, fumetto, videogioco ecc. Mi piacerebbe sapermi muovere in tutti questi campi. Ahimè, l'età non me lo consente. Qualcun altro lo farà.

L'apparente contraddizione tra la critica teorica agli e-book e il loro utilizzo pratico si risolve, comunque, nell'archivio di Evangelisti: la presenza di diversi software e-book reader e manager, come Calibre[113], Adobe Digital Editions[114] e My Kindle Content[115], testimoniano come l'autore, pur mantenendo le sue riserve concettuali, non rinunciasse agli strumenti disponibili per la fruizione di contenuti digitali, aperto alle potenzialità della convergenza digitale. Il suo auspicio, in questo, era che lo scrittore tornasse a essere "artista globale", capace di utilizzare tutte le possibilità espressive offerte dai nuovi media. Questa visione, nella sua esperienza, si concretizza in numerosi progetti sperimentali: dai fumetti (come *La furia di Eymerich* (2003, Francesco Mattioli), *Lazarus Ledd Extra. I cristalli di Eymerich* (2003, Ade Capone e Arturo Lozzi) e i volumi di *Nicolas Eymerich inquisitore* (2003-2007, Jorge Zentner e David Sala)) alle sceneggiature radiofoniche (*La scala per l'inferno*, 1998; *Il castello di Eymerich*, 2000; *La furia di Eymerich*, 2001; *Il processo di Wilhelm Reich*, 2006), fino ai videogiochi realizzati con Ivan Venturi.

L'utilizzo degli spazi digitali da parte di Evangelisti non era casuale ma rispondeva a una precisa visione del ruolo dello scrittore contemporaneo. La sua strategia comunicativa mirava a dimostrare che, anche tramite la letteratura di genere, è possibile stimolare un impegno civile nel lettore, che «gode dell'avventura, di una narrazione ricca e ben costruita, complessa, ma scopre che è finalizzata a usare la

---

[112] A partire dal 2012 TiconBlu e Imagimotion hanno prodotto il videogioco *Nicolas Eymerich, Inquisitore: La Peste*, ispirato ai romanzi di Evangelisti. Il primo episodio, *Inquisitore*, è stato lanciato sul mercato italiano nel novembre 2012 con doppiaggio e sottotitoli italiani, introducendo per la prima volta nel settore videoludico anche una versione completamente in latino (Mameli Andrea 2012). Dal luglio 2013, Microïds ha iniziato a distribuire l'edizione internazionale del gioco, caratterizzata da audio disponibile in inglese e italiano e sottotitoli in inglese, francese, tedesco, spagnolo e italiano. *Nicolas Eymerich, Inquisitore: La Peste* si è distinto anche per la duplice modalità di fruizione: come tradizionale avventura grafica e come audiogame, rendendolo accessibile anche agli utenti con disabilità visive (Michela Trigari 2012). Nel 2014 è stata commercializzata la prima edizione italiana del secondo capitolo della serie, dal titolo *Il villaggio* («Avventura, non solo cinema» 2013) .

[113] https://calibre-ebook.com/.

[114] https://www.adobe.com/it/solutions/ebook/digital-editions/download.html.

[115] https://www.amazon.com/b/ref=ruby_redirect?ie=UTF8&node=16571048011.

piacevolezza della paraletteratura e il suo "massimalismo" genetico per attuare un uso improprio (o proprio) di armi dell'intrattenimento di massa con un'intenzione sovversiva» (Sebastiani 2018, 218, 219).

Come evidenzia l'analisi di Sebastiani, «il campo di battaglia è ancora una volta l'immaginario», e l'autore vi agisce sfidando il lettore a «ricostruire il grande affresco storico della contemporaneità e dei suoi possibili futuri» (Sebastiani 2018, 218, 219). Per Evangelisti, la letteratura non ha un potere rivoluzionario in sé, ma può innescarlo, può dare consapevolezza e far comprendere al lettore che «[s]e non gli piace quel che vede, deve trovare il modo di cambiarlo. Non ha che una possibilità: deve ribellarsi a Eymerich, nelle forme e nei modi che ha a disposizione, o che ritiene opportuni» (Sebastiani 2018, 219).

Questa consapevolezza del potere trasformativo della narrazione si riflette anche nel rapporto di Evangelisti con la tecnologia, che testimonia come un intellettuale possa affrontare criticamente le trasformazioni del proprio tempo. Il suo approccio ai media digitali metteva i riflettori su questioni tanto centrali quanto conflittuali dell'era digitale: ripensare i contenuti in funzione dei nuovi media, individuare modelli economici sostenibili per la creazione digitale, riconoscere la fragilità degli ecosistemi digitali e le potenzialità della multimedialità. La sua eredità dimostra che innovazione e pensiero critico non si escludono, e che la fascinazione per il nuovo può convivere con la consapevolezza dei rischi. In un'epoca di accelerazione tecnologica costante, il suo esempio offre strumenti per navigare criticamente le trasformazioni in corso, mantenendo viva la tensione tra possibilità creative e responsabilità intellettuale.

Parte II - Modellazione

4. Elementi costitutivi per la descrizione del digitale d'autore

La Parte I ha affrontato la prima domanda di ricerca (RQ1) – come si configura un archivio d'autore contemporaneo e quali pratiche adottano gli autori nella produzione, gestione e conservazione dei propri materiali – attraverso un percorso che dal quadro teorico generale si è mosso verso l'indagine delle pratiche degli autori contemporanei e la presentazione del caso di Valerio Evangelisti. Questa ricognizione ha evidenziato la complessità del digitale d'autore: materiali distribuiti su supporti eterogenei, stratificati su decenni di evoluzione tecnologica, caratterizzati da pratiche gestionali idiosincratiche e da una ricchezza di metadati senza precedenti rispetto all'analogico. Delineato il quadro fenomenologico, si pone ora la seconda domanda di ricerca (RQ2): come rappresentare il digitale d'autore, analizzando la complessità e le relazioni contestuali?

Questo capitolo funge da raccordo fra l'analisi fenomenologica del digitale d'autore e la modellazione della sua rappresentazione attraverso l'individuazione delle dimensioni che hanno implicazioni dirette sulla pratica descrittiva. L'analisi si concentra sulla stratificazione tra materialità e intangibilità del documento digitale (cap. 4.1), sulla presenza di metadati nativi e le loro tipologie (cap. 4.2), sulle multiple dimensioni della *provenance* (cap. 4.3) e sulla complessità dei contesti archivistici (cap. 4.4). Queste quattro dimensioni, pur senza pretesa di esaustività rispetto alle sue problematicità, circoscrivono le peculiarità essenziali del digitale d'autore che devono essere considerate per lo sviluppo di un approccio descrittivo. Tali dimensioni costituiranno i requisiti teorici che guideranno sia la valutazione critica dei modelli esistenti, sia lo sviluppo della proposta ontologica presentata nei capitoli successivi.

4.1 Caratteristiche del documento digitale

La distinzione tra caratteristiche intrinseche ed estrinseche dei documenti rappresenta un pilastro metodologico consolidato in archivistica, paleografia e diplomatica per l'analisi dei documenti cartacei. Secondo la definizione che ne dà Alessandro Pratesi, si definiscono caratteri estrinseci «quelli che si riferiscono alla fattura materiale del documento e ne costituiscono l'apparenza esteriore, potendosi esaminare indipendentemente dal contenuto» (Pratesi 2018, 64); mentre sono intrinseci quei caratteri «che si riferiscono al contenuto del documento, inteso sempre, però sotto l'aspetto formale» (Pratesi 2018, 73).

Nell'analisi dei documenti cartacei, questa dicotomia permette di separare gli elementi contenutistici del documento (il dato storico/informativo, le formule, il linguaggio etc.) dagli aspetti materiali (supporto, formato, legatura etc.): «permettendo di ricostruire il processo genetico caso per caso, e consentono al tempo stesso una minuziosa analisi comparativa, rivelando analogie e differenze che chiariscono la misura esatta della coincidenza tra regola normativa e applicazione pratica» (Pratesi 2018, 63).

Applicata al contesto *born-digital*, questa distinzione conserva la sua validità concettuale e risulta forse ancor più necessaria, pur richiedendo una ridefinizione che tenga conto delle peculiarità dei documenti digitali e del complesso rapporto tra contenuto informativo e materialità del medium.

La necessità di tale ridefinizione emerge dalla trasformazione del concetto stesso di "primary record". Se tradizionalmente un *primary record* poteva essere definito come «a physical object produced or used at the particular past time that one is concerned with in a given instance» (Modern Language Association 1995), questa concezione nel digitale non può più essere assunta come coestensiva con quella di "oggetto fisico" (Kirschenbaum 2013, 4). Nella lettura di Kenneth Thibodeau, ogni oggetto digitale è al contempo un oggetto fisico, un oggetto logico e un oggetto concettuale, e le sue proprietà a ciascuno di questi livelli possono essere significativamente diverse (Thibodeau 2002, 3)[116]. Tale rapporto si manifesta, dunque, attraverso una stratificazione di livelli di astrazione:

- **Livello fisico**: la memorizzazione dei dati consiste in sequenze di bit codificate su supporti hardware (come settori magnetici, celle a stato solido ecc.). Come oggetto fisico, un oggetto digitale rappresenta essenzialmente un'iscrizione di segni su un medium: una serie di convenzioni definiscono la relazione tra il sistema di segni e il supporto fisico (Thibodeau 2002). Tali convenzioni variano a seconda del medium: esistono differenze evidenti tra la registrazione su dischi magnetici e su dischi ottici, e persino all'interno dello stesso tipo di supporto. Ad esempio, nei dischi rigidi (HDD) ogni bit corrisponde all'orientamento magnetico di domini microscopici, letti tramite sensori GMR (Ennen et al. 2016); nella DRAM, il bit è determinato dalla carica elettrica di un condensatore (Rana 2024); nelle memorie Flash NAND, il bit è definito dalla quantità di elettroni intrappolati nel gate flottante di un transistor, manipolato tramite *tunneling* (Bez et al. 2003). In ciascun caso, i bit occupano uno spazio fisico definito, ma il sistema di memorizzazione non conosce il loro significato: l'iscrizione fisica in sé non comporta morfologia,

[116] La tripartizione proposta da Kenneth Thibodeau può essere accostata, per analogia, al modello OSI (Open Systems Interconnection) delle reti di comunicazione (International Organization for Standardization 1989): il livello fisico dell'oggetto digitale corrisponde al Physical Layer (livello 1) di OSI; il livello logico ai livelli 2-4, che strutturano i dati; e il livello concettuale/informativo ai livelli 5-7, deputati alla presentazione e all'interfaccia applicativa.

sintassi o semantica, e non distingue se i bit rappresentino un documento, un'immagine o qualsiasi altro tipo di informazione (Thibodeau 2002, 3–4).

- **Livello logico**: i bit organizzati in file e cartelle assumono significato secondo la logica del software applicativo. Questi elementi, con attributi come nome, dimensione, permessi e *timestamp*, offrono una rappresentazione della materialità fisica. A questo livello, una serie di regole definisce come l'informazione è codificata, interpretata e trasformata tra formati, indipendentemente dal supporto fisico. Ciò che distingue il livello logico non è la collocazione fisica dei dati, ma la capacità di ricostruire correttamente l'informazione così come è stata concepita dall'applicazione. Inoltre, l'accesso e l'interpretazione dell'oggetto logico richiedono software in grado di rispettare le regole del tipo di dati, assicurando che l'informazione venga correttamente ricostruita indipendentemente dalla struttura o dal supporto fisico su cui è memorizzata (Thibodeau 2002, 4–5).
- **Livello informativo/concettuale**: riguarda l'interpretazione del contenuto informativo dei file (testo, immagini, codice, audio, ecc.). L'accesso richiede strumenti applicativi in grado di decodificare la struttura logica e presentare l'informazione in forma comprensibile. Un oggetto concettuale rappresenta un'entità significativa nel mondo reale, riconoscibile come unità di informazione (Thibodeau 2002, 5–6). Si tratta di un livello relativo al contenuto informativo: si riferisce a ciò che l'oggetto contiene o di cui tratta ed è intrinseco a un oggetto informativo (Gilliland 2016).

Questa natura del documento digitale, che è dunque simultaneamente materiale e intangibile, non è di immediata comprensione, a causa di quella che Jean-François Blanchette definisce l'«immaterial trope» (Blanchette 2011, 1043), una retorica dell'immaterialità che ha plasmato per decenni l'immaginario collettivo sul digitale, propagandando l'idea dell'informazione digitale come essenzialmente incorporea e svincolata dalla materialità (Sundqvist 2021, 1).

Secondo Sherry Turkle, questo paradigma ha iniziato a consolidarsi a partire dall'introduzione del Macintosh negli anni Ottanta (Turkle 1995, 35). Il suo design offriva all'utente «a scintillating surface on which to float, skim, and play» (Turkle 1995, 34), favorendo così la possibilità di gestire e generare informazioni senza dover necessariamente comprendere i meccanismi sottostanti. Questa caratteristica ha posto le basi per quella che Turkle definisce una visione "postmodernista", in cui il computer appare come una tecnologia opaca (Turkle 1995, 23): la conoscenza tecnica, non più necessaria al suo utilizzo, tende progressivamente a scomparire insieme alla consapevolezza della fisicità del digitale. A rafforzare

questo processo hanno contribuito i fenomeni di miniaturizzazione dei computer, l'introduzione di tecnologie come gli schermi LCD (Simon 2007) e i successivi avanzamenti nel design delle interfacce. In questo senso, Kirschenbaum sottolinea come: «computers are unique in the history of writing technologies in that they present a premeditated material environment built and engineered to propagate an illusion of immateriality» (Kirschenbaum 2007, 135).

Da un lato, effettivamente, il contenuto di un file può essere inteso come indipendente dalla sua materialità fisica: la stessa informazione può essere conservata su un disco magnetico, su un SSD, in un server remoto o su un cloud, senza che ciò comporti una trasformazione del suo contenuto informativo. La possibilità di copiare, trasferire e replicare un file senza perdita d'informazione dimostra che il suo contenuto è distinto dal mezzo che lo ospita. Tuttavia, pur essendo un'entità indipendente da un singolo supporto, il contenuto rimane l'interpretazione di una sequenza di bit organizzata secondo un formato e una codifica, che necessita di essere memorizzata su un supporto, per quanto esso sia intercambiabile.

La tensione tra queste dimensioni si riflette in modo emblematico, ad esempio, nel ruolo dei formati dei file. I formati non sono meri contenitori neutri, ma dispositivi che definiscono modalità specifiche di organizzazione dei dati. In questo senso, introducono vincoli e possibilità che incidono sull'accessibilità, sulla manipolazione e sulla preservazione dell'informazione digitale (Kirschenbaum 2007, 135-49). I sistemi operativi e le applicazioni identificano il formato di un file attraverso diversi meccanismi: analizzando le informazioni contenute al suo interno, interrogando i metadati registrati nel *file system* oppure, più semplicemente, basandosi sull'estensione del nome del file (Trace 2011, 24). I formati si collocano tra la materialità dei bit e la visualizzazione dell'informazione e, in questo ruolo intermedio, traducono sequenze di segnali fisici in contenuti potenzialmente intelligibili. In altre parole, il formato non solo determina come i dati sono organizzati e interpretati, ma influisce direttamente su quali aspetti dell'informazione vengono conservati, modificati o persi. Ad esempio, la scelta di un algoritmo di compressione lossy, come nel caso del formato JPEG, costituisce un compromesso deliberato tra fedeltà dell'informazione ed efficienza di storage (Hudson et al. 2017, 25). Diversi livelli di compressione possono produrre output visivamente simili all'occhio umano, ma da un punto di vista informatico e informativo generano entità profondamente diverse, poiché molti dati originari vengono eliminati in modo irreversibile (Blanchette 2011, 1045). Il formato determina anche le modalità di interazione con l'informazione: una riproduzione TIFF di un documento rimane non interpretabile per un motore di ricerca testuale, mentre la stessa immagine processata tramite *optical character recognition* (OCR) e salvata in un formato testuale strutturato diventa ricercabile, analizzabile e manipolabile in modi radicalmente differenti (Blanchette 2011, 1045). Va rilevato che oggi molti sistemi operativi e dispositivi

mobili integrano funzionalità OCR automatiche (come la funzione Live Text introdotta da Apple con iOS 15[117] o le capacità di Google Cloud con IA) che permettono la ricerca testuale direttamente nelle immagini senza modificare il formato del file originale, estraendo e indicizzando il testo dinamicamente. Il principio epistemologico, tuttavia, resta invariato: anche in questo caso è necessaria un'operazione interpretativa e computazionale, l'OCR, che traduce il contenuto visivo in informazione testuale interrogabile. L'accesso e la manipolazione dell'informazione dipendono infatti non solo dal formato del file ma dalla totalità di quei processi che Vallverdú i Segura invita a leggere sotto la lente della *Computational Epistemology*, sostenendo come l'interazione con l'informazione digitale sia sempre mediata da operazioni di elaborazione che ne determinano le possibilità d'uso (Vallverdú i Segura 2009).

Inoltre, la stabilità a lungo termine di un file dipende dalla sopravvivenza dell'ecosistema tecnico che ne consente l'interpretazione. Formati proprietari o obsoleti rischiano di diventare illeggibili, trasformando sequenze binarie formalmente integre in dati sostanzialmente inutilizzabili (Allegrezza 2024). In questo senso, l'inaccessibilità tecnica del "contenitore" entra in conflitto con la necessità archivistica di garantire l'accesso ai contenuti informativi.

Infatti, l'accesso e la fruizione del contenuto sono sempre e inevitabilmente radicati in una complessa infrastruttura hardware e software. L'indipendenza informativa del file è una condizione puramente potenziale che per attualizzarsi richiede un ecosistema tecnologico specifico. Come sottolinea Kirschenbaum, infatti, il bit possiede una sua ineliminabile "materialità", che si articola su due livelli: una "materialità forense", legata alla sua iscrizione fisica su un supporto (ad esempio, l'orientamento magnetico di un dominio su un hard disk), e una "materialità formale", concernente gli apparati computazionali necessari a interpretarlo (Kirschenbaum 2007, 10-15). Senza un hardware in grado di leggere quel supporto, un sistema operativo che gestisca il *file system*, un software applicativo capace di decodificare il formato specifico e, non ultimo, l'energia elettrica per alimentare il tutto, la sequenza di bit rimane illeggibile. L'accesso al significato è mediato da una catena di componenti materiali, ciascuna distinta ma inseparabile dall'insieme: dai bit iscritti sul supporto fino al dispositivo con cui l'utente legge il dato, ogni elemento contribuisce in modo imprescindibile alla fruizione dell'informazione. Il documento è dunque intangibile in quanto mediato da astrazioni digitali; mediazione che, però, non elimina la materialità, piuttosto la rende invisibile, celandola in architetture stratificate e distribuite.

Questa dipendenza infrastrutturale non è soltanto sincronica, ossia legata alla coesistenza di diversi ecosistemi tecnologici, ma anche profondamente diacronica: l'evoluzione continua di hardware,

[117] https://support.apple.com/en-il/guide/iphone/iph37fdd714b/ios.

software, formati e supporti introduce un fattore temporale cruciale. Leggere e interpretare file conservati su floppy disk degli anni Ottanta, infatti, richiede contesti radicalmente diversi rispetto all'accesso a documenti prodotti con sistemi più recenti.

Un ulteriore contesto in cui il contenuto digitale mostra chiaramente la propria dipendenza dalla materialità è il fenomeno della *data remanence*, ossia la persistenza di tracce informative anche dopo tentativi di cancellazione (Skorobogatov 2005; Aissaoui et al. 2017; Shu et al. 2017). Questo fenomeno si verifica quando rimangono residui di dati su un supporto di memorizzazione nonostante le operazioni di cancellazione standard, come la rimozione di file o la formattazione del dispositivo (Kirschenbaum 2007, 60). Tali tracce possono essere recuperate utilizzando tecniche forensi avanzate, rappresentando un rischio per la sicurezza e la privacy delle informazioni sensibili (Bellekens et al. 2015; Ries 2018). Per contrastare la *data remanence*, laddove essa costituisca un rischio per la sicurezza, sono state sviluppate diverse tecniche, classificate nelle tre categorie principali di *shredding*, *degaussing* e *physical destruction*, la cui scelta dipende strettamente dalle caratteristiche fisiche del supporto e dal livello di sicurezza richiesto (Conrad et al. 2010).

Se dal punto di vista della sicurezza informatica la *data remanence* costituisce una criticità da neutralizzare, in ambito archivistico e filologico essa rappresenta invece una risorsa metodologica di straordinario valore. Proprio la persistenza di queste tracce consente infatti di recuperare versioni cancellate e varianti testuali, restituendo così la stratificazione del processo creativo e permettendo di ricostruire la genesi dei testi con un livello di dettaglio altrimenti irraggiungibile (Ries 2018; Crasson et al. 2025). Tuttavia, questa potenzialità solleva inevitabili questioni etiche: il recupero e l'analisi di materiali che l'autore ha deliberatamente tentato di eliminare impone una riflessione sulla legittimità di tali pratiche e sulla necessità di definire protocolli deontologici che bilancino l'interesse della ricerca con il rispetto della volontà autoriale.

4.2 Metadati

I metadati sono una forma di descrizione archivistica, che a sua volta è una forma di rappresentazione delle informazioni (Pacheco et al. 2023). Prima di affrontare il tema nel contesto nativo digitale, è necessario comprendere cosa si intenda in questo contesto per metadati, poiché il concetto è impiegato in molte discipline con interpretazioni differenti. Furner (2020), ad esempio, ha rilevato che esistono 96 standard ISO che offrono 46 diverse definizioni del termine. Pacheco, Guardado Da Silva e Vieira De Freitas (2023) presentano una ricostruzione articolata della storia e delle diverse accezioni attribuite al termine "metadato": la prima attestazione, secondo Furner, risale al 1968, quando il *computer scientist*

Philip Bagley scriveva che uno degli elementi fondamentali di un linguaggio di programmazione è «the ability to associate explicitly with a data element a second data element which represents data "about" the first data element. This second data element we might term a "metadata element"» (Bagley 1968, 223).

Nella letteratura archivistica vi sono opinioni contrastanti su cosa possa essere considerato metadato. Seguendo l'interpretazione più ampia, i metadati possono essere intesi come qualunque dato strutturato che descriva proprietà di altri dati (Gladney 2007, 7). Nel contesto della descrizione di risorse digitali, la Digital Preservation Coalition (DPC) individua come elementi costitutivi la presenza di una struttura formale e, conseguentemente, la loro idoneità a essere elaborati automaticamente, definendoli «data about a digital resource that is stored in a structured form suitable for machine processing» (Digital Preservation Coalition 2024). In questa prospettiva, la definizione della DPC si ricollega alla letteratura che identifica in contesto e struttura caratteristiche essenziali dei record (Duranti 1997; US National Archives 2016) le quali inevitabilmente si riflettono anche nei metadati che li rappresentano.

Nelle definizioni finora discusse, i metadati sono tendenzialmente concepiti come il prodotto di operazioni di descrizione archivistica. Questo modello rimane applicabile anche al contesto digitale, ma con una differenza sostanziale: i file sono accompagnati fin dalla loro creazione da metadati generati nativamente, che si accumulano e si aggiornano lungo l'intero ciclo di vita del documento. Pur non riportando informazioni canoniche, è fondamentale valorizzare questi metadati, perché costituiscono il nuovo humus informativo e contestuale dei documenti digitali. La loro presenza introduce complessità inedite rispetto alla tradizione analogica, poiché alla curatela descrittiva si affianca una componente generata da sistemi, applicazioni e utenti nelle fasi precedenti. Ne risulta un insieme eterogeneo di informazioni, derivato dall'analisi di caratteristiche differenti dello stesso oggetto digitale, che costituisce una sfida centrale per la modellazione: distinguere, correlare e interpretare correttamente metadati di diversa origine.

Esistono varie tipologie di metadati nativi dei documenti digitali, tra cui i metadati di sistema e i metadati *embedded* (The Sedona Conference 2013), due categorie che operano a livelli diversi dell'architettura digitale e seguono logiche di generazione e aggiornamento completamente differenti. Ciò che caratterizza entrambe le tipologie è la notevole varietà di valori, formati e codifiche che possono assumere a seconda dell'ambiente tecnologico in cui vengono generati, gestiti e interpretati.

I metadati di sistema, o *file system*, sono informazioni generate e gestite automaticamente dal sistema operativo per ogni file o cartella presente nel sistema: data di creazione, di ultima modifica e di ultimo

accesso, attributi relativi alla posizione nel *file system* e permessi di accesso (Larry e Lars 2011, 180). Questi metadati sono registrati e mantenuti dal *file system* stesso, indipendentemente dal contenuto specifico del file.

Pur apparendo come informazioni standardizzate, i metadati di sistema presentano in realtà variazioni determinanti che riflettono la diversità degli ecosistemi tecnologici. Un esempio emblematico è rappresentato dalle informazioni temporali dei file. I sistemi operativi moderni sono conformi allo standard POSIX[118] e gestiscono le informazioni temporali dei file attraverso tre *timestamp* fondamentali (IEEE/Open Group Std 2024): `mtime` (*modification time*) che documenta l'ultima modifica del contenuto del file secondo la prospettiva del *file system*; `atime` (*access time*) che registra l'ultimo accesso in lettura al contenuto del file; `ctime` (*change time*) che traccia l'ultima modifica dei metadati del file stesso, distinguendosi dalle modifiche al contenuto.

Il loro funzionamento, tuttavia, non è universale e dipende dalle opzioni di *mount* del *file system* (Kerrisk 2010, 165-67) e dal sistema operativo. Ad esempio, il *file system* NTFS memorizza i valori temporali in formato UTC, risultando quindi indipendente da cambiamenti di fuso orario o dall'ora legale, mentre il *file system* FAT conserva i valori temporali in base all'ora locale del computer. Ad esempio, un file creato alle 15:00 (Pacific Standard Time, PST) a Los Angeles corrisponde alle 23:00 UTC. Un volume NTFS memorizza questo valore in UTC e lo converte automaticamente nel fuso locale del sistema che lo visualizza; di conseguenza, lo stesso file apparirà con orario 18:00 (Eastern Standard Time, EST) se aperto a New York. Un volume FAT, invece, registra direttamente l'orario locale (15:00) senza riferimenti all'UTC, rendendo il *timestamp* dipendente dal contesto in cui è stato creato (Microsoft Corporation 2021; Bouma et al. 2023). Inoltre, non tutti i *file system* sono in grado di memorizzare i tempi di creazione e di ultimo accesso, né lo fanno con la stessa granularità. Ad esempio, nei *file system* FAT16 o FAT32 la risoluzione del tempo di creazione è pari a 10 millisecondi, quella del tempo di scrittura a 2 secondi, mentre l'ultimo accesso è registrato con una risoluzione giornaliera. Diversamente, in NTFS l'aggiornamento del tempo di ultimo accesso può essere posticipato fino a un'ora dopo l'ultimo effettivo accesso e la risoluzione del *timestamp* può arrivare sino a 100 nanosecondi (Microsoft Corporation 2021; Bouma et al. 2023).

Parallelamente, un'altra dimensione significativa è rappresentata dai metadati *embedded*, ossia metadati incorporati direttamente nei file dal software che li ha generati o modificati, oppure aggiunti manualmente dall'utente (Larry e Lars 2011, 181-84). Essi comprendono testi, numeri o altre

[118] https://pubs.opengroup.org/onlinepubs/9799919799/nframe.html

informazioni inserite in modo diretto o indiretto all'interno del file, e che di norma non sono visibili nella visualizzazione standard del documento da parte dell'utente (The Sedona Conference 2013, 174).

La varietà e ricchezza di questi metadati dipende strettamente dalla tipologia di documento e dal software utilizzato. Nei documenti testuali DOCX, ad esempio, questi includono formato del file, tipo MIME, codifica dei caratteri e lingua, data di creazione originaria, identificazione del creatore e dell'ultimo agente che ha modificato il file, software utilizzato per la generazione o modifica, registro delle revisioni effettuate, tempo di lavoro, numero di pagine, parole, caratteri, ecc. (ECMA International 2016). Analogamente, nelle immagini JPEG i metadati EXIF possono includere informazioni tecniche sulla fotografia come modello e marca della fotocamera, impostazioni di scatto (ISO, apertura, tempo di esposizione), data e ora di acquisizione, coordinate GPS della geolocalizzazione, oltre a dati aggiunti dal software di editing utilizzato per modificare l'immagine (Library of Congress 2023; Canon s.d.). Ogni applicazione può aggiungere metadati specifici secondo i propri standard e convenzioni, creando un panorama informativo estremamente variegato.

L'estrazione e l'interpretazione di metadati *embedded* e di sistema è tradizionalmente avvenuta nell'ambito dell'informatica forense (*digital forensics*), per scopi investigativi e legali (Fernando 2021; Oh et al. 2022; Balkibayeva 2024). Tuttavia, il loro valore in termini archivistici come risorsa potenzialmente utile per la conservazione, la descrizione e l'analisi critica dei materiali nativi digitali, è pienamente da valorizzare (Rogers 2019). La novità, rispetto all'analogico, si riscontra anche nella loro quantità, determinando una densità di informazioni a livello documento senza precedenti. Come osservava già Margaret Hedstrom nel 1993, «in the electronic era, the descriptive paradigm will shift from the current practice of augmenting scarce descriptive information to one of selecting from an abundance of metadata» (Hedstrom 1993, 59). Questa abbondanza solleva questioni strategiche: se da un lato ci possono essere dubbi sull'utilità di conservare tutti questi metadati a fini di preservazione (essendo potenzialmente ri-estraibili con nuovi e sempre più sofisticasti strumenti) e sull'eticità di questa scelta (in considerazione delle implicazioni della sostenibilità della *digital preservation*) (Pendergrass et al. 2019) , dall'altro essi rappresentano una risorsa fondamentale per la descrizione archivistica. La presenza di metadati nativi può velocizzare e, in parte, sostituire la curatela manuale, e il loro livello di dettaglio consente di ripensare le funzioni stesse della descrizione nel contesto digitale. Infatti, come evidenzia una nota della DPC a cura di Jenny Bunn, la definizione dei bisogni informativi si sta formalizzando verso un livello sempre più dettagliato, riflettendo l'adattamento ai sistemi informatici e le crescenti aspettative degli utenti per il trattamento automatico dei metadati archivistici per la ricerca (Bunn 2021). In questo senso, la granularità dei dati permette approcci computazionali che inaugurano

nuove possibilità di ricerca, come l'applicazione di tecniche di *distant reading*, un approccio teorizzato da Franco Moretti per analizzare la letteratura attraverso strumenti quantitativi e computazionali, studiando migliaia di testi simultaneamente invece di singole opere (Moretti 2000). Contrapposto al tradizionale *close reading*, il *distant reading* utilizza grafici, mappe e analisi statistiche per rivelare pattern e tendenze letterarie su vasta scala ed è divenuto quadro di riferimento metodologico per una ingente quantità di studi e sperimentazioni (Underwood 2019; Gavin 2022; Jockers 2023). Il *distant reading*, così come l'applicazione di modelli AI, possono essere utilizzati per identificare pattern e strutture latenti che sfuggirebbero a occhio nudo nei dati archivistici (Grandjean 2016; Donig et al. 2023), così come per migliorare le operazioni di esplorazione, ricerca e accesso alle risorse (Colavizza et al. 2021; Jaillant e Caputo 2022; Di Marcantonio 2024; Jaillant e Zhao 2025).

Occorre considerare, inoltre, che il portato informativo di un metadato non si esaurisce al dato che veicola, ma comprende anche la sua origine e le modalità con cui è stato generato o estratto. I file, infatti, vanno concepiti come oggetti intrinsecamente dipendenti da un «whole ecosystem of devices and code, from the manufacture of the screen itself to the metadata underlying the operating system» (Levy 2022). Questa interdipendenza sistemica solleva il problema di come rappresentare non solo i metadati in sé, ma anche le loro relazioni con gli ecosistemi tecnologici con cui sono venuti a contatto, perché presenterano informazioni codificate secondo standard e convenzioni diverse a seconda dell'ambiente in cui sono stati generati, gestiti ed estratti.

4.3 Provenance

Il principio di provenienza (*provenance*) costituisce uno dei fondamenti teorici dell'archivistica moderna, che continua ad evolvere (Douglas 2010). La sua apparente linearità, che prevede di mantenere insieme i documenti prodotti da un soggetto produttore e di rispettarne l'ordinamento originale, ha subito profonde revisioni critiche. Già nel contesto europeo del XIX secolo, come documenta Lodolini (1981), emergeva una tensione tra la necessità di ricostruire l'ordine originario delle carte e le esigenze pratiche di riordinamento, evidenziando come il principio non possa essere applicato meccanicamente ma richieda sempre un'interpretazione contestuale. Tuttavia, la tradizione archivistica per lungo tempo è rimasta vincolata ad visione monodimensionale della *provenance*, che tende a semplificare i processi di creazione e gestione documentale privilegiando relazioni lineari tra un singolo creatore e una serie chiusa e completa di documenti (Bak 2024, 848). Terry Cook nel 1993 critica questa impostazione, osservando che una relazione così diretta e netta raramente riflette la complessità reale dei flussi documentali, nei quali spesso intervengono molteplici attori in momenti diversi (Cook 1993). Secondo Cook, che si basa

anche sui lavori di Scott (1966) e Barr (1987), la *provenance* è frequentemente multipla e prolungata nel tempo, coinvolgendo più soggetti e contesti nella produzione e gestione dei documenti (Cook 1993, 33). Nel glossario allegato ad ISAD(G), l'ICA estende il concetto di *provenance* definendola come «the relationship between records and the organizations or individuals that created, accumulated and/or maintained and used them in the conduct of personal or corporate activity» (2000, 11), sottolineando il legame tra i documenti e i soggetti che li hanno prodotti, accumulati, custoditi o utilizzati nell'ambito delle proprie attività. Tom Nesmith (2002) amplia questa prospettiva, descrivendo la *provenance* come l'intera storia del documento, nozione che include anche i successivi custodi e le azioni, ad esempio quelle degli archivisti, che hanno inciso sul ciclo di vita dei materiali.

Nel contesto digitale, le tracce della storia del fondo si amplificano ulteriormente, richiedendo un ripensamento che tenga conto delle specificità tecnologiche e delle nuove forme di mediazione che caratterizzano il digitale (Bak 2024). Possiamo iniziare a individuare alcuni snodi a partire da questo presupposto: nei documenti digitali, la possibilità di leggere e interpretare l'informazione non dipende più solamente dal supporto fisico, ma richiede l'intervento di strati tecnologici che, a partire da un supporto, mediano l'accesso ai dati. Nei documenti analogici, la leggibilità è una proprietà mediata direttamente dal supporto, che veicola il contenuto informativo in modo immediatamente percettibile dai sensi umani. L'accesso all'informazione (sebbene non la sua comprensione) è condizionato esclusivamente dal degrado fisico-chimico che minaccia la sopravvivenza materiale del supporto, senza richiedere apparati interpretativi aggiuntivi. Al contrario, nei documenti digitali l'informazione, pur inscritta in forma fisica, si presenta come codice binario, un substrato opaco per l'uomo senza la mediazione di molteplici livelli tecnologici. *File system*, sistemi operativi e software applicativi operano come filtri interpretativi e non sono né neutrali né trasparenti, applicando ciascuno le proprie regole e convenzioni che si sovrappongono in livelli di astrazione successivi. L'utente non accede mai al documento digitale nella sua forma fisica e binaria, ma a una sua rappresentazione condizionata dalle capacità degli strumenti impiegati. Poiché la lettura e l'interpretazione diretta del codice binario risultano per noi impossibili, dipendiamo completamente da questi mediatori per accedere all'informazione.

Questa dipendenza rende ancora più evidente una questione che la disciplina archivistica ha dovuto affrontare anche in ambito analogico: l'impossibilità di una descrizione completamente oggettiva e neutrale[119]. Come osserva Elizabeth Yakel, «through the process of selection of information for inclusion

[119] Una nuova tendenza negli studi archivistici ricontestualizza la soggettività dell'archivista come risorsa epistemica: anziché aspirare a un'imparzialità ormai riconosciuta come "untenable, sometimes harmful" (King 2024, 512), si valorizza la

and choice of access points, archivists reveal and conceal, making finding aids political statements» (2011, 19). Similmente, Richard Gartner afferma: «there is nothing objective about metadata: it always makes a statement about the world, and this statement is subjective in what it includes, what it omits, where it draws its boundaries and in the terms it uses to describe it» (2016, 4).

Se per descrivere l'analogico l'archivista compie scelte soggettive nella selezione e rappresentazione delle informazioni, nel contesto digitale l'interpretazione si moltiplica attraverso i diversi strati tecnologici che si frappongono tra il documento, la sua fruizione e la sua analisi da parte dell'archivista (Tomasi 2022). Ad esempio, contrariamente alle aspettative di oggettività che potrebbero derivare dall'automazione, l'estrazione dei metadati tramite strumenti software costituisce a tutti gli effetti un'operazione interpretativa, in quanto soggetta alle specificità algoritmiche e implementative di ciascun tool utilizzato (Gorini e Giagnolini 2025). Tra gli strumenti *open source* più utilizzati per l'estrazione automatica di metadati vi sono: Apache Tika[120], un *framework* per l'estrazione di contenuti e metadati da una vasta gamma di formati; ExifTool[121], specializzato nella lettura e scrittura di metadati EXIF, IPTC, XMP e altri; MediaInfo[122] e FFmpeg[123], principali strumenti per l'analisi e la manipolazione di file multimediali audio e video; DROID[124], sviluppato da The National Archives del Regno Unito e impiegato principalmente per l'identificazione dei formati digitali; infine, il modulo os di Python[125] e le utility di sistema come `stat` nei sistemi Unix-like, che consentono l'accesso programmatico ai metadati dei file tramite funzioni di interrogazione del *file system*. Gli esiti dell'estrazione possono variare anche significativamente a livello di valore del metadato quando si applicano tool diversi allo stesso file, generando conflittualità informativa che deve essere adeguatamente contestualizzata (Kulmukhametov et al. 2021). Ciò si spiega in considerazione del fatto che il medesimo valore può essere ottenuto dai

posizionalità dell'archivista intesa come l'insieme di prospettive, formazione personale, identità sociale e punto di vista che inevitabilmente influenzano l'interpretazione del materiale archivistico. Questa consapevolezza, supportata dalla teoria della "situated knowledge" di Haraway (Haraway 1988, 581) e del "Feminist Standpoint Appraisal" di Caswell (Caswell 2021, 1), riconosce che certe intuizioni sono possibili solo da specifiche prospettive marginali, mentre posizioni dominanti e privilegiate possono nascondere o distorcere informazioni cruciali. La proposta metodologica consiste nel sostituire l'ideale misconosciuto dell'imparzialità con la trasparenza sulla posizionalità dell'archivista (che deve rimanere pur sempre critica), creando quello che King definisce "archival meta-metadata" (2024, 510): informazioni su autorialità, revisioni e contesto di produzione degli strumenti di ricerca. Tale approccio permette di esplicitare quelle dimensioni contestuali e interpretative che una presunta neutralità tenderebbe a occultare, impoverendo così sia l'informazione disponibile agli utenti sia la capacità di valutare criticamente le descrizioni stesse nel loro divenire storico.

[120] https://tika.apache.org/.

[121] https://exiftool.org/.

[122] https://mediaarea.net/en/MediaInfo.

[123] https://ffmpeg.org/.

[124] https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

[125] https://docs.python.org/3/library/os.html.

singoli tool attraverso strategie di recupero fondamentalmente distinte: la dimensione del file può essere determinata interrogando il *file system* attraverso il campo `st_size` dell'*inode* oppure leggendo l'*header* del documento (*Content-Length*) (come fa Apache Tika). Analogamente, la data di creazione viene estratta da strutture completamente diverse a seconda del formato: nei file JPEG viene letta dal campo *CreateDate* nell'*header* EXIF, nei documenti DOCX dagli elementi XML, nei PDF dal dizionario Info del documento, richiedendo ciascuna un parser specifico e specializzato (Harvey 2003; Apache Software Foundation 2024). Oppure la determinazione della durata di un video: strumenti come MediaInfo estraggono il valore direttamente dai campi di metadata del *container* (es., MP4, AVI), che forniscono una durata pre-calcolata e spesso approssimata. Al contrario, strumenti come FFmpeg, calcolano la durata analizzando l'intero flusso video, frame by frame, determinando il valore massimo tra tutti gli stream presenti, il che può portare a risultati più accurati ma computazionalmente costosi (FFmpeg Team, 2024; MediaInfo, 2024). Ciascuno strumento, inoltre, può avere regole interne aggiuntive per la normalizzazione dell'informazione estratta, che comportano ulteriori differenze di rappresentazione finale.

La comprensione delle logiche di estrazione e normalizzazione risulta quindi cruciale per interpretare correttamente le informazioni e ricostruire la storia dei documenti digitali. Lo studio effettuato assieme ad Adele Gorini sugli archivi di Silvia Avallone e Paolo Di Paolo, conservati presso il Centro Manoscritti di Pavia nell'ambito del progetto PAD, ha mostrato come sia fondamentale disporre di più punti di vista e strumenti per contestualizzare correttamente le discrepanze nei metadati relative a dimensioni temporali, autoriali e testuali (Gorini e Giagnolini 2025).

Nel caso dell'Archivio Avallone, ad esempio, la gran parte dei documenti presentava date di modifica estratte da Apache Tika e ExifTool che differivano dai corrispondenti *timestamp* del *file system* di esattamente due ore. Sebbene siano note le possibili discrepanze tra *file system* metadata e *embedded* metadata (dovute, ad esempio, a operazioni come download, copia o trasferimento dei file che modificano la posizione del documento senza alterarne il contenuto), la differenza esatta di due ore ha suggerito che la causa più plausibile risiedesse nella gestione dei fusi orari. In particolare, i metadati estratti da strumenti da Apache Tika e ExifTool sono normalizzati in *Coordinated Universal Time* (UTC), mentre i metadati del *file system* (ad esempio FAT) tendenzialmente riflettono l'orario locale del sistema al momento della creazione o modifica del file. Durante il periodo estivo, in Italia vige il Central

European Summer Time (CEST, UTC+2) e i valori locali risultano due ore avanti rispetto all'UTC, spiegando il delta osservato tra le due fonti[126] (Gorini e Giagnolini 2025).

Un'altra discrasia osservata nello studio riguarda i metadati relativi al tempo totale di lavoro sui file in formato DOC. La documentazione di Word non specifica chiaramente come questo valore venga calcolato, e i dati estratti possono apparire di difficile interpretazione. La differenza tra Apache Tika ed ExifTool è emblematica: il primo restituisce valori grezzi in tick (unità di 100 nanosecondi), senza conversione esplicita in secondi o minuti e senza che l'unità di misura sia specificata, mentre il secondo converte automaticamente i tick in unità leggibili. Ad esempio, il file "aggiunte.doc" dall'archivio Di Paolo mostra un tempo di editing di 600.000.000 tick in Apache Tika, mentre ExifTool lo riporta come 1 minuto, corrispondente correttamente alla definizione di tick di Microsoft Word (Gorini e Giagnolini 2025). Questo non rappresenta un errore di Apache Tika, piuttosto una differenza nell'interpretazione e nella normalizzazione del dato[127].

Questi casi evidenziano come l'analisi dei metadati nativi richieda un approccio multiprospettico, che integri strumenti diversi e tenga conto delle specificità del formato, del software di creazione e del contesto tecnologico. Tale analisi costituisce dunque un caso emblematico della dipendenza interpretativa dai contesti tecnologici e mostra come le rappresentazioni documentarie siano inevitabilmente situate entro specifiche infrastrutture tecniche, le cui logiche condizionano in maniera determinante la ricostruzione della storia dei materiali digitali. Alla luce di queste dipendenze, il principio archivistico della provenienza necessita una riconcettualizzazione che vada oltre i suoi confini tradizionali, verso una provenienza di tipo multidimensionale (Alfier 2017, 38). Secondo Greg Bak, che riprende la visione di Tom Nesmith (2002), si può intendere la *provenance* nel digitale «as the entire history of the record, a definition in which *provenance* is complex, multiple and broad, encompassing many actors including technology designers and creators, record creators and record keepers and archivists» (2024, 849). Per analizzare questa complessità, Bak declina la digital *provenance* in tre dimensioni: quella tecnica, che documenta le storie e le idee che informano il design di hardware e software; quella socio-tecnica, che cattura le conoscenze tacite e le pratiche delle comunità di utenti che

[126] Un'altra differenza rilevante nella registrazione delle date tra strumenti e *file system* riguarda la modalità di rappresentazione: Apache Tika ed ExifTool tendono a normalizzare i valori temporali in UTC, mentre i metadati di *file system* utilizzano *timestamp* Unix espressi come numeri in notazione scientifica. Ad esempio, il valore in notazione scientifica 1.75233369117012E9 corrisponde a 1752333691.17012 secondi dall'epoch Unix (1° gennaio 1970), ossia 12 luglio 2025, ore 15:21:31 UTC (17:21:31 CEST in Italia).

[127] Tali differenze non si riscontrano invece nei file DOCX, che utilizzano un formato XML, la cui apertura rende i metadati più leggibili e uniformi tra diversi strumenti di estrazione.

si formano attorno alle tecnologie; e quella sociale, che riconosce i significati culturali e simbolici che i sistemi digitali assumono all'interno di specifici contesti storici e sociali (Bak 2024). Tuttavia, la definizione di *provenance* va distinta dalla sua rappresentazione: la prima descrive la natura complessiva del fenomeno, mentre la seconda è una costruzione che la traduce in termini descrittivi che, per quanto avanzati, possono restituirne solo una parte (Bak 2024, 849). Poiché la *provenance* abbraccia l'intera storia dei documenti, gli archivisti sono chiamati a compiere scelte in parte situazionali, determinate dal mandato e dalle risorse dell'istituzione, dalla natura dei documenti e dalle comunità a cui essi sono destinati. In questa prospettiva, diventa possibile articolare quattro dimensioni principali della *provenance* da considerare nella rappresentazione dei materiali nativi digitali, nella consapevolezza che ulteriori componenti possono rivelarsi rilevanti in funzione dei mandati istituzionali, delle caratteristiche documentarie specifiche o delle esigenze delle comunità di riferimento:

1. **La storia conservativa del supporto fisico**. Anche nel digitale, definire la *provenance* comprende la ricostruzione della storia del supporto (come hard disk, SSD, nastri, ecc.) e la sua integrità, in continuità con quanto avviene negli archivi tradizionali. Per contestualizzare al meglio l'origine dei materiali non è sufficiente garantire l'integrità della sequenza di bit: occorre documentare la provenienza fisica e la catena di custodia dei supporti da cui i dati sono stati estratti. Come osserva Federico Valacchi:

   > gli spazi della memoria si sono delocalizzati, sono scivolati lungo il piano inclinato di geografie impalpabilmente binarie ma non si sfugge alla materialità ferrigna dell'hardware e dei supporti che ospitano i bit. L'idea di spazio non può essere abbandonata, non si può rinunciare alla percezione e alla tutela fisica di ciò che leggiamo, vediamo ed elaboriamo a partire da un'immaterialità solo presunta (2022, 3).

   Questa esigenza, che è definita *chain of custody* in termini forensi (Prayudi e Sn 2015) , implica il tracciamento del percorso di un oggetto digitale attraverso i diversi supporti e strati di astrazione, mantenendo un legame verificabile con i supporti originali che lo hanno veicolato. Documentare i supporti e la loro catena di custodia significa quindi non solo tracciare la storia archivistica dei file digitali, ma anche riconoscere la persistenza di una dimensione spaziale che resta essenziale per ancorare la memoria.

2. **Individuazione degli agenti coinvolti nel ciclo di vita del documento**. La provenienza digitale deve distinguere gli agenti responsabili delle attività che gravitano attorno al documento,

riconoscendo diversi livelli di agentività. Si può parlare di responsabilità umana diretta quando l'azione è compiuta intenzionalmente da una persona (o da un ente attraverso le persone che ne ricoprono le funzioni), ad esempio nella creazione del documento o nella redazione manuale di metadati o testi descrittivi; di agentività semiautomatica quando l'intervento automatico è mediato dall'uomo, come la revisione di un testo prodotto da un modello AI[128]; e di responsabilità pienamente automatica quando il processo è interamente demandato a procedure computazionali, come nel caso del calcolo degli hash. Questa distinzione è finalizzata a rappresentare le diverse modalità di generazione e trasformazione del documento digitale e per attribuirne correttamente le responsabilità.

3. **Documentazione dei processi effettuati sui materiali**. La necessità di documentare tutti gli asserti che vengono prodotti sui materiali digitali richiama il principio secondo cui «ogni azione va documentata o anche ogni interpretazione condotta sui dati va descritta» (Tomasi 2022, 89). La documentazione dei processi effettuati sui materiali digitali richiede, innanzitutto, la tracciabilità completa di hardware e software coinvolti nell'intero ciclo di vita del documento. Questa esigenza si estende ben oltre la semplice registrazione degli strumenti utilizzati, richiedendo la documentazione sistematica di tutti gli interventi curatoriali: dall'estrazione dei metadati tecnici al calcolo degli hash crittografici per la verifica dell'integrità, dalle operazioni di migrazione formato alla creazione di copie di preservazione. Ciascuna di queste attività introduce variabili che possono influenzare le proprietà osservabili delle risorse o la loro rappresentazione descrittiva, rendendo essenziale la tracciabilità degli agenti coinvolti, delle date di esecuzione, delle versioni software utilizzate, dei parametri operativi applicati, e della documentazione tecnica di riferimento.

   Parte di queste informazioni possono essere ricavate attraverso l'analisi dei metadati nativi, che contengono informazioni sul contesto di creazione dei file, sulle tecnologie impiegate e sulle eventuali modifiche subite. Tuttavia, i metadati nativi rappresentano solo una delle fonti informative necessarie: la loro stessa interpretazione dipende dagli strumenti di estrazione utilizzati. La possibilità di confrontare molteplici estrazioni effettuate con software diversi, di

[128] Un aspetto di crescente rilevanza nella pratica archivistica riguarda l'integrazione dell'IA nei sistemi archivistici, nella ricerca e nella produzione dei contenuti (Colavizza et al. 2021). L'impiego di tale tecnologia richiede lo sviluppo di *framework* descrittivi specializzati, in grado di documentare sia le specificità dei materiali prodotti sia i processi generativi che li hanno originati. La sistematicità di questa documentazione si configura come condizione necessaria per assicurare la trasparenza metodologica e permettere una valutazione critica dell'affidabilità delle informazioni raccolte (Kilbride 2024).

consultare la documentazione tecnica dei software di produzione dei documenti, e di accedere alle specifiche dei formati risulta cruciale per contestualizzare correttamente la provenienza dei dati. Ciò significa che anche i processi di estrazione e gestione dei metadati devono essere documentati con la stessa granularità riservata ad altri interventi curatoriali, poiché anch'essi introducono variabili che possono determinare quali informazioni vengono rilevate, come vengono normalizzate, e quali interpretazioni vengono applicate ai dati estratti.

La documentazione sistematica di questi processi costituisce la base epistemica per valutare l'affidabilità delle informazioni prodotte, per riprodurre gli interventi quando necessario, per comprendere le dipendenze tecnologiche che possono influenzare la preservazione a lungo termine, e per rendere esplicite le scelte curatoriali che, stratificandosi nel tempo, costruiscono la rappresentazione digitale degli archivi e ne determinano l'accessibilità futura.

4. **Documentazione dell'integrità.** Riprendendo Heather MacNeil, che ha definito la descrizione archivistica un «apparatus of authenticity» (2005, 271) e ne ha proposto l'uso come attestazione aggregata per l'autenticazione dei materiali nativi digitali, e in continuità con le indicazioni dell'Authenticity Task Force di InterPARES 1 (2002, 11), Michael Forstrom sottolinea come la descrizione non sia soltanto uno strumento di accesso e mediazione, ma anche una documentazione capace di certificare lo stato e l'integrità delle collezioni (Forstrom 2009). La descrizione assume il valore di un'attestazione dello stato dei materiali nel momento in cui viene redatta, includendo informazioni fondamentali come titolo, autore, acquisizioni e materiali correlati, oltre a eventuali migrazioni o riformattazioni. Nel digitale, questo principio si traduce nella necessità di integrare nella descrizione, in particolare, il dato dell'hash dei documenti, ovvero il risultato di un'operazione crittografica che genera un'impronta digitale univoca del file. Il valore così ottenuto, se ricalcolato in momenti successivi e confrontato con quello originario, consente di verificare l'integrità del documento: anche la minima alterazione del file, infatti, produrrebbe una modifica dell'hash. In questo senso, l'hash rappresenta un elemento tecnico imprescindibile per attestare e preservare l'autenticità dei materiali digitali nel tempo, configurandosi come parte integrante della *provenance* nella misura in cui documenta lo stato dei materiali al momento della loro descrizione.

4.4 Contesti

Tutti gli elementi fin qui trattati concorrono a definire i contesti archivistici. La documentazione delle caratteristiche degli oggetti e dei processi che li hanno generati costituiscono, nel loro insieme,

componenti fondamentali di tali contesti. Allo stesso tempo, il concetto di contesto può estendersi per includere relazioni e circostanze di ordine più ampio, quali quelle culturali e sociali che permeano l'intero ciclo di vita dei materiali.

La rilevanza dei contesti è riconosciuta nella disciplina da almeno trent'anni. Nel 2005 Nesmith osservava come gli archivisti abbiano vissuto un «pronounced contextual turn» (2005, 259), vale a dire «a movement toward a deeper appreciation of the role of contextual knowledge about records in archival work" (2005, 260). Tale svolta ha portato archivisti e ricercatori a collocare i documenti all'interno di cornici istituzionali, culturali e storiche sempre più ampie, suggerendo che i contesti sono virtualmente illimitati. A tale riguardo, Nesmith pone due interrogativi: «(1) What are its dimensions and characteristics? and (2) How may its features be incorporated into archival work?» (Nesmith 2005, 260).

Ketelaar sottolinea la dimensione storica del contesto archivistico, osservando che «the [contextual] history of records follows the logic of archives as transactional and process-bound information» (2023, 36), e identifica i contesti come «the why, who, what, and how of archiving, all determined by societal challenges and technologies» (2023, 36). In linea con questa prospettiva, Dilley lo definisce come «an articulation concerning a set of connections and disconnections thought to be relevant to a specific agent that is socially and historically situated, and to a particular purpose» (2002, 454).

La prospettiva della "societal provenance", teorizzata da Nesmith (2006), amplia ulteriormente il concetto di contesto archivistico includendo le dimensioni sociali, culturali e comunitarie che attraversano i confini istituzionali. In questa visione, i documenti non appartengono esclusivamente alla memoria di singole istituzioni, ma si inseriscono in cornici di memoria collettiva che richiedono approcci descrittivi capaci di valorizzare connessioni trasversali tra diverse tipologie di patrimonio. Similmente, il movimento dei *community archives* (Bastian 2020) sottolinea come le comunità stesse debbano essere riconosciute come attori nel processo di contestualizzazione, portando prospettive che arricchiscono e talvolta contestano le narrazioni istituzionali tradizionali.

In riferimento ai documenti elettronici, Duranti e Preston (2008, 13) individuano cinque contesti che concorrono alla loro creazione e gestione:

1. *provenancial*, relativo al soggetto produttore, al suo mandato, alla sua struttura e alle sue funzioni;
2. *juridical-administrative*, relativo al sistema legale e organizzativo in cui opera il soggetto produttore;
3. *procedural*, relativo al processo o alla funzione durante i quali il documento digitale viene generato;

4. *documentary*, relativo al fondo archivistico di appartenenza e alla sua struttura interna;
5. *technological*, relativo all'ambiente digitale in cui il documento è creato e mantenuto.

Con un approccio più diacronico, Zou propone una distinzione nelle tre dimensioni interagenti di contesto di creazione (*creation*), di descrizione (*description*) e di utilizzo (*use*) (Zou e Park 2024). Il contesto della creazione, in particolare, comprende l'ambiente originario in cui il documento è prodotto e utilizzato, includendo il produttore, il processo di creazione e il contesto giuridico-amministrativo e tecnologico in cui è avvenuto, nonché le dimensioni culturali e sociali richiamate dal concetto di "societal provenance" di Nesmith (2006). Il *framework* di Duranti e Preston può essere letto, in questo senso, come una articolazione dettagliata del contesto di creazione delineato da Zou, che ha il merito di mostrare come queste tre dimensioni non siano compartimenti separati, ma realtà intrecciate. Il modo in cui un documento viene creato condiziona le possibilità di descrizione; le scelte descrittive influenzano gli usi futuri; le pratiche di uso, a loro volta, retroagiscono sulle strategie descrittive e sulle interpretazioni della creazione. Si delinea così una concezione dinamica e circolare del contesto archivistico, che abbraccia la complessità dei documenti e restituisce alla disciplina la profondità e la ricchezza di significati del patrimonio documentario.

Francesca Tomasi sottolinea come, alla luce delle *digital humanities*, i tre pilastri fondamentali della concezione di contesto («soggetto produttore o *creator*, rispetto del principio di provenienza o *provenance*, posizionamento dell'entità nella gerarchia documentaria» (2022, 79)) si estendano e prevedano l'assunzione di nuovi punti di vista sulla rappresentazione, gestione e organizzazione della conoscenza. I primi due pilastri individuati da Tomasi, ossia i soggetti produttori e il principio di provenienza, sono stati trattati nei capitoli precedenti; resta da considerare la dimensione gerarchica. In questa prospettiva, il digitale reintroduce in modo concreto la struttura ad albero negli archivi, una configurazione dalla quale l'archivistica aveva progressivamente cercato di emanciparsi per via delle problematicità di una simile rappresentazione (Michetti 2009; Vilar e Šauperl 2017). Nell'ambiente digitale, tuttavia, l'albero non si limita a essere una convenzione descrittiva, ma diviene una struttura effettiva: le *directories* dei sistemi informatici organizzano i materiali secondo logiche gerarchiche, spesso molto profonde, con implicazioni dirette sulla creazione, sull'organizzazione e sull'accessibilità dei documenti.

In questo processo di sedimentazione, il principio di provenienza tradizionale viene messo alla prova. I record elettronici hanno molteplici potenziali "original orders": la questione centrale non è solo identificare quale sia l'ordine originario, ma comprendere quali funzioni esso assolve nelle diverse fasi

del ciclo di vita dei documenti (Zhang 2012). Un file può avere una posizione originaria nel computer del produttore, una posizione diversa al momento dell'acquisizione (all'interno dei supporti di copia e preservazione) e un'ulteriore posizione nella copia destinata alla consultazione. Inoltre, la visualizzazione dei contenuti una cartella può cambiare da sistema operativo a sistema operativo, e il produttore può avere preferenze di ordinamento (ad esempio, per data di ultima modifica, alfabetico, per tipologia di file) che non lasciano traccia stabile. Come evidenziato nel capitolo 2.3, nei contesti di archivi personali e di piccole organizzazioni, gli individui prendono decisioni personali e tendenzialmente non uniformi sull'organizzazione dei file, affidandosi a strutture costruite idiosincraticamente, oppure, volontariamente o involontariamente, accettando le convenzioni organizzative di default proposte dai sistemi operativi.

La struttura ad albero fornisce un contesto organizzativo immediato, ma non necessariamente il più adatto: come rilevato dall'indagine sulle abitudini gestionali degli scrittori presentata nel capitolo 2, la difficoltà nel ritrovare file personali anche sui propri computer testimonia l'inadeguatezza di questo contesto se preso isolatamente. Nel digitale, l'archivista dovrebbe tradurre la complessità in rappresentazioni che mediano tra diversi livelli contestuali: origine, funzione, tipologia, contesto istituzionale e dimensioni culturali. Le logiche tradizionali di serie e sottoserie, ad esempio, risultano spesso impraticabili per gli archivi di persona: le cartelle possono riflettere preferenze individuali o vincoli tecnici, e raggiungere profondità tali da rendere inapplicabile la terminologia canonica. Inoltre, assumerle come unico contesto archivistico solo perché "native" costituisce un fraintendimento metodologico, poiché, come abbiamo visto, non è necessariamente corretto supporre che quello sia l'ordine originario.

La centralità del concetto di contesto trova eco nella riflessione più ampia di Tomasi sul suo significato nell'ambito più esteso delle discipline umanistiche:

> non c'è disciplina che non trovi nella nozione di contesto la chiave di lettura e di interpretazione dei dati più appropriata per rispondere ai propri quesiti di ricerca. Un dato privo di contesto è un dato non altrimenti interpretabile, tanto sul piano computazionale quanto su quello umanistico, nelle sue declinazioni che qui ci interessano, come quelle linguistica, letteraria, biblioteconomica e archivistica (2022, 78).

In questa dimensione si inserisce la crescente convergenza tra archivi, biblioteche e musei, testimoniata dalle esperienze MAB (Musei, Archivi e Biblioteche) o GLAM (Galleries, Libraries, Archives, Museums). Questa convergenza non è meramente tecnica o organizzativa, ma risponde a una concezione

più ampia del patrimonio culturale come ecosistema interconnesso, in cui le tradizionali distinzioni disciplinari tendono ad attenuarsi. L'integrazione contestuale in ottica GLAM risponde a una duplice esigenza: da un lato, riconoscere che la comprensione piena di un documento spesso richiede l'accesso a risorse eterogenee conservate da tipologie diverse di istituzioni culturali; dall'altro, rispondere alle aspettative degli utenti contemporanei, abituati a navigare in modo fluido e interconnesso, senza essere vincolati dalle tradizionali segmentazioni disciplinari. Come avvertono Higgins, Hilton e Dafis: «there is a risk, if we do not examine the new ways in which readers consume resources in this environment, that our "correct" hierarchical catalogues will be nothing more than carefully curated silos of irrelevance» (Higgins et al. 2014, 14).

In risposta a questo quadro, l'avvento delle tecnologie del Web Semantico si sono affermate come prospettiva privilegiata, in quanto in grado di fornire «una serie di interconnessioni orizzontali e trasversali che servono ad arricchire l'esperienza conoscitiva» (Tomasi 2022, 94). In questa prospettiva, come sottolinea Tomasi, «Il contesto è tanto verticale (l'albero inteso in termini tassonomici), quanto orizzontale, (le relazioni di pari livello)», e il Semantic Web permette un «passaggio dall'albero strettamente gerarchico a un modello a grafo etichettato e tipizzato» in grado di valorizzare entrambe le dimensioni (2022, 94). Questa mediazione richiede strumenti descrittivi in grado di catturare contesti trasversali, permettendo all'utente di navigare materiali secondo percorsi multipli che collegano, ad esempio, documenti d'archivio con opere conservate in musei, edizioni librarie correlate e materiali visuali provenienti da diverse istituzioni (Bailey 2013; Niu 2015).

In ambito computazionale, in particolare, è proprio l'attribuzione di contesto che «consente ai dati di diventare informazione e all'informazione di diventare conoscenza» (Tomasi 2022, 78). Integrando queste sfaccettature in modelli flessibili e aperti all'integrazione interdisciplinare, l'archivista può fornire una lettura della complessità stratificata e relazionale del patrimonio culturale, restituendo ai materiali digitali la loro ricchezza di significati e inserendoli in reti di relazioni che riflettono più fedelmente la natura.

Definire quali contesti siano più rilevanti rimane onere dell'archivista, che può valutarli e valorizzarli a partire dalla conoscenza diretta dei materiali. La sfida consiste nel bilanciare il rigore metodologico proprio di ciascuna disciplina con l'apertura necessaria a creare connessioni tra domini diversi, riconoscendo che il contesto archivistico nel XXI secolo è necessariamente un contesto integrato, multiplo e dialogico.

5. La rappresentazione semantica nella pratica archivistica

Il capitolo 4 ha identificato le quattro dimensioni del digitale d'autore con implicazioni dirette sulla pratica descrittiva, impostando un orizzonte per rispondere alla RQ2: come rappresentare il digitale d'autore, restituendone la complessità e le relazioni contestuali? La fenomenologia delineata nei capitoli precedenti presenta un quadro complesso in cui i paradigmi consolidati della descrizione archivistica mostrano evidenti limiti di applicabilità (Langdon 2016; Bunn 2021).

In questo scenario, i LOD emergono come un approccio promettente per le esigenze di descrizione dei materiali nativi digitali. Basati sul Resource Description Framework (RDF) e sui principi del Web Semantico (Berners-Lee et al. 2001; Bizer et al. 2011), i LOD superano le limitazioni degli schemi tradizionali, consentendo una rappresentazione modulare ed esplicita e favorendo la condivisione di informazioni leggibili dalle macchine, configurandosi come fonti privilegiate per l'indicizzazione dei contenuti sul Web. La struttura dei LOD si fonda sul paradigma delle triple RDF, ciascuna composta da un soggetto, un predicato e un oggetto. Questo formalismo consente di descrivere con rigore entità e relazioni: il soggetto identifica una risorsa, il predicato ne specifica una proprietà o relazione, e l'oggetto rappresenta il valore associato o un'altra risorsa collegata. L'insieme di tali triple genera un grafo di conoscenza, i cui nodi sono definiti da identificatori univoci (URI). Le ontologie svolgono un ruolo centrale in questo processo, formalizzando le entità e le relazioni che strutturano il grafo.

Il presente capitolo intende offrire una panoramica degli strumenti per la modellazione archivistica disponibili, delineando lo stato dell'arte delle ontologie esistenti per costituire una base metodologica finalizzata alla descrizione del digitale d'autore. Con rappresentazione semantica si intende, in questo capitolo, la modellazione del dominio che si traduce nell'uso di ontologie formali. Verranno analizzate le caratteristiche delle principali ontologie utilizzate nella pratica archivistica (cap. 5.1) – RiC-O, ArCo, ArchOnto (su base CIDOC CRM) e PREMIS – con particolare attenzione agli allineamenti potenziali tra RiC-O e, rispettivamente, ArCo e ArchOnto (cap. 5.2). Il capitolo illustrerà, inoltre, due approcci per la trasformazione di inventari tradizionali in *knowledge base* RDF, evidenziando i benefici e i margini di innovazione che ne derivano (cap. 5.3).

5.1 Ontologie per gli archivi

L'archivio può essere inteso come un sistema complesso, costituito da elementi eterogenei che, ciascuno con le proprie specificità, concorrono a delineare una rappresentazione stratificata delle attività e delle entità coinvolte nel ciclo di vita dei documenti. Una tale rappresentazione non può prescindere dall'esplicitazione delle relazioni che intercorrono sia tra le singole unità archivistiche, sia tra queste e i molteplici contesti – istituzionali, culturali, amministrativi – che ne hanno influenzato la genesi e l'evoluzione (Damiani 2022).

Tra i principali e più influenti tentativi di formalizzare tale complessità si colloca, nel 1994, l'introduzione del General International Standard Archival Description (ISAD(G)) da parte dell'International Council on Archives (ICA). Lo standard ha favorito l'adozione di pratiche condivise e strutturate a livello internazionale, segnando una tappa cruciale nella storia della descrizione archivistica. Nel processo di standardizzazione promosso dall'ICA hanno fatto seguito, rispettivamente nel 1996, 2007 e 2008, l'International Standard Archival Authority Records for Corporate Bodies, Persons and Families **(**ISAAR(CPF)), l'International Standard for Describing Functions (ISDF) e l'International Standard for Describing Institutions with Archival Holdings (ISDIAH). Questi standard, sebbene abbiano incontrato una minore diffusione e applicazione rispetto ad ISAD(G), hanno offerto un contributo importante nella promozione di buone pratiche per la descrizione archivistica e l'interoperabilità. Tuttavia, come evidenziato da Pierluigi Feliciati:

> dopo circa 25 anni lo scambio di dati è stato raggiunto molto parzialmente, adottando solo modelli per consentire l'*harvesting* di descrizioni da parte di aggregatori nazionali o continentali. […] Certo, una normalizzazione degli inventari in rete c'è stata, ma i vincoli derivanti dalla concezione dello strumento descrittivo come opera chiusa, al massimo arricchita da indici dei nomi, dei luoghi etc. non sono mai stati superati con sistemi di integrazione dei dati (2021, 94).

Il modello gerarchico sotteso a ISAD(G), fondato su una logica verticale che privilegia la sequenzialità e l'ordine interno dei fondi, ha mostrato presto i propri limiti nel rappresentare le articolazioni trasversali e i legami orizzontali tra documenti, produttori e contesti (Vitali 2014; Damiani 2022). L'applicazione combinata di ISAD(G) con ISAAR(CPF), ISDF e ISDIAH avrebbe dovuto, almeno in teoria, mitigare tali criticità, offrendo un sistema integrato capace di restituire una visione più relazionale del patrimonio. In pratica, però, ciò non è avvenuto. Come recentemente riconosciuto dallo stesso ICA (2023), questi standard sono stati sviluppati nell'arco di oltre vent'anni, da gruppi di lavoro diversi, in risposta a esigenze metodologiche e tecnologiche del loro tempo, risultando così non pienamente coerenti tra loro

e poco chiari nelle modalità di integrazione. Il risultato è una descrizione che, anche se formalmente articolata, rimane frammentaria e incapace di restituire in modo organico la rete di relazioni che caratterizza i processi archivistici. L'interoperabilità tra le descrizioni archivistiche e la loro integrazione con l'ampio serbatoio informativo relativo a oggetti, eventi, agenti, luoghi e date presente restano, così, sostanzialmente inespresse (Feliciati 2021).

Le rigidità di questi standard emergono con particolare evidenza nei casi di ISAAR(CPF), ISDIAH e ISDF, i quali tendono a rappresentare gli agenti (persone, enti, famiglie) solo come produttori o conservatori, trascurando la dimensione plurale e dinamica delle interazioni sociali, materiali e istituzionali che contribuiscono alla formazione e alla gestione dei fondi. Si tratta di un'impostazione che si rivela già parziale nell'archivio analogico e che diventa del tutto insufficiente in quello digitale, dove le relazioni tra soggetti, processi e documenti si moltiplicano, si frammentano e si distribuiscono su scale spazio-temporali fluide, sfuggendo a schemi descrittivi lineari.

Come osservato da Giorgia Di Marcantonio (2023, 3), queste problematiche metodologiche presentano elementi di grande continuità con il passato, richiamando le difficoltà concettuali e operative che hanno caratterizzato anche l'elaborazione della Guida generale degli Archivi di Stato italiani. Le riflessioni di Claudio Pavone sulla complessa esperienza redazionale evidenziavano l'inadeguatezza degli strumenti descrittivi nel restituire «la reale vischiosità insita nei fondi documentari» e le criticità di un processo di ordinamento definito "metafisico" caratterizzato dalla ricerca di un compromesso tra la configurazione teorica ideale dell'archivio e la sua effettiva strutturazione documentaria (Di Marcantonio 2023, 4).

Parallelamente a queste riflessioni, negli ultimi decenni si è affermata una visione olistica e relazionale del patrimonio culturale, incentivata dall'avvento del digitale, che ha messo progressivamente in discussione la tradizionale separazione tra beni archivistici, librari e museali. In questa prospettiva, i fondi archivistici non sono più concepiti come entità chiuse e autonome, ma come nodi di una rete di strati di contesti interconnessi, in cui si intrecciano materiali, attori e pratiche provenienti da domini differenti.

Questa evoluzione ha avuto un riflesso anche nella gestione delle risorse culturali, sempre più orientata a modelli integrati. Archivi, biblioteche e musei, ossia le istituzioni del dominio MAB (o GLAM/LAM), hanno progressivamente adottato il paradigma dei LOD per migliorare la rappresentazione dei dati e potenziarne accessibilità e interoperabilità (Van Hooland e Verborgh 2014; Tomasi 2022). Il passaggio da modelli *document-centric* a modelli *data-centric* consente infatti di valorizzare in maniera più dinamica e flessibile le relazioni tra entità e contesti (Daquino 2021).

Le comunità internazionali dei musei e delle biblioteche hanno avviato già da oltre due decenni un processo di riflessione metodologica e sviluppo concettuale che ha portato alla definizione di modelli semantici. Nel settore museale, tale percorso ha condotto all'elaborazione di CIDOC Conceptual Reference Model (CIDOC CRM) a partire dal 1999, mentre nell'ambito bibliografico si è giunti alla formulazione dell'IFLA Library Reference Model object-oriented (LRMoo) (IFLA LRMoo Working Group e CIDOC CRM SIG 2024), ultimo stadio di una linea evolutiva che prende avvio dal modello Functional Requirements for Bibliographic Records (FRBR) (Riva et al. 2023) .

La comunità archivistica ha colto questa esigenza in maniera più asincrona, vedendo lo sviluppo di diverse iniziative locali prima di uno sviluppo internazionale (Tomasi e Daquino 2015; Henttonen e Kilkki 2017; Guernaccini et al. 2019). È in questo contesto che l'ICA ha istituito nel 2012 l'Expert Group on Archival Description (EGAD), con l'incarico di elaborare un nuovo standard capace di integrare e superare ISAD(G), ISAAR(CPF), ISDF e ISDIAH, rispondendo alle esigenze di una rappresentazione più articolata, dinamica e relazionale (Clavaud e Wildi 2021). Da questo lavoro è nato Records in Contexts (RiC), un modello strutturato in quattro componenti complementari: Records in Contexts - Foundations of Archival Description (RiC-FAD), Records in Contexts - Conceptual Model (RiC-CM)[129], Records in Contexts - Ontology (RiC-O, attualmente disponibile nella *release* 1.1) e Records in Contexts - Application Guidelines (RiC-AG)[130]. Dopo una lunga fase di bozze, le versioni 1.0 delle prime tre componenti sono state pubblicate alla fine del 2023, mentre lo sviluppo delle linee guida applicative, avviato all'inizio del 2024, ha portato alla pubblicazione della versione 0.1 nell'ottobre 2025 (International Council on Archives, Expert Group on Archival Description s.d.).

RiC-O, in particolare, è un'ontologia formale sviluppata in OWL 2 come implementazione tecnica di RiC-CM che estende e raffina includendo logica formale. RiC-O è progettata come un'ontologia di dominio che fornisce un vocabolario generico e regole formali per la creazione di dataset RDF che descrivano in modo coerente qualsiasi tipo di risorsa archivistica come Linked Data, supportando l'interrogazione tramite SPARQL e l'esecuzione di inferenze. Il suo scopo principale è fornire un vocabolario controllato e un insieme di regole formali per descrivere, in modo coerente e interoperabile,

[129] La prima bozza di RiC-CM, pubblicata nel 2016, ha suscitato un ampio interesse all'interno della comunità archivistica internazionale. Come ricorda Feliciati (2021), l'Expert Group on Archival Description (EGAD) ha ricevuto 64 documenti di osservazioni e commenti, elaborati da organismi e singoli archivisti provenienti da 19 paesi, per un totale di circa 220 pagine. Anche la comunità archivistica italiana ha contribuito attivamente al dibattito: sotto il coordinamento dell'Istituto Centrale per gli Archivi (ICAR) e dell'Associazione Nazionale Archivistica Italiana (ANAI), sono stati prodotti una serie di contributi, successivamente raccolti in un numero monografico dei "Quaderni del Mondo degli archivi", dal titolo *Records in Contexts: A conceptual model for archival description: Il contributo italiano* (n. 2, luglio 2017) (Vitali 2017).

[130] https://ica-egad.github.io/RiC-AG/

le risorse archivistiche e le loro entità contestuali, modellando le complesse relazioni che le legano. A differenza degli standard ICA tradizionali come ISAD(G), ISAAR(CPF), ISDF e ISDIAH il suo approccio non è gerarchico ma si basa su un modello a rete che collega entità attraverso relazioni semanticamente definite. La *release* attuale e più aggiornata è la 1.1, pubblicata nel maggio 2025 che consolida e amplia la prima *release* stabile (v. 1.0, dicembre 2023) introducendo aggiunte e modifiche significative.

Le caratteristiche fondamentali di RiC-CM e RiC-O includono una modellazione basata su entità e relazioni che permette di descrivere non solo la risorsa archivistica in sé, ma tutte le entità del suo contesto (agenti, attività, mandati, luoghi, date) e le loro relazioni (Figura 5.1).

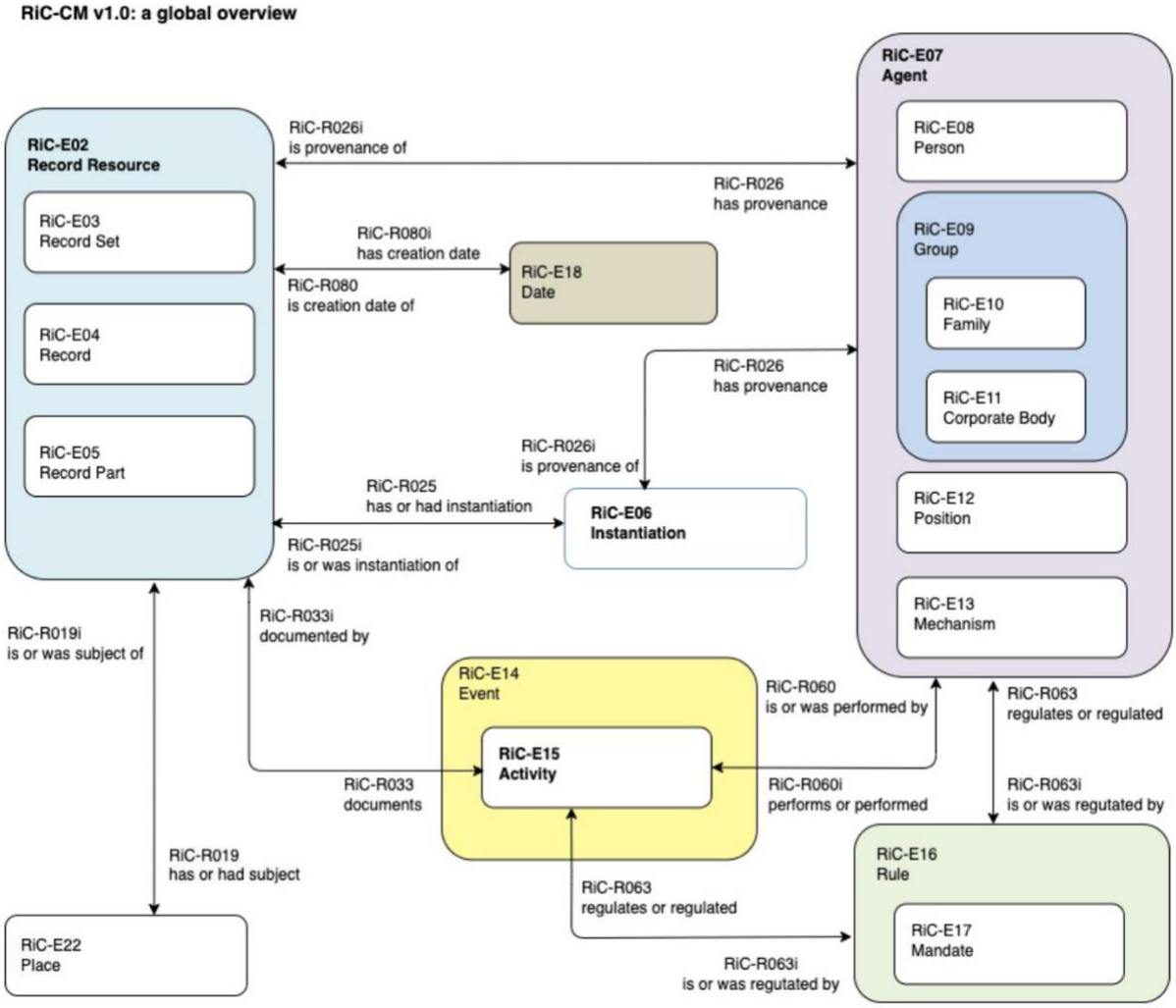


Figura 5.1. Panoramica di RiC-CM pubblicata da ICA-EGAD nel settembre 2023.

Complessivamente, nella sua versione 1.1, RiC-O è costituito da 110 classi, 485 *object property* e 76 *data property*. Il principio cardine nella rappresentazione di qualsiasi risorsa archivistica è la distinzione concettuale tra `rico:RecordResource` (l'entità intellettuale, ossia il contenuto concettuale) e le sue istanziazioni `rico:Instantiation` (le sue forme fisiche o digitali). Un singolo record, ad esempio, può avere multiple istanziazioni: un originale cartaceo, una sua digitalizzazione, una copia microfilm e versioni derivative di vario tipo.

Accanto a `rico:RecordResource` e `rico:Instantiation`, che costituiscono il nucleo della modellazione, RiC-O definisce altre classi fondamentali per descrivere l'ecosistema documentario. In particolare:

- `rico:RecordSet`, che rappresenta aggregazioni concettuali di record consentendo di esprimere i legami gerarchici e aggregativi prescindendo dai canonici livelli di descrizione (che è comunque possibile specificare tipizzando `rico:RecordSet` attraverso `ricoRecordSetType`);
- `rico:Agent`, che comprende la modellazione di persone, famiglie ed enti coinvolte, a titoli diversi, nella rappresentazione delle risorse archivistiche. Gli agenti sono connessi ad altre entità tramite relazioni diverse, come quelle di creazione, responsabilità, proprietà o partecipazione;
- `rico:Activity`, che modella le azioni e i processi (ad esempio un atto amministrativo, un trasferimento o una campagna di digitalizzazione) che coinvolgono agenti e producono o trasformano `rico:RecordResource` e `rico:Instantiation`;
- `rico:Mandate`, che permette di rappresentare i vincoli normativi, giuridici o istituzionali che regolano le attività degli agenti;
- `rico:Date` e `rico:Place`, che specificano rispettivamente coordinate spazio-temporali, fornendo il contesto essenziale per l'interpretazione delle risorse archivistiche.

Il rapporto tra queste classi è reso esplicito da un articolato sistema di *object property* che consente di rappresentare relazioni molteplici, sia di tipo verticale (ad esempio `rico:isRecordSetOf` per indicare appartenenze gerarchiche) sia di tipo orizzontale e reticolare (ad esempio `rico:hasSuccessor`, `rico:hasRelatedRecord`, `rico:isAssociatedWithPlace`).

Una differenza significativa tra RiC e gli standard ICA esistenti è che questi ultimi modellano principalmente uno strumento di corredo, tendenzialmente un inventario (*finding aid*) organizzando la descrizione in una gerarchia predefinita a cui associare metadati. Al contrario, RiC-CM (e quindi RiC-O) modella le entità stesse (il record, il suo creatore, l'attività che l'ha prodotto, il luogo, ecc.) e le loro

relazioni, senza presupporre uno specifico prodotto finale di descrizione. Dalla riflessione fatta da EGAD e il relativo stato dell'arte, è emersa l'esigenza di una descrizione archivistica flessibile, con finalità e livelli di granularità delle informazioni che possono variare notevolmente. L'ontologia è stata quindi progettata con un'architettura estensibile, in grado di adattarsi a contesti e requisiti descrittivi diversificati. L'approccio basato su entità consente inoltre di generare dinamicamente viste multiple e personalizzate sui medesimi dati, superando i limiti delle rappresentazioni gerarchiche tradizionali e offrendo prospettive complementari sulle risorse archivistiche.

EGAD sottolinea però anche ciò che RiC-O non è: non costituisce un set di regole per la composizione del contenuto descrittivo, non è una specifica di implementazione per sistemi di gestione documentale e non è un modello per la gestione fisica dei record (sebbene possa essere esteso per integrarlo) (International Council on Archives, Expert Group on Archival Description 2025b).

La *roadmap* per RiC-O include l'articolazione delle classi `rico:Event` e `rico:Activity` e del sistema di classi `rico:Relation`, l'aggiunta di suggerimenti da parte della comunità e di mappature OWL tra alcune classi o proprietà e componenti di altri modelli (tra cui PREMIS, Schema.org, PROV-O, IFLA-LRM e RDA, CIDOC-CRM), e lo sviluppo di vocabolari multilingue per fornire agli utenti strumenti per popolare gli attributi RiC-CM che hanno valori controllati (International Council on Archives, Expert Group on Archival Description 2025a). Questa evoluzione risponde alla necessità di supportare l'applicazione di RiC-O in contesti internazionali e multilingue e di integrarsi con l'ecosistema globale delle ontologie per i beni culturali.

Questa evoluzione risponde a esigenze di ampio respiro, ma l'adozione di RiC-O rappresenta, nella pratica, una sfida significativa per la comunità archivistica, perché comporta un autentico cambio di paradigma che investe non solo le dimensioni tecniche, ma anche i fondamenti concettuali della disciplina (Feliciati 2021). La piena assimilazione non può essere considerata né immediata né scontata e richiede il coinvolgimento diretto degli archivisti insieme a un articolato processo di formazione (Di Marcantonio 2023). La transizione è resa più complessa anche dalle criticità emerse durante lo sviluppo iniziale di RiC, che ha presentato diverse zone d'ombra, rivelandosi «singhiozzante e per alcuni tratti auto-referenziale» (Feliciati 2021, 100). Diverse criticità sono state sollevate da autorevoli voci della comunità archivistica internazionale, che hanno sottolineato come lo sviluppo di RiC non sia stato sufficientemente comunicato e condiviso con la comunità archivistica durante le sue fasi iniziali di elaborazione (InterPARES Trust 2016; Gillean 2017; Feliciati 2021). Questa mancanza di dialogo ha reso obiettivamente difficile fornire commenti costruttivi su un draft ormai giunto a maturazione, limitando le possibilità di contributo critico da parte della comunità. Il processo è stato criticato per la

mancanza di trasparenza nei criteri di selezione degli esperti di EGAD, per non aver coinvolto rappresentanti di tutti i continenti (in particolare Africa e Asia), e per non aver condotto un'analisi preliminare dell'effettivo livello di applicazione degli standard ICA nei diversi paesi 05/03/2026 20:52:00.

I limiti delle prime versioni sono stati ulteriormente esacerbati dalla scarsità di esempi applicativi, dalla mancanza di materiali didattici e di un IRI stabile, fattori che hanno ostacolato la sperimentazione pratica e rallentato l'adozione diffusa (Feliciati 2021, 100). Per superare tali difficoltà, l'EGAD è attualmente impegnato nella redazione delle *Records in Contexts - Application Guidelines* (RiC-AG), l'ultimo documento previsto dallo standard, volto a fornire indicazioni operative per l'implementazione di RiC. La prima bozza è stata pubblicata alla conferenza annuale ICA del 2025 a Barcellona e intende offrire un quadro metodologico condiviso e strumenti pratici a sostegno della comunità archivistica. Il quadro di sviluppo ha registrato comunque alcuni miglioramenti significativi. Tra questi: l'aumento della condivisione delle attività tramite il *repository* pubblico GitHub dedicato al progetto, la creazione del Google Group "Records_in_Contexts_users", amministrato da EGAD, che funge da forum principale per discussioni, domande e scambio di esperienze, e la pagina dedicata alla raccolta su base comunitaria di risorse su RiC (articoli, dataset, tools e applicazioni)[131].

Dal punto di vista implementativo, Di Marcantonio (2023) esprime preoccupazioni riguardo alle potenziali disparità di adozione a causa di diverse potenzialità di investimento tecnico e di formazione da parte delle istituzioni e della comunità. Lo stesso EGAD, d'altronde, nella bozza di RiC-CM del 2016 affermava che «many institutions will simply not have the resources to immediately embrace RiC-CM» (Experts Group on Archival Description 2016). Come osservato da Di Marcantonio, se uno degli obiettivi è creare «bacini di conoscenza comprensibili alle macchine, per favorire un accesso integrato alle risorse», il risultato potrebbe essere paradossalmente opposto: «o la maggior parte dei destinatari del RiC integreranno il nuovo 'super standard' nei processi descrittivi, oppure continueremo ad avere una mappatura del tutto parziale e scarsamente efficace della conoscenza» (Di Marcantonio 2023, 6-7).

Nonostante queste criticità, l'interesse verso RiC-O è in crescita. Diverse istituzioni, ricercatori e aziende hanno iniziato a esplorarne le potenzialità, sperimentandone l'uso sia per la pubblicazione di dati archivistici in LOD, sia per l'arricchimento di metadati esistenti o per la progettazione di nuovi sistemi di descrizione (Mikhaylova e Metilli 2023; Santos e Revez 2023; García-González e Bryant 2023; De Coulon 2024; Rajh 2024; Pashkeeva et al. 2024). L'implementazione rimane dunque in una fase ancora

[131] https://ica-egad.github.io/RiC-ResourceList/.

iniziale, ma segnata da un'accelerazione che lascia intravedere il ruolo centrale che RiC-O potrebbe assumere negli anni a venire.

È necessario riconoscere che le transizioni tecnologiche e concettuali di questa portata richiedono tempi di sedimentazione naturalmente estesi, seguendo dinamiche di adozione graduale che caratterizzano storicamente ogni innovazione nel campo dell'*information management*. Le criticità evidenziate, pur assolutamente necessarie per orientare il dibattito scientifico, devono essere contestualizzate in una prospettiva evolutiva che consideri l'implementazione di RiC come un processo incrementale, destinato a svilupparsi inizialmente presso le istituzioni di maggiori dimensioni e risorse, per poi eventualmente diffondersi attraverso meccanismi di trasferimento tecnologico e condivisione di *best practices*. In questo scenario, le tecnologie di IA potrebbero accelerare la transizione, supportando la conversione dei metadati, l'automazione della creazione di triple RDF e lo sviluppo di interfacce di visualizzazione. Allo stesso tempo, sistemi di raccomandazione, algoritmi di disambiguazione semantica e interfacce conversazionali faciliterebbero l'accesso anche agli utenti meno esperti ai complessi grafi di conoscenza archivistica.

Se RiC rappresenta oggi il riferimento più istituzionalizzato, essendo promosso dall'ICA, è stato preceduto da una serie di iniziative finalizzate a esplorare il potenziale dei LOD nel dominio. Parallelamente al suo sviluppo, infatti, si è distinto un altro progetto rilevante nel panorama archivistico, ossia l'Entity and Property Inference for Semantic Archives (EPISA)[132]. Avviato nel 2019 come sforzo congiunto tra informatici, esperti in scienze dell'informazione e archivisti dell'Arquivo Nacional da Torre do Tombo (l'Archivio Nazionale Portoghese), EPISA è stato un progetto quadriennale (2019-2022) guidato dall'Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), con l'Università di Évora e la Direção-Geral do Livro, dos Arquivos e das Bibliotecas (DGLAB) come partner, con l'obiettivo principale l'integrazione dell'istituzione nella rete globale dei LOD (I. Koch et al. 2019). Il progetto è nato come risposta ai limiti del sistema informativo archivistico del DGLAB, DigitArq[133], caratterizzato da descrizioni con componenti testuali complesse, dense e scarsamente interconnesse, che rendevano difficile per gli utenti individuare i documenti di interesse ed esplorarne le relazioni. L'emergere dei modelli LOD nel settore dei beni culturali ha spinto DGLAB a ripensare il proprio approccio alla descrizione archivistica, con l'obiettivo di migliorare la struttura dei record e l'esperienza di fruizione degli utenti (I. D. Koch 2025, 64).

[132] https://episa.inesctec.pt/.
[133] https://digitarq.arquivos.pt/.

Considerando il panorama allora esistente delle ontologie applicate al dominio dei beni culturali, e lo stato ancora embrionale di RiC, il progetto ha portato allo sviluppo di ArchOnto, un modello concettuale per la descrizione delle risorse archivistiche fondato su CIDOC CRM e su ulteriori ontologie complementari (I. Koch et al. 2020). Il progetto EPISA ha quindi articolato la propria strategia su due fronti: da un lato, l'elaborazione di ArchOnto, con l'obiettivo di superare le limitazioni descrittive tradizionali; dall'altro, la progettazione e implementazione di un sistema di migrazione delle descrizioni archivistiche esistenti, codificate secondo lo standard ISAD(G), verso il nuovo *framework* semantico. Questa duplice dimensione ha richiesto lo sviluppo di regole di mappatura specifiche e di algoritmi di estrazione delle entità per gestire un prototipo dedicato alla transizione di record già presenti in DigitArq, garantendo al contempo la preservazione dei dati pregressi e l'apertura verso nuove possibilità di interrogazione e collegamento degli stessi.

ArchOnto[134] ha una struttura modulare che integra cinque ontologie (Figura 5.2) per trattare aspetti complementari del dominio archivistico, e prevede complessivamente 127 classi, 315 *object property*, 56 *data property*. Le cinque ontologie sono:

- CIDOC CRM[135]: il nucleo di base di ArchOnto, che fornisce i concetti e le proprietà principali per catturare le caratteristiche essenziali dei documenti d'archivio (evento, data, luogo, persona, gruppo). La scelta di CIDOC CRM è stata motivata dalla sua stabilità, dalla comunità internazionale attiva e dalle applicazioni già consolidate in altri domini;
- CIDOC CRM PC[136]: adottata per rappresentare relazioni n-arie, ovvero quelle che collegano più di due entità (ad esempio, per rappresentare il ruolo di una persona in un dato evento, che implica il collegamento di un evento, una persona e un ruolo)[137];

[134] https://purl.org/episa/archonto.

[135] https://cidoc-crm.org/html/cidoc_crm_v7.1.3.html.

[136] http://purl.org/episa/archonto/cidoc_crm_pc.

[137] Nei linguaggi del Web Semantico, una proprietà è una relazione utilizzata per collegare due individui o un individuo e un valore. Tuttavia, in alcuni casi, il modo più naturale e conveniente per rappresentare certi concetti è utilizzare relazioni per collegare un individuo a più di un solo individuo o valore. Queste relazioni sono chiamate relazioni n-arie. Ad esempio, potremmo voler rappresentare le proprietà di una relazione, come la nostra certezza su di essa, le date di esistenza, e così via. Un altro esempio è la rappresentazione di relazioni tra più individui, come un acquirente, un venditore e un oggetto che è stato acquistato (W3C Semantic Web Best Practices and Deployment Working Group 2006). Durante la fase sperimentale di ArchOnto, al posto di CIDOC CRM PC è stata inizialmente utilizzata una N-ary Ontology sviluppata ad hoc, poiché la versione stabile della PC Ontology è stata pubblicata solo a progetto inoltrato. La sostituzione è stata agevole grazie all'aderenza della N-ary Ontology alle raccomandazioni di CIDOC CRM, garantendo la compatibilità concettuale (I. D. Koch 2025, 87). La documentazione della N-ary Ontology è disponibile all'IRI https://purl.org/episa/archonto/n_ary.

- DataObject[138]: ontologia ausiliaria per gestire valori *literal* e la loro validazione in ArchOnto. Questo modulo è stato sviluppato per assicurare coerenza e uniformità nei formati di elementi semplici come date, stringhe testuali o numeri;
- Ontologia ISAD[139]: comprende solo *data property* che corrispondono agli elementi standard ISAD(G). L'ontologia è stata sviluppata per proporre una rappresentazione intermedia delle informazioni già codificate in ISAD(G). Questa ontologia è stata utilizzata per migrare i documenti già esistenti in DigitArq verso ArchOnto, con il ruolo di garantire una transizione graduale verso il modello più granulare. Il suo utilizzo non è previsto per l'elaborazione di nuove descrizioni;
- Link2DataObject[140]: effettua il collegamento tra le ontologie CIDOC CRM e DataObject. È costituita da un'unica proprietà utilizzata per realizzare il collegamento (`L2DO hasValue`). La separazione di questo modulo da DataObject ha esclusive motivazioni pratiche: in caso di sostituzione di DataObject con un'altra ontologia, il collegamento con CIDOC CRM rimarrebbe intatto, preservando l'integrità complessiva del modello.

Oltre a queste cinque ontologie, ArchOnto implementa le sue proprie classi e proprietà, create come estensioni a CIDOC CRM. Le classi sono utilizzate per specializzare concetti già presenti in CIDOC CRM ma che necessitano di adattamenti per includere vocabolari controllati in chiave archivistica. Le tre proprietà principali, `ARP12 has description level`, `ARP8 upper level` e `ARP9 lower level`, strutturano i legami gerarchici fra unità descrittive, consentendo di rappresentare fedelmente la stratificazione tradizionale delle descrizioni archivistiche.

[138] https://purl.org/episa/archonto/data_object.
[139] https://purl.org/episa/archonto/isad_ontology.
[140] https://purl.org/episa/archonto/link2data_object.

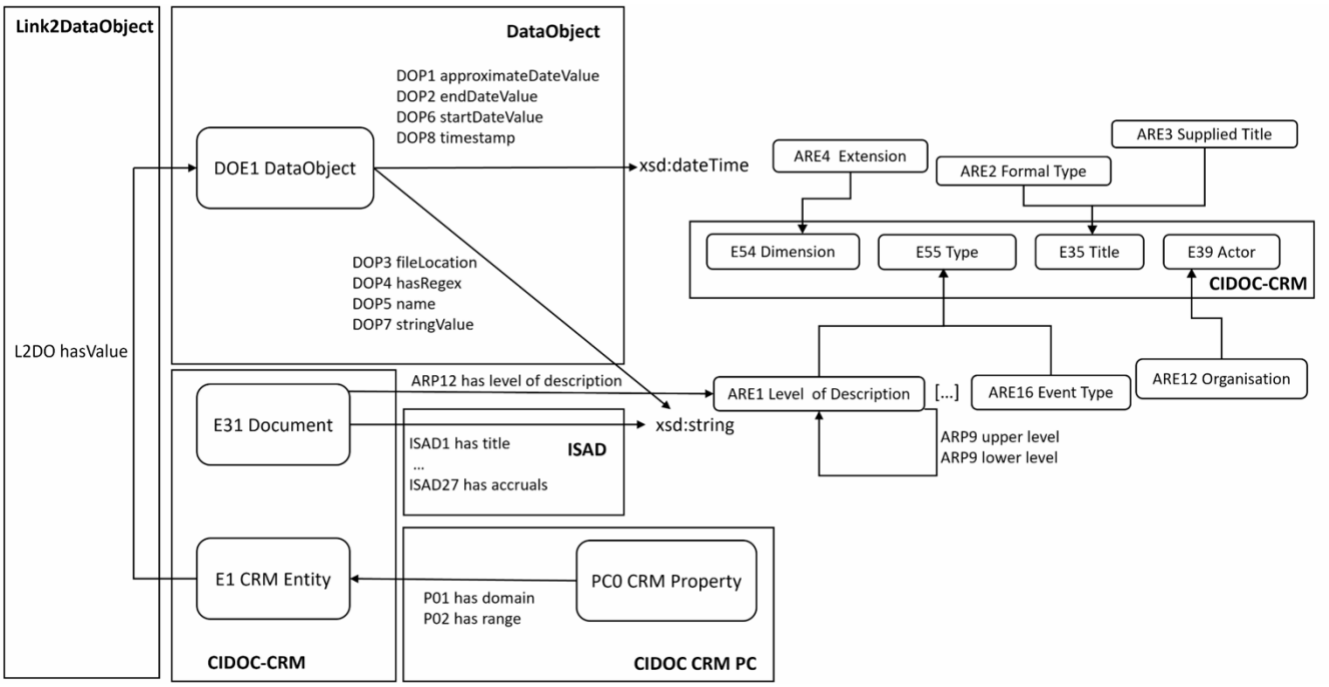


Figura 5.2. Rappresentazione della struttura di ArchOnto elaborata da Inês Koch.

La rappresentazione di una risorsa archivistica in ArchOnto si articola attorno a un sistema stratificato di entità che catturano sia gli aspetti fisici che concettuali della documentazione. Al centro di questo sistema si trova la classe `E31 Document` di CIDOC CRM, che rappresenta l'unità di base e funge da punto di convergenza per tutte le altre informazioni descrittive. Il primo elemento strutturante è il livello di descrizione (`ARE1 Level of Description`), che definisce la posizione del documento nella gerarchia archivistica tradizionale. Questa classe specializza il concetto più generico di `E55 Type` di CIDOC CRM; la gerarchia viene mantenuta attraverso le proprietà `ARP8 upper_level` e `ARP9 lower_level`, che stabiliscono le relazioni di livello che consentendo di navigare verticalmente nella struttura archivistica.

Per l'identificazione del documento, ArchOnto utilizza un sistema che comprende sia i titoli che gli identificativi. I titoli si dividono in due categorie principali: `ARE2 Formal Title` per i titoli formali originali del documento, e `ARE3 Supplied Title` per quelli attribuiti dall'archivista. Gli identificatori (`E42 Identifier`) sono tipizzati attraverso `ARE5 Identifier Type`, permettendo di distinguere tra *Reference Code*, *Physical Location*, *Original Numbering* e altre tipologie identificative essenziali per la gestione archivistica.

Gli agenti correlati ai documenti sono `E21 Person` per gli individui e `E74 Group` (con la sua specializzazione `ARE12 Organisation`) per le entità collettive. Particolarmente importante è il sistema di rappresentazione dei ruoli tramite `ARE8 Role Type`, che permette di specificare la funzione svolta da

una persona in un determinato evento (produttore, autore materiale, destinatario, etc.), utilizzando relazioni n-arie per collegare persona, evento e ruolo.

La dimensione spazio-temporale è articolata attraverso `E53 Place` per le localizzazioni geografiche, tipizzate da `ARE14 Place Type`, ed `E52 Time-Span` per gli aspetti temporali. Le date sono ulteriormente qualificate da `ARE6 Date Type` e `ARE9 Date Certainty`, permettendo di distinguere tra date esatte, inferite, predominanti e di specificarne il grado di certezza. Il sistema temporale si integra con gli eventi (`E5 Event`, `E7 Activity`) per contestualizzare cronologicamente le azioni documentate.

Gli eventi costituiscono il cuore concettuale del modello, seguendo l'approccio *event-centered* di CIDOC CRM. Attraverso `E5 Event` ed `E7 Activity`, tipizzati da `ARE16 Event Type`, viene rappresentata la dimensione processuale della documentazione. Eventi come la produzione (`E12 Production`), la nascita (`E67 Birth`), o attività più specifiche, vengono collegati ai documenti che li testimoniano, alle persone che vi partecipano e ai luoghi dove si svolgono.

ArchOnto distingue sistematicamente tra gli aspetti fisici e concettuali del documento. La dimensione fisica è rappresentata attraverso `E22 Human-Made Object`, con le sue proprietà materiali (`E57 Material` per il supporto) e dimensionali (`E54 Dimension`, specializzato in `ARE4 Extension`). La dimensione concettuale è catturata da `E33 Linguistic Object`, che include aspetti come la lingua (`E56 Language`) e le caratteristiche intellettuali del contenuto.

Il popolamento di ArchOnto è stato sperimentato in modi diversi, sia attraverso la creazione manuale di individui a partire dai record ISAD(G), sia tramite processi automatici di migrazione con regole di trasformazione tra gli elementi di ISAD(G) e le corrispondenti entità dell'ontologia (Koch 2025, 90-98, 127-29). Un primo esperimento ha riguardato la validazione manuale di un campione di 25 record, con l'analisi approfondita della scheda descrittiva del fondo "Juízo da Índia e Mina", che ha permesso di verificare l'aderenza del modello a descrizioni archivistiche preesistenti (Figura 5.3).

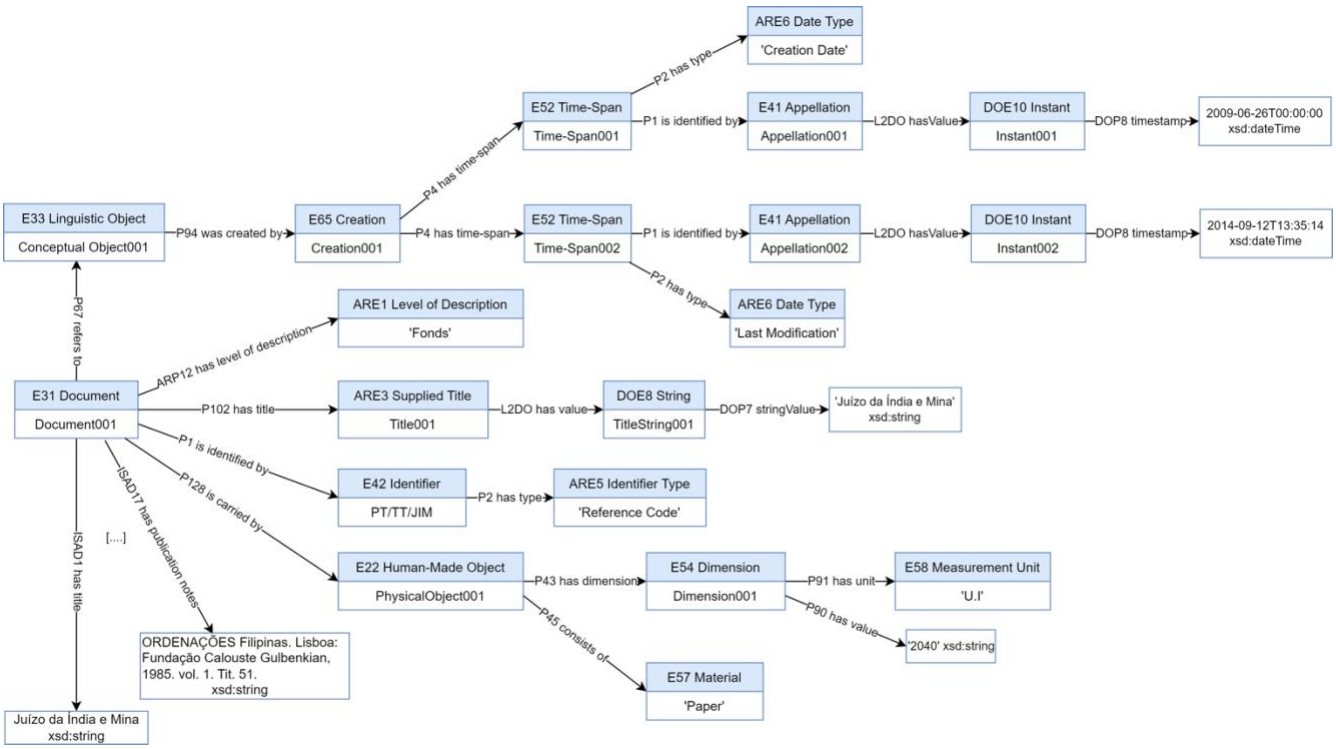


Figura 5.3. Esempio di rappresentazione della descrizione di livello "Fondo" del fondo Juízo da Índia e Mina.

Successivamente si è affrontata la migrazione automatica da ISAD(G) ad ArchOnto di alcuni set di dati, in particolare delle serie dei Registos de Baptismos dell'Archivio Distrettuale di Porto (1216 record di battesimo dal 1644 al 1911) e dei Registos de Passaportes Deferidos dell'Archivio Distrettuale di Bragança (102 record di passaporti dal 1914 al 1918) (I. D. Koch 2025, 58).

Secondo quanto analizzato da Inês Koch, ArchOnto presenta un incremento significativo dell'espressività descrittiva per il dominio archivistico di CIDOC CRM, con un aumento del 16% nelle classi utilizzabili e del 13% nelle proprietà disponibili (2025, 129-30). La migrazione automatica di oltre 1.300 record archivistici ha confermato la scalabilità del modello e la sua capacità di preservare l'integrità semantica delle descrizioni esistenti, incrementando del 6% il numero totale di asserzioni rispetto alla rappresentazione in CIDOC CRM base (I. D. Koch 2025, 130).

L'esperienza portoghese, attraverso l'implementazione di descrizioni conformi ad ArchOnto nella piattaforma EPISA e la sua parziale integrazione nei sistemi nazionali, attesta la maturità del modello, dimostrando concretamente la fattibilità della transizione verso paradigmi semantici nel dominio archivistico, incluso il recupero del pregresso.

In Italia, lo stato dell'arte delle ontologie per gli archivi affonda le sue radici in progetti come ReLOAD (Repository for Linked Open Archival Data) [141], promosso dall'Archivio Centrale dello Stato, dall'Istituto per i beni culturali della Regione Emilia-Romagna e da regesta.exe. Dal 2011 al 2014 ReLOAD ha sviluppato EAC-CPF Ontology[142], Ontology for Achival description (OAD) [143], Ontology for cultural organizations services and access (OCSA) [144], modelli che hanno posto le basi per formalizzare soggetti produttori, autorità e relazioni archivistiche in ottica LOD. Sulla base di questi modelli, ReLOAD ha sperimentato la pubblicazione in LOD di descrizioni archivistiche di fondi dell'Archivio Centrale dello Stato, della Camera dei Deputati e dell'Istituto per i beni culturali della Regione Emilia-Romagna[145], dimostrando da un lato la fattibilità tecnica e il valore informativo dei LOD per l'accessibilità, la valorizzazione e l'integrazione delle descrizioni archivistiche; dall'altro la complessità a far dialogare modelli e pratiche esistenti (EAD, EAC-CPF, tracciati nazionali) con gli standard del Web Semantico (Guernaccini et al. 2019).

Parallelamente, l'Istituto Centrale per gli Archivi (ICAR) sviluppava l'Ontologia del Sistema Archivistico Nazionale (SAN), denominata SAN LOD[146]**,** con l'obiettivo di tradurre in OWL/RDF lo schema concettuale dei tracciati di scambio CAT SAN alla base del portale nazionale degli archivi (Sistema Archivistico Nazionale 2014). Pubblicata nel 2014, l'ontologia SAN definisce formalmente soggetti produttori, soggetti conservatori, complessi archivistici e strumenti di ricerca, inserendoli in un modello interoperabile con vocabolari come SKOS[147] e Geonames[148]. Con 27 classi e oltre 84 proprietà,

[141] Il sito web del progetto non è più disponibile, ma la documentazione si può ancora rintracciare nelle notizie d'archivio del magazine di regesta.exe (Regesta.exe, redazione 2013) e nel *repository* GitHub dedicato al progetto: https://github.com/regestaexe/ReloadProject/tree/master.

[142] EAC-CPF Ontology è finalizzata alla descrizione di tutti i soggetti (enti, persone e famiglie) coinvolti nella produzione degli archivi. Tramite questa ontologia possono essere descritti non solo i soggetti produttori in ottica ISAAR(CPF), ma anche tutti quei soggetti che in senso più ampio costituiscono o concorrono a formare il contesto di produzione, raccolta o sedimentazione di un archivio. La documentazione dell'ontologia è disponibile al link: http://culturalis.org/eac-cpf/.

[143] OAD, espressa in OWL (Ontology Web Language), prende in considerazione tutti gli elementi della descrizione archivistica previsti dallo standard ISAD(G) integrandoli con altri elementi informativi non previsti dallo standard citato - come le voci di indice - e con i collegamenti ai soggetti produttori e conservatore. La documentazione dell'ontologia è disponibile al link: http://culturalis.org/oad/.

[144] OCSA è finalizzata alla descrizione dei servizi, le attrezzature, le attività di qualunque soggetto (persona, famiglia, ente pubblico o privato) che si configuri e agisca in qualità di istituto culturale/soggetto conservatore. La documentazione dell'ontologia è disponibile al link: http://culturalis.org/ocsa/1.0/.

[145] I file RDF contenenti le descrizioni sono disponibili nel *repository* del progetto al link: https://github.com/regestaexe/ReloadProject/tree/master/data.

[146] La documentazione dell'ontologia è disponibile al link: http://www.maas.ccr.it/SAN-LOD/lode/.

[147] https://www.w3.org/TR/skos-reference/.

[148] https://www.geonames.org/.

SAN LOD rappresenta il primo tentativo ministeriale di fornire un modello semantico condiviso per l'ecosistema archivistico italiano.

In questo quadro si è progressivamente affermata ArCo (Architettura della Conoscenza)[149], la rete di ontologie promossa dall'Istituto Centrale per il Catalogo e la Documentazione (ICCD) in collaborazione con l'Istituto di Scienze e Tecnologie della Cognizione - ISTC (CNR) e l'Università di Bologna. ArCo non è una singola ontologia ma un ecosistema modulare che nasce con l'obiettivo di rappresentare in RDF il Catalogo Generale dei Beni Culturali, fornendo vocabolari specifici per descrivere oggetti culturali, contesti, eventi, collocazioni, schede catalografiche e aspetti denotativi della descrizione (Carriero et al. 2019). L'approccio adottato è *pattern-based* e orientato al riuso: ArCo importa e si allinea con vocabolari nazionali (Cultural-ON[150], OntoPiA[151] – con Level-0 ispirata a DOLCE-Zero[152]) e con modelli internazionali (DOLCE+DnS [153], CIDOC CRM [154], Europeana Data Model (EDM) [155], BIBFRAME[156], FRBR[157], FaBiO[158], FEntry[159], OAEntry), posizionandosi come ponte semantico tra il modello catalografico italiano e l'ecosistema globale dei beni culturali.

L'architettura modulare articola la rappresentazione del patrimonio culturale attraverso diversi componenti specializzati, ciascuno dedicato a specifiche dimensioni descrittive. L'attuale versione (1.0) è costituita da sette moduli: il modulo ArCo[160] importa tutti gli altri moduli e modella le informazioni centrali sui beni culturali. La gerarchia è organizzata attorno alla classe `arco:CulturalProperty`, distinta in beni culturali intangibili e tangibili, ulteriormente suddivisi in mobili e immobili, con tipologie specifiche articolate su più livelli. Vengono inoltre modellati i materiali componenti di lotti archeologici e introdotte categorie trasversali di classificazione (es. beni fotografici, musicali, ecc.). Il modulo Core[161] costituisce l'infrastruttura concettuale fondamentale del sistema, fornendo classi di base per identificativi, tipi, concetti e classificazioni che fungono da dorsale semantica per l'intero network e

[149] https://dati.beniculturali.it/arco/index.php.
[150] https://dati.beniculturali.it/cultural-ON/ITA.html.
[151] https://github.com/italia/dati-semantic-assets/wiki.
[152] https://schema.gov.it/lodview/onto/l0.
[153] https://www.cnr.it/en/institutes-databases/database/684/dolce-dns-ontology-library.
[154] https://cidoc-crm.org/.
[155] https://pro.europeana.eu/page/edm-documentation.
[156] https://www.loc.gov/bibframe/.
[157] https://sparontologies.github.io/frbr/current/frbr.html.
[158] https://sparontologies.github.io/fabio/current/fabio.html.
[159] https://essepuntato.it/fentry/current/fentry.html.
[160] https://w3id.org/arco/ontology/arco.
[161] https://w3id.org/arco/ontology/core.

garantiscono coerenza terminologica tra i diversi domini. Il modulo Context-Description[162] gestisce gli aspetti contestuali della descrizione attraverso la modellazione di date, responsabilità, acquisizioni, situazioni legali e interventi conservativi, con particolare attenzione alla temporalità delle informazioni descrittive che permette di tracciare l'evoluzione delle caratteristiche dei beni nel tempo. La dimensione fisica e tecnica trova rappresentazione nel modulo Denotative-Description[163], che modella materiali, tecniche esecutive, misurazioni e aspetti morfologici dei beni culturali, mentre la localizzazione spazio-temporale è affidata al modulo Location[164] che integra coordinate geografiche, indirizzi e collocazioni temporali seguendo standard geografici nazionali e internazionali. Il modulo Catalogue[165] formalizza le pratiche catalografiche consolidate attraverso strutture per schede, record e normative ICCD, mantenendo aderenza agli standard ministeriali e facilitando la migrazione di dati catalografici esistenti. Infine, l'ontologia Cultural Event[166] modella gli eventi culturali intesi come eventi che coinvolgono un bene culturale.

Questa architettura modulare permette utilizzi selettivi dei componenti in base alle necessità specifiche, supportando sia implementazioni complete che adozioni parziali per domini specializzati, mantenendo allo stesso tempo interoperabilità semantica attraverso l'infrastruttura condivisa.

La prima versione di ArCo risale al marzo 2018 con una vocazione orientata primariamente alla rappresentazione beni catalogati dall'ICCD (opere d'arte, reperti, monumenti). Questo ha reso le prime versioni poco rappresentative delle specificità archivistiche. Per rispondere a questo divario è stato sviluppato e rilasciato nel 2023 un ulteriore modulo, arco-archive, giunto alla versione 0.2 nel dicembre 2023 e testato sul caso d'uso del Portale delle Fonti per la storia della Repubblica italiana[167], con l'obiettivo di fornire una rappresentazione semantica strutturata delle risorse archivistiche relative al periodo storico 1943-1953.

[162] https://w3id.org/arco/ontology/context-description.
[163] https://w3id.org/arco/ontology/denotative-description.
[164] https://w3id.org/arco/ontology/location.
[165] https://w3id.org/arco/ontology/catalogue.
[166] https://w3id.org/arco/ontology/cultural-event.
[167] https://portalefontirepubblicaitaliana.cnr.it/

Complessivamente il modulo archive risulta composto da 310 classi, 608 *object property* e 181 *data property*. Il modulo adotta un approccio minimalista basato su sei classi principali che catturano gli elementi essenziali della descrizione archivistica attraverso una strategia di integrazione modulare con l'ecosistema ArCo (Figura 5.4). Il modello si articola attorno alla classe centrale `ArchivalResource`, che estende il concetto di entità culturale per rappresentare qualsiasi risorsa archivistica indipendentemente dal suo livello gerarchico. Questa classe stabilisce connessioni con i moduli esterni del network ArCo attraverso proprietà che delegano funzionalità specifiche ai componenti specializzati: il modulo context-description fornisce il *framework* per datazione (`hasDating`), responsabilità (`hasResponsibility`), acquisizioni (`hasAcquisition`) e situazioni giuridiche (`hasLegalSituation`), mentre il modulo arco-lite gestisce responsabilità semplificate attraverso `hasAuthor` e `hasHolder` per casi d'uso che non richiedano la complessità del sistema contestuale completo.

La gestione delle aggregazioni archivistiche è affidata alla classe `ArchivalCollection`, che modella collezioni di risorse unite da vincoli archivistici includendo proprietà specifiche per la descrizione del sistema di organizzazione (`arrangementSystem`) e della consistenza fisica (`linearShelfSpace`), marcate come provvisorie per indicare la natura evolutiva del modello. La relazione tra collezioni e risorse è formalizzata attraverso proprietà ereditate dal sistema CIS che stabiliscono appartenenze e gerarchie, mentre la localizzazione spazio-temporale è delegata interamente al modulo Location attraverso `hasTimeIndexedTypedLocation`, che fornisce geolocalizzazione temporalizzata senza richiedere implementazioni specifiche nell'ontologia archivistica.

La proprietà `hasArchivalResourceLevel` specializza la relazione generica `core:hasType` mantenendo compatibilità semantica con il modulo core, che fornisce l'infrastruttura concettuale di base attraverso `hasType`, `hasIdentifier` e `isDescribedBy`, garantendo coerenza con l'intero network ArCo.

La modellazione degli strumenti di ricerca è gestita attraverso la classe `FindingAid`, collegata alle risorse mediante la proprietà `hasFundingAid`[168] che specializza `core:hasConcept`.

L'integrazione con il modulo Denotative-Description fornisce supporto per misurazioni fisiche attraverso `hasMeasurementCollection`, particolarmente rilevante per la consistenza delle risorse archivistiche, mentre la connessione con il sistema catalografico avviene attraverso una entità `Record` che descrive la collezione o il documento, permettendo la referenziazione di descrizioni catalografiche esistenti e supportando l'annotazione ICCD attraverso `iccdNormTag`. L'ontologia adotta una strategia di

[168] Si tratta, con ogni probabilità, di un errore di battitura per "FindingAid".

dipendenza controllata dalle ontologie italiane per la gestione di agenti, concetti temporali e localizzazioni geografiche, delegando queste funzionalità ai rispettivi moduli specializzati attraverso riferimenti diretti a `l0/Agent` e `l0/Topic` senza sviluppare specializzazioni specifiche del dominio.

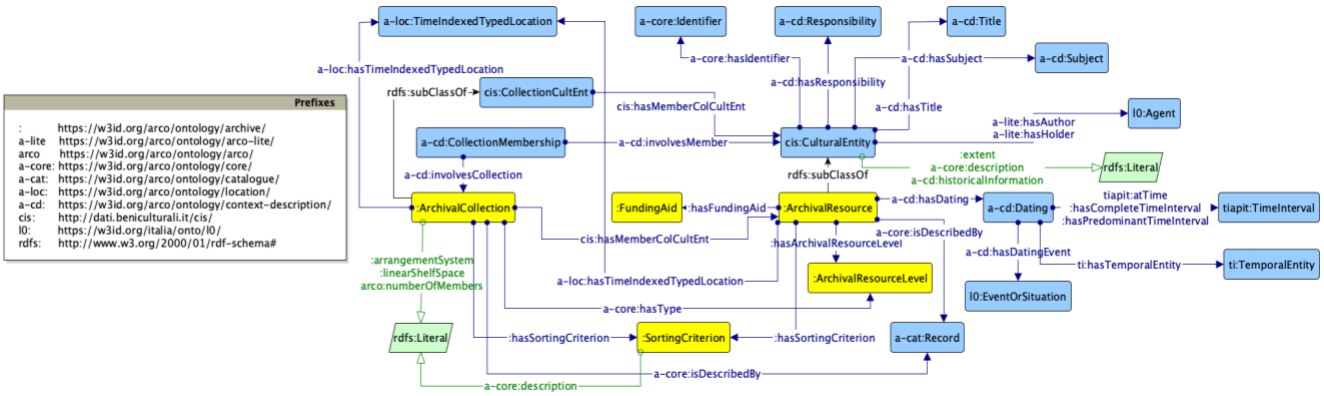


Figura 5.4. Rappresentazione grafica del modulo Archives di ArCo.

Dalla primavera 2025 l'Istituto Centrale per gli archivi (ICAR) sta promuovendo l'iniziativa ArCo4Archives con l'intento di estendere e stabilizzare ArCo per il dominio archivistico, facendo leva sul modello concettuale IcarImport2[169] e predisponendo bozze e materiali su canali pubblici per il confronto comunitario[170]. Su queste basi è iniziato un progetto di raffinamento del modulo ArCo archive che ha portato a un'evoluzione del modello concettuale, mantenendo la compatibilità con l'ecosistema ArCo ma introducendo alcune estensioni significative. Il modello risultante presenta una maggiore granularità descrittiva, con l'introduzione di specializzazioni locali per agenti e aspetti temporali che integrano le funzionalità precedentemente delegate ai moduli esterni. Sono state aggiunte classi specifiche per la gestione della digitalizzazione (`ArchivalDigitalizationProject`, `WebReference`) che rispondono alle esigenze contemporanee di gestione del patrimonio archivistico digitale, mentre il controllo degli accessi è stato formalizzato attraverso classi dedicate (`ConditionOfAccess`, `ConditionOfUse`). L'obiettivo di ArCo4Archives è creare un modello di riferimento per la produzione di un *knowledge graph* archivistico sulla base dei dati del Sistema Informativo Archivistico (SIA) (Istituto Centrale per gli Archivi - ICAR 2025).

Dal punto di vista metodologico, l'evoluzione di ArCo presenta alcuni tratti distintivi e raccomandabili, come la tendenza allo sviluppo di modelli modulari e *pattern-based* che facilitano il riuso e la coerenza

[169] https://icar.cultura.gov.it/attivita-e-progetti/progetti-icar-1/icar-import-2.
[170] https://github.com/IstitutoCentraleArchivi-ICAR/ICAR-Semantics.

concettuale; la produzione di documentazione aperta e guide (es. ArCo Primer Guide v. 1.0), così come di dump RDF ed endpoint LOD per abilitare l'uso e la sperimentazione da parte della comunità. Questi punti di forza hanno permesso ad ArCo di emergere come il più ampio e aggiornato *knowledge graph* italiano sul patrimonio culturale. Accanto a questi, permane la criticità di un'adozione locale e non pienamente armonizzata con modelli internazionali, in particolare in vista della graduale attestazione di RiC-O come modello di riferimento per la rappresentazione formale degli archivi. Il modulo ArCo archive del 2023, in particolare, pone una criticità nell'impostazione ancora strettamente analogica, che riflette gli standard passati (ISAD(G) in particolare) e non sembra pienamente ricettivo dell'autocritica recentemente effettuata dallo stesso ICA sulla rappresentatività degli standard tradizionali, la quale ha effettivamente condotto a un cambio di approccio, almeno parziale, nello sviluppo di RiC-O. È in questo contesto che si inserisce ArCo4Archives, che rappresenta un tentativo di revisione del modulo archivistico attraverso un confronto con gli sviluppi metodologici più recenti del settore: l'iniziativa si presenta come un progetto in via di maturazione, le cui potenzialità sono destinate a emergere pienamente attraverso il processo di confronto interdisciplinare e sperimentazione applicativa che caratterizzerà i prossimi sviluppi.

Se i modelli analizzati finora si concentrano principalmente sulla rappresentazione delle entità in termini di descrizione archivistica, un aspetto altrettanto cruciale riguarda la rappresentazione dei requisiti necessari a garantirne la sopravvivenza e l'affidabilità nel tempo. In una dimensione di confine tra descrizione e conservazione si colloca PREMIS (Preservation Metadata: Implementation Strategies), lo standard di riferimento per i metadati per la preservazione digitale. Per lungo tempo non è stato considerato uno standard specificamente rivolto agli archivi, data la sua focalizzazione sulla preservazione digitale; è stato tuttavia progressivamente riconosciuto come tale grazie alla crescente adozione da parte delle istituzioni GLAM e, soprattutto, in ragione del ruolo sempre più centrale del digitale nella produzione documentale. Questi fattori ne hanno consolidato la funzione di riferimento nella documentazione e nella gestione del ciclo di vita degli oggetti digitali.

All'inizio degli anni Duemila diventa evidente la necessità di uno standard di metadati dedicato alla conservazione a lungo termine delle risorse digitali. Con questa finalità, tra il 2003 e il 2005, sotto il patrocinio di Online Computer Library Center (OCLC)[171] e Research Libraries Group (RLG)[172], un gruppo di lavoro composto da oltre trenta rappresentanti di istituzioni culturali, enti governativi e organizzazioni private di diversi paesi ha elaborato il PREMIS Data Dictionary for Preservation

[171] https://www.oclc.org/.
[172] https://www.rlg.org/.

Metadata. Il rapporto, conclusivo dei lavori del gruppo, comprendeva sia un dizionario dei dati sia linee guida e raccomandazioni sui metadati di conservazione; una seconda versione è stata pubblicata nel marzo 2008.

Sin dalle sue origini, PREMIS è stato progettato con un principio chiave: garantire la massima flessibilità di implementazione. Per questo motivo, il Data Dictionary è stato concepito in maniera tecnicamente neutrale, senza presupporre specifici sistemi di archiviazione digitale, strategie di conservazione o processi di gestione dei metadati, nell'ottica di favorirne l'adozione in una grande varietà di contesti di preservazione (Lindlar 2022).

Il PREMIS Data Dictionary, giunto alla versione 3.0 nel 2015, fornisce un vocabolario strutturato e le specifiche implementative in formato XML/Schema, pensate per l'interscambio tra *repository* digitali (PREMIS Editorial Committee 2015b). Con la crescente importanza del Web Semantico, è emersa tuttavia la necessità di una formalizzazione in linguaggi adatti a garantire interoperabilità semantica e capacità inferenziali. A questo scopo è stata sviluppata la PREMIS Ontology, rappresentazione in OWL/RDF del modello, resa disponibile dalla Library of Congress[173].

Il PREMIS Data Dictionary definisce un modello concettuale basato su cinque entità fondamentali e tra loro interrelate (Figura 5.5); l'`Intellectual Entity` rappresenta l'unità concettuale di contenuto, come un libro o una collezione; l'`Object` incarna l'oggetto digitale vero e proprio, articolandosi nei tre livelli di `Representation`, `File` e `Bitstream`, dove i `File` costituiscono l'unità minima di conservazione, essendo sequenze di byte note al sistema operativo; gli `Agents` identificano i soggetti (persone, organizzazioni o software) che intervengono attivamente nei processi conservativi; gli `Events` documentano le azioni significative che modificano o verificano lo stato degli oggetti digitali; infine, i `Rights` codificano le autorizzazioni giuridiche essenziali per operare legalmente sugli oggetti. Questa struttura integrata permette di descrivere compiutamente non solo gli aspetti statici degli oggetti digitali, ma anche la dinamica dei processi e il contesto normativo che ne garantisce la preservazione a lungo termine.

[173] https://id.loc.gov/ontologies/premis-3-0-0.

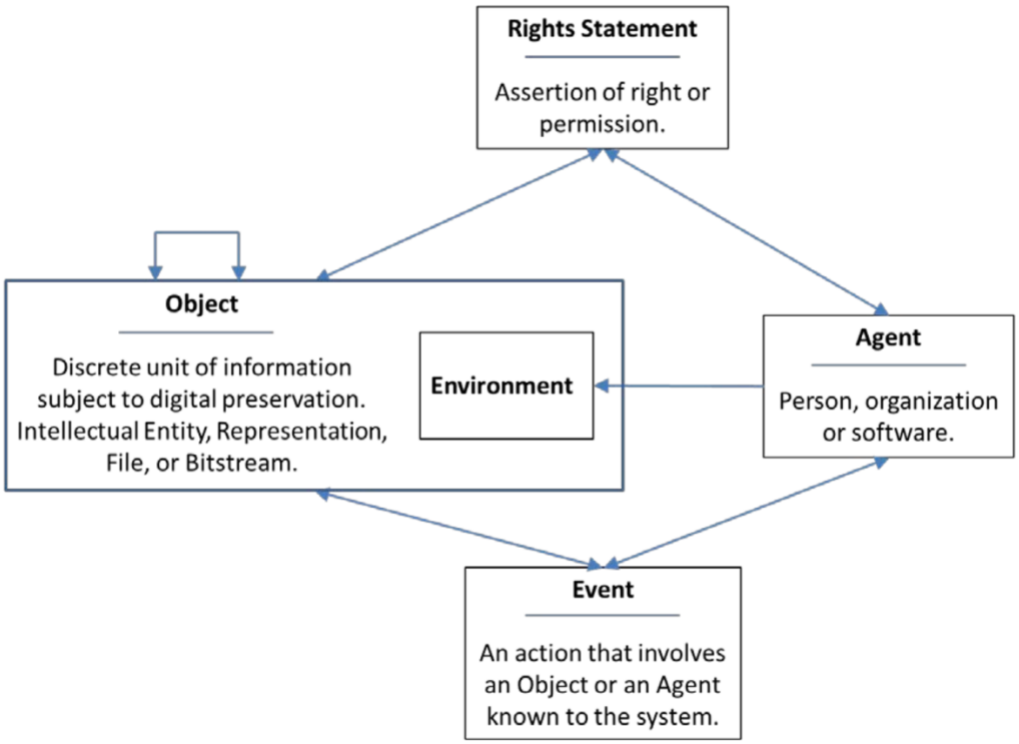


Figura 5.5. PREMIS Data Model (PREMIS Editorial Committee 2015a, 6).

L'ontologia formalizza le entità del Data Dictionary in OWL e comprende complessivamente 68 classi, 67 *object property* e 23 *data property*. La classe generale `premis:Object` rappresenta le unità di informazione soggette a conservazione ed è articolata nelle sottoclassi `premis:IntellectualEntity`, `premis:Representation`, `premis:File` e `premis:Bitstream`. La classe `premis:Event` descrive azioni che incidono sulla capacità di un *repository* di preservare oggetti digitali e include proprietà come `premis:outcome` e `premis:outcomeNote`, che documentano i risultati dell'evento. La classe `premis:Agent` modella attori umani e non umani, ulteriormente specializzati in `premis:Person`, `premis:Organization`, `premis:HardwareAgent` e `premis:SoftwareAgent`. La gestione dei diritti è rappresentata dalla classe `premis:RightsBasis`, che si articola in `premis:Copyright`, `premis:License`, `premis:Statute` e `premis:InstitutionalPolicy`, affiancata da `premis:RightsStatus`, che ne individua l'applicazione a uno specifico oggetto. Vi è poi `premis:PreservationPolicy`, che documenta decisioni e regole istituzionali sulla gestione della conservazione, e che comprende anche le `premis:SignificantProperties`, ovvero le caratteristiche ritenute essenziali da mantenere. Altre classi di particolare rilievo sono `premis:Fixity` e `premis:Signature`, che permettono di documentare l'integrità e l'autenticità degli oggetti.

Accanto alle classi, l'ontologia definisce un ricco insieme di proprietà RDF/OWL che traducono le regole del Data Dictionary in vincoli semantici. Alcune tra le più significative sono `premis:identifier`, che collega un'entità a un identificatore univoco; `premis:relationship`, che documenta relazioni tra oggetti digitali; `premis:Fixity`, che associa un file o un bitstream a un valore di checksum; `premis:storedAt`, che lega un oggetto al suo luogo di archiviazione; `premis:dependency` e `premis:purpose`, che descrivono le dipendenze funzionali tra oggetti; `premis:rightsStatus`, che connette un oggetto allo stato dei diritti; `premis:act` e `premis:restriction`, che definiscono le azioni consentite o vietate da una regola. Queste proprietà non servono solo a descrivere stati statici, ma a modellare il ciclo di vita dinamico degli oggetti digitali, in linea con il paradigma OAIS (ISO 14721:2012).

Negli ultimi anni l'attenzione della comunità legata a PREMIS si è concentrata in particolare sulla modellazione dei diritti legati agli oggetti digitali, comprendendo aspetti come la privacy, la proprietà e l'accesso. Il sottogruppo dedicato del PREMIS Editorial Committee ha presentato nel 2022, in occasione del convegno iPres a Glasgow, un white paper che delinea un approccio innovativo alla rappresentazione dei diritti, che confluirà nella versione 4.0 del PREMIS Data Dictionary, che prevede un sostanziale aggiornamento della *Rights Entity*.

La PREMIS Ontology rappresenta un'evoluzione decisiva, consentendo di passare da uno scambio puramente sintattico di metadati a una gestione semantica e inferenziale, in linea con le esigenze del Web Semantico. Per archivi e biblioteche digitali, lo standard consente di documentare in modo sistematico le informazioni necessarie a garantire la conservazione, l'integrità e l'autenticità degli oggetti digitali nel tempo. I metadati mappati da PREMIS dovrebbero inoltre essere considerati non più esclusivamente tecnici, ma parte integrante dei metadati descrittivi, rappresentando strumenti fondamentali per comprendere, contestualizzare e rendere pienamente fruibili le risorse digitali native, favorendo la loro interoperabilità, il riuso e la partecipazione attiva a reti informative globali.

Alla luce della ricerca condotta, dunque, emergono quattro principali modelli di riferimento – RiC-O, ArCo archive, ArchOnto e PREMIS – che è possibile ipotizzare come basi per la rappresentazione del materiale digitale d'autore. La Tabella 1 ne riassume sinteticamente le caratteristiche principali.

Caratteristiche	RiC-O	ArCo Archive	ArchOnto	PREMIS
Namespace IRI	https://www.ica.org/standards/RiC/ontology	https://w3id.org/arco/ontology/archive	ttps://purl.org/episa/archonto/1.0	http://www.loc.gov/premis/rdf/v3/
Fonte / Organizzazione	ICA	ICCD, ISTC-CNR, Università di Bologna	INESC TEC	Library of Congress
Versione	1.1 (2025)	0.2 (2023)	1.0 (2025)	3.0 (2015)
Classi	110	310	127	68
Object properties (OP)	485	608	315	67
Data properties (DP)	76	181	56	23
Livelli archivistici supportati	Tutti i livelli	Fondo, subfondo, complesso archivistico, serie, sottoserie, sottogruppo, unità archivistica, sottofascicolo, unità documentaria.	Tutti i livelli	n.a.
Stato	Stabile	Instabile	Instabile	Stabile
Focus	Descrizione risorse archivistiche	Descrizione risorse archivistiche	Descrizione risorse archivistiche	Preservazione risorse digitali
Adozione	Istituzioni archivistiche internazionali	Portale delle Fonti, ICCD	Arquivo Nacional da Torre do Tombo (ANTT) District Archive of Guarda District Archive of Porto District Archive of Bragança.	Istituzioni internazionali

Tabella 5.1 Confronto dei principali modelli di riferimento per la rappresentazione archivistica: RiC-O, ArCo Archive, ArchOnto e PREMIS.

PREMIS, essendo specificamente orientato al materiale digitale e non necessariamente archivistico, non è immediatamente comparabile ai modelli generali per la descrizione, ma verrà comunque considerato nello sviluppo del modello dedicato al *born-digital* descritto nel capitolo 5 alla luce del suo ruolo centrale nel contesto conservativo.

CIDOC CRM costituisce lo standard di riferimento per la rappresentazione semantica del patrimonio culturale a livello internazionale. Tuttavia, la sua architettura concettuale riflette principalmente le esigenze del dominio museale, mentre la complessità strutturale del modello ne aumenta la difficoltà di

implementazione rispetto a RiC-O. Considerato che l'adozione di RiC-O ha già evidenziato criticità nel settore archivistico, dovute al profondo cambio di paradigma che implica rispetto alla tradizione, l'applicazione di CIDOC CRM presenterebbe barriere ancora più elevate. L'estensione ArchOnto, pur fornendo una specializzazione per il dominio archivistico, introduce ulteriori livelli di astrazione che incrementano la complessità della modellazione.

RiC-O, d'altra parte, pur essendo un modello recente e ancora poco diffuso rispetto a CIDOC CRM, nasce con l'intento di rispondere in modo più specifico alle esigenze archivistiche. La sua progettazione da parte dell'ICA ne garantisce una legittimazione istituzionale e una solida base di sviluppo. Inoltre, l'architettura del modello è concepita per essere estensibile: è possibile proporre integrazioni ed estensioni alla comunità per una loro eventuale incorporazione nell'ontologia. Tuttavia, l'implementazione pratica evidenzia ancora limiti di maturità, soprattutto in termini di interoperabilità semantica e di effettiva adozione da parte delle comunità professionali.

ArCo4Archives, costruito sulla consolidata esperienza di ArCo, ha il potenziale per affermarsi come punto di riferimento nazionale per la rappresentazione LOD degli archivi italiani. Al contempo, un allineamento con le pratiche internazionali ne rafforzerebbe il ruolo e incrementerebbe l'interoperabilità, evitando che i dati del patrimonio culturale italiano rimangano isolati rispetto alle risorse archivistiche internazionali.

Alla luce di queste considerazioni, il capitolo 5.2 è dedicato alle prospettive comparative sulla modellizzazione semantica negli archivi, con l'obiettivo di porre le basi per un'armonizzazione tra RiC-O, ArCo e CIDOC CRM, nell'ottica di una rappresentazione coerente e integrata dei dati archivistici a livello nazionale e internazionale.

5.2 Prospettive comparative sulla modellazione semantica negli archivi

L'allineamento ontologico rappresenta un passaggio cruciale per garantire l'interoperabilità semantica tra modelli differenti, poiché consente di mettere in relazione concetti e strutture che condividono finalità comuni di descrizione, gestione e valorizzazione del patrimonio culturale. Un allineamento accurato non solo contribuisce a migliorare la qualità e la coerenza dei dati, riducendo ambiguità e ridondanze, ma ne agevola anche la riusabilità e l'integrazione in scenari più ampi di ricerca e fruizione, favorendo così l'interconnessione tra dataset eterogenei e la costruzione di ecosistemi informativi realmente interoperabili (San Emeterio de la Parte et al. 2025). In questa direzione, il rischio di "isolamento" rispetto a modelli e ontologie esistenti, come ha osservato Feliciati a proposito della bozza di RiC-O (0.2)

(Feliciati 2021, 100), costituisce un ostacolo concreto alla definizione di un linguaggio comune e condiviso.

Questo capitolo è dedicato all'analisi comparativa fra i modelli RiC-O, ArCo e CIDOC CRM (via ArchOnto). Diversi approcci sono stati proposti per valutare e confrontare modelli semantici, siano essi automatici, semi-automatici o manuali (Liu et al., 2021; Rudwan e Fonou-Dombeu, 2025; Amini et al., 2025). Per entrambi i casi di studio presentati in questo capitolo si è ritenuto opportuno procedere con un allineamento manuale, poiché la natura contestuale dei modelli analizzati richiede un'interpretazione critica che gli strumenti automatizzati non sarebbero stati in grado di garantire con il medesimo livello di consapevolezza. Tale scelta metodologica ha inoltre permesso di approfondire ulteriormente lo studio dei modelli, in vista dell'individuazione di quello più adeguato alla rappresentazione dei materiali nativi digitali.

La sezione 5.2.1 è dedicata all'allineamento ontologico fra RiC-O e ArCo, considerando, in particolare, il modulo archive[174]. In questo contesto, l'analisi ha previsto lo sviluppo concreto di un allineamento semantico attraverso lo studio approfondito delle definizioni concettuali e delle dipendenze ontologiche tra classi e proprietà dei due modelli. Ciò ha permesso di confrontare un'ontologia sviluppata a livello nazionale (ArCo) con un modello internazionale più trasversale (RiC-O), fornendo ipotesi di allineamento per armonizzare un'implementazione locale con quello che si sta affermando come prossimo standard internazionale.

La sezione 5.2.2, invece, propone un'analisi comparativa di RiC-O e ArchOnto basata sull'applicazione dei due modelli allo stesso caso di studio. Questo approccio consente di evidenziare le differenze e le corrispondenze nel modo in cui i modelli rappresentano i documenti archivistici e i loro contesti. Pur non producendo un allineamento formale, la comparazione fornisce elementi metodologici e operativi utili per un futuro allineamento tra i modelli.

5.2.1 RiC-O e ArCo

L'allineamento ontologico tra ArCo e RiC-O rappresenta un primo passo per avvicinare un modello sviluppato a livello nazionale a uno internazionale, nell'ottica di facilitare l'interoperabilità dei dati e la condivisione coerente delle informazioni tra istituzioni culturali (Banek et al. 2008; Moraitou et al. 2019).

[174] L'attività di allineamento fra RiC-O e ArCo è stata sviluppata durante il tirocinio presso l'Istituto centrale per la digitalizzazione del patrimonio culturale - Digital Library (1° ottobre 2023-31 marzo 2025) con la supervisione della dott.ssa Margherita Porena.

La procedura è stata supportata dall'utilizzo del software *open source* Protégé Desktop[175], un ambiente di editing ontologico basato su OWL che consente di esplorare e modificare le ontologie in un unico spazio di lavoro attraverso un'interfaccia utente personalizzabile. Ciò ha consentito di studiare agevolmente RiC-O ed ArCo nelle loro specificità e di ottenere un file di allineamento al termine del processo[176].

L'allineamento è stato sviluppato a partire dall'analisi dei concetti fondamentali del dominio archivistico nelle classi dei due modelli, estendendosi successivamente alle *object property* e alle *data property*. Gli allineamenti sono stati condotti sia come allineamenti diretti, tra concetti semanticamente equivalenti, sia come allineamenti gerarchici, tra concetti collegati da una relazione di estensione o restrizione semantica.

I risultati sintetici degli allineamenti sono riportati nelle tabelle:

- Tabella 5.2: allineamenti diretti fra classi;
- Tabella 5.3: allineamenti gerarchici fra classi;
- Tabella 5.4: allineamenti diretti fra *object property*;
- Tabella 5.5: allineamenti gerarchici fra *object property*;
- Tabella 5.6: allineamenti diretti fra *data property*.

[175] https://protege.stanford.edu/software.php.

[176] L'allineamento è stato effettuato sul file di base RiC-O version 1.1, importando classi e proprietà dal network di ontologie di ArCo, con specifico riferimento al modulo archive. Il file è disponibile nel *repository* dedicato al progetto, consultabile al link: https://github.com/LuciaGiagnolini12/rico-arco-alignment. Nel testo, in riferimento a classi e proprietà, si fa riferimento al seguente uso di prefissi: rico: https://www.ica.org/standards/RiC/ontology#; arco-core: https://w3id.org/arco/ontology/core/; arco-cd: https://w3id.org/arco/ontology/context-description/; a-loc: https://w3id.org/arco/ontology/location/; arco-dd: https://w3id.org/arco/ontology/denotative-description/; arco-archive: https://w3id.org/arco/ontology/archive/; a-cat: https://w3id.org/arco/ontology/catalogue/; arco-lite: https://w3id.org/arco/ontology/arco-lite/; cis: http://dati.beniculturali.it/cis/; cpv: https://w3id.org/italia/onto/CPV/; onto-language: https://w3id.org/italia/onto/Language/; l0: https://w3id.org/italia/onto/l0/; clv: https://w3id.org/italia/onto/CLV/; TI: https://w3id.org/italia/onto/TI/.

Classe RiC-O	Classe ArCo
rico:Identifier https://www.ica.org/standards/RiC/ontology#Identifier	arco-core:identifier https://w3id.org/arco/ontology/core/Identifier
rico:Language https://www.ica.org/standards/RiC/ontology#Language	onto-language:Language https://w3id.org/italia/onto/Language
rico:Person https://www.ica.org/standards/RiC/ontology#Person	cpv:Person https://w3id.org/italia/onto/CPV/Person
rico:Title https://www.ica.org/standards/RiC/ontology#Title	arco-cd:Title https://w3id.org/arco/ontology/context-description/Title
rico:Activity https://www.ica.org/standards/RiC/ontology#Activity	l0:Activity https://w3id.org/italia/onto/l0/Activity

Tabella 5.2. Allineamenti diretti fra classi di RiC-O e ArCo.

Classe RiC-O	Classe ArCo
rico:Record (sottoclasse) https://www.ica.org/standards/RiC/RiC-O_v0-2.html#Record	cis:CulturalEntity (superclasse) http://dati.beniculturali.it/cis/CulturalEntity
rico:Instatiation (sottoclasse) https://www.ica.org/standards/RiC/RiC-O_v0-2.html#Instantiation	cis:CultualEntity (superclasse) http://dati.beniculturali.it/cis/CulturalEntity
rico:CarrierExtent (sottoclasse) https://www.ica.org/standards/RiC/ontology#CarrierExtent	arco-dd:TechnicalCharacteristic (superclasse) https://w3id.org/arco/ontology/denotative-description/TechnicalCharacteristic
rico:PlaceType (superclasse) https://www.ica.org/standards/RiC/ontology#PlaceType	a-loc:LocationType (sottoclasse) https://w3id.org/arco/ontology/location/LocationType

Tabella 5.3. Allineamenti gerarchici fra classi di RiC-O e ArCO.

Object property RiC-O	Object property ArCo
rico:isOrWasIdentifierOf https://www.ica.org/standards/RiC/ontology#isOrWasIdentifierOf	arco-core:isIdentifierOf https://w3id.org/arco/ontology/core/isIdentifierOf
rico:hasOrHadIdentifier https://www.ica.org/standards/RiC/ontology#hasOrHadIdentifier	arco-core:hasIdentifier https://w3id.org/arco/ontology/core/hasIdentifier

rico:isEventAssociatedWith https://www.ica.org/standards/RiC/ontology#isEventAssociatedWith	arco-core:involves https://w3id.org/arco/ontology/core/involves
rico:isAssociatedWithEvent https://www.ica.org/standards/RiC/ontology#isAssociatedWithEvent	arco-core:isInvolvedIn https://w3id.org/arco/ontology/core/isInvolvedIn
rico:hasOrHadTitle https://www.ica.org/standards/RiC/ontology#hasOrHadTitle	arco-cd:hasTitle https://w3id.org/arco/ontology/context-description/hasTitle
rico:isOrWasTitleOf https://www.ica.org/standards/RiC/ontology#isOrWasTitleOf	arco-cd:isTitleOf https://w3id.org/arco/ontology/context-description/isTitleOf
rico:hasBirthPlace https://www.ica.org/standards/RiC/ontology#hasBirthPlace	cpv:hasBirthPlace https://w3id.org/italia/onto/CPV/hasBirthPlace
rico:isBirthPlaceOf https://www.ica.org/standards/RiC/ontology#isBirthPlaceOf	cpv:isBirthPlaceof https://w3id.org/italia/onto/CPV/isBirthPlaceOf
rico:hasOrHadSpouse https://www.ica.org/standards/RiC/ontology#hasOrHadSpouse	cpv:isConsortOf https://w3id.org/italia/onto/CPV/isConsortOf
rico:isChildOf https://www.ica.org/standards/RiC/ontology#isChildOf	cpv:isChildOf https://w3id.org/italia/onto/CPV/isChildOf
rico:hasChild https://www.ica.org/standards/RiC/ontology#hasChild	cpv:isParentOf https://w3id.org/italia/onto/CPV/isParentOf
rico:knows https://www.ica.org/standards/RiC/ontology#knows	cpv:knows https://w3id.org/italia/onto/CPV/knows
rico:hasDeathPlace https://www.ica.org/standards/RiC/ontology#hasDeathPlace	cpv:hasDeathPlace https://w3id.org/italia/onto/CPV/hasDeathPlace
rico:isDeathPlaceof https://www.ica.org/standards/RiC/ontology#isDeathPlaceOf	cpv:isDeathPlaceOf https://w3id.org/italia/onto/CPV/isDeathPlaceOf
rico:performsOrPerformed https://www.ica.org/standards/RiC/ontology#performsOrPerformed	arco-cd:isActivityOperatorOf https://w3id.org/arco/ontology/context-description/isActivityOperatorOf

rico:isOrWasPerformedBy https://www.ica.org/standards/RiC/ontology#isOrWasPerformedBy	arco-cd:hasActivityOperator https://w3id.org/arco/ontology/context-description/hasActivityOperator
rico:hasOrHadLanguage https://www.ica.org/standards/RiC/ontology#hasOrHadLanguage	onto-language:hasLanguage https://w3id.org/italia/onto/Language/hasLanguage
rico:isOrWasLanguageOf https://www.ica.org/standards/RiC/ontology#isOrWasLanguageOf	onto-language:isLanguageOf https://w3id.org/italia/onto/Language/isLanguageOf
rico:precededOrPrecedes https://www.ica.org/standards/RiC/ontology#precedesOrPreceded	l0:precedes https://w3id.org/italia/onto/l0
rico:followsOrFollowed https://www.ica.org/standards/RiC/ontology#followsOrFollowed	l0:follows https://w3id.org/italia/onto/l0
rico:hasAuthor https://www.ica.org/standards/RiC/ontology#hasAuthor	arco-lite:hasAuthor https://w3id.org/arco/ontology/arco-lite/hasAuthor
rico:isAuthorOf https://www.ica.org/standards/RiC/ontology#isAuthorOf	arco-lite:isAuthorOf https://w3id.org/arco/ontology/arco-lite/isAuthorOf
rico:hasOrHadPart https://www.ica.org/standards/RiC/ontology#hasOrHadPart	arco-core:hasPart https://w3id.org/arco/ontology/core/hasPart
rico:IsOrWasPartOf https://www.ica.org/standards/RiC/ontology#isOrWasPartOf	arco-core:isPartOf https://w3id.org/arco/ontology/core/isPartOf
rico:hasOrHadHolder https://www.ica.org/standards/RiC/ontology#hasOrHadHolder	arco-lite:hasHolder https://w3id.org/arco/ontology/arco-lite/hasHolder
rico:isOrWasHolderOf https://www.ica.org/standards/RiC/ontology#isOrWasHolderOf	arco-lite:isHolderOf https://w3id.org/arco/ontology/arco-lite/isHolderOf

Tabella 5.4. Allineamenti diretti fra object property di RiC-O e ArCo.

Object property RiC-O	Object property ArCo
rico:hasOrHadPhysicalLocation (superproprietà) https://www.ica.org/standards/RiC/ontology#hasOrHadPhysicalLocation	a-loc:atLocation (sottoproprietà) https://w3id.org/arco/ontology/location/atLocation
rico:isOrWasPhysicalLocationOf (superproprietà) https://www.ica.org/standards/RiC/ontology#isOrWasPhysicalLocationOf	a-loc:isLocationIn (sottoproprietà) https://w3id.org/arco/ontology/location/isLocationIn
rico:describesOrDescribed (superproprietà) https://www.ica.org/standards/RiC/ontology#describesOrDescribed	a-cat:describes (sottoproprietà) https://w3id.org/arco/ontology/catalogue/describes
rico:isOrWasDescribedBy (superproprietà) https://www.ica.org/standards/RiC/ontology#isOrWasDescribedBy	a-cat:isDescribedBy (sottoproprietà) https://w3id.org/arco/ontology/catalogue/isDescribedBy
rico:hasOrHadSubject (superproprietà) https://www.ica.org/standards/RiC/ontology#hasOrHadSubject	arco-cd:hasSubject (sottoproprietà) https://w3id.org/arco/ontology/context-description/hasSubject
rico:isOrWasSubjectOf (superproprietà) https://www.ica.org/standards/RiC/ontology#isOrWasSubjectOf	arco-cd:isSubjectOf (sottoproprietà) https://w3id.org/arco/ontology/context-description/isSubjectOf
rico:hasExtent (sottoproprietà) https://www.ica.org/standards/RiC/ontology#hasExtent	arco-dd:hasTechnicalCharacteristic (superproprietà) https://w3id.org/arco/ontology/denotative-description/hasTechnicalCharacteristic
rico:isExtentOf (sottoproprietà) https://www.ica.org/standards/RiC/ontology#isExtentOf	arco-dd:isTechnicalCharacteristicOf (superproprietà) https://w3id.org/arco/ontology/denotative-description/isTechnicalCharacteristicOf
rico:isOrWasTypeOf (superproprietà) https://www.ica.org/standards/RiC/ontology#isOrWasTypeOf	arco-core:isTypeOf (sottoproprietà) https://w3id.org/arco/ontology/core/isTypeOf
rico:hasOrHadType(superproprietà) https://www.ica.org/standards/RiC/ontology#hasOrHadType	arco-core:hasType (sottoproprietà) https://w3id.org/arco/ontology/core/hasType

Tabella 5.5. Allineamenti gerachici fra object property di RiC-O e ArCo.

Data property RiC-O	Data property ArCo
rico:modificationDate https://www.ica.org/standards/RiC/ontology#modificationDate	TI:modified https://w3id.org/italia/onto/TI/modified
rico:Date https://www.ica.org/standards/RiC/ontology#date	TI:date https://w3id.org/italia/onto/TI/date
rico:generalDescription https://www.ica.org/standards/RiC/ontology#generalDescription	arco-core:description https://w3id.org/arco/ontology/core/description
rico:name https://www.ica.org/standards/RiC/ontology#name	l0:name https://w3id.org/italia/onto/l0/name
rico:physicalOrLogicalExtent https://www.ica.org/standards/RiC/ontology#physicalOrLogicalExtent	arco-archive:extent https://w3id.org/arco/ontology/archive/extent

Tabella 5.6. Allineamenti diretti fra data property di RiC-O e ArCo.

Di seguito, si presentano le considerazioni principali che hanno guidato le decisioni di allineamento, evidenziando i casi in cui è stato possibile effettuare corrispondenze e quelli in cui, per ragioni semantiche o strutturali, gli allineamenti sono stati esclusi.

Concetto di oggetto culturale

Come abbiamo visto, in RiC-O un oggetto culturale (più specificamente, una risorsa archivistica) è rappresentato distintamente con due classi, a seconda che si identifichi il suo contenuto informativo, per il quale si utilizza la classe `rico:RecordResource` o una sua istanza fisica, per cui si utilizza la classe `rico:Instantiation`. `rico:RecordResource` possiede tre sottoclassi che ne restringono e specificano il significato:

- `rico:Record`: «discrete information content formed and inscribed, at least once, by any method on any carrier in any persistent, recoverable form by an Agent in the course of life or work activity»[177].
- `rico:RecordSet`: «one or more records that are grouped together by an Agent based on the records sharing one or more attributes or relations».

[177] Le descrizioni di classi, *object property* e *data property* di RiC-O qui riportate provengono dalla documentazione ufficiale dell'ontologia, disponibile a link: https://www.ica.org/standards/RiC/RiC-O_1-1.html.

- `rico:RecordPart`: «component of a Record with independent information content that contributes to the intellectual completeness of the Record».

In ArCo non esistono distinzioni fra il contenuto informativo e la sua istanza fisica come in RiC-O, ma si fa riferimento direttamente all'oggetto culturale, in cui convivono entrambi gli aspetti. Per rappresentare un oggetto culturale, ArCo ha modellato diverse classi a seconda delle iniziali esigenze di ICCD, sulla base della strutturazione dei dati delle schede di catalogo. Il modulo ArCo archive individua i documenti archivistici come `cis:CulturalEntity` e, in particolare, come `arco-archive:ArchivalResource`. `arco-archive:ArchivalResource` rappresenta una generica risorsa archivistica, a prescindere dal suo livello di descrizione, che invece viene specificato tramite `arco-archive:ArchivalResourceLevel`.

Inoltre, è importante sottolineare che il termine e il concetto di "record" nelle due ontologie è utilizzato in maniera diversa: in RiC-O il record rimanda al contenuto informativo iscritto all'interno di una risorsa istanziata come, ad esempio, il contenuto testuale e informativo di un documento; in ArCo il record è il record descrittivo-catalografico di una risorsa.

Se, da un lato, `arco-archive:ArchivalResource` appare utilizzata in maniera analoga a `rico:RecordResource`, in quanto rappresenta una risorsa che si specifica progressivamente a livello di contenuto informativo, dall'altro lato il riferimento alle caratteristiche fisiche dell'oggetto la avvicina piuttosto a `rico:Instantiation`. Considerando l'uso di `rico:RecordResource` e di `rico:Instantiation`, si osserva che, pur condividendo gran parte delle *object* e *data property*, alcune proprietà fondamentali per la descrizione archivistica hanno come dominio o range esclusivamente l'una o l'altra classe. Ne consegue che `arco-archive:ArchivalResource` non può essere allineata in modo diretto né con `rico:RecordResource` né con `rico:Instantiation`. L'unica ipotesi plausibile è dunque un allineamento di tipo gerarchico, in cui `rico:RecordResource` e `rico:Instantiation` siano considerate sottoclassi di `arco-archive:ArchivalResource`, che funge da concetto più astratto capace di ricomprendere entrambe le dimensioni.

Concetti di soggetto produttore, autore e provenance

In RiC-O non esiste un concetto di attribuzione autoriale (*authorship attribution*) nel senso proprio dell'ambito artistico-letterario. Nella descrizione archivistica, infatti, l'attribuzione delle informazioni ricade generalmente sull'archivista, che è generalmente il primo a determinare e registrare i dati relativi all'oggetto. RiC-O prevede tuttavia la classe `rico:CreationRelation`, che collega un `rico:Record` o una `rico:Instantiation` a un `rico:Agent` responsabile della loro creazione, specificando il ruolo

svolto (ad esempio autore, incisore o tipografo) tramite la *object property* `rico:CreationWithRole`, che mette in relazione `rico:CreationRelation` e `rico:RoleType`.

All'interno di questo quadro, `rico:AuthorshipRelation` è una specializzazione di `rico:CreationRelation` che limita la relazione all'autore e al contenuto informativo (`rico:Record`), e dunque non alle istanze fisiche. Tale classe non va confusa con `rico:AccumulationRelation`, che esprime invece l'acquisizione di un oggetto da parte di un soggetto nell'ambito delle proprie attività o con finalità di collezione. Con `rico:AuthorshipRelation` RiC-O intende esplicitare che il soggetto produttore di un archivio può non coincidere con l'autore dei documenti che esso contiene.

Le relazioni `rico:CreationRelation` e `rico:AuthorshipRelation` sono classi n-arie che formalizzano una relazione orientata; sono collegate al record o all'istanza tramite le proprietà `rico:creationRelationHasSource` e `rico:authorshipRelationHasSource`, e all'agente tramite `rico:CreationRelationHasTarget` e `rico:AuthorshipRelationHasTarget`.

Su queste basi, un possibile allineamento con ArCo riguarda `arco-cd:Responsibility`, che generalizza il concetto di `arco-cd:AuthorshipAttribution` (classe deprecata nella versione più recenti di ArCo), estendendolo oltre il dominio ABAP (Archeologia, Belle Arti e Paesaggio). `arco-cd:Responsibility` rappresenta infatti l'attribuzione di una responsabilità a un agente in relazione al ruolo svolto o a un intervento effettuato su un'entità culturale.

Per individuare il corretto punto di contatto con RiC-O, occorre considerare che `rico:CreationRelation` è sottoclasse di `rico:OrganicProvenanceRelation`, la quale collega una `rico:RecordResource` o una `rico:Instantiation` a un agente che la crea, accumula, riceve o invia. Ne consegue che l'allineamento più coerente sia di tipo gerarchico: `arco-cd:Responsibility` come concetto più ampio, e `rico:OrganicProvenanceRelation` come sua specializzazione. La prima, infatti, aspetti di provenienza intesi da RiC-O e ulteriori forme di responsabilità, come ad esempio il restauro di un bene culturale.

Concetto di evento

Il concetto di evento in ArCo è modellato tramite l'ontologia OntoPiA, che definisce `l0:EventOrSituation` come «un'entità che si sviluppa nel tempo, nel mondo fisico o sociale; esempi sono fenomeni atmosferici, concerti, viaggi, processi istituzionali. Rappresenta inoltre un contesto relazionale costruito da un osservatore sulla base di un frame. Gli eventi si oppongono agli oggetti, in

quanto scorrono nel tempo aggregando più oggetti, mentre gli oggetti tendono a mantenersi stabili entro un certo intervallo temporale»[178].

In maniera analoga, RiC-O definisce `rico:Event` come un evento che può essere causato dalla natura, da un agente o da una combinazione di entrambi, caratterizzato da confini temporali e spaziali. Un evento può coinvolgere attivamente agenti, incidere su qualsiasi entità, essere puntuale o esteso nel tempo, comprendere altri eventi come parti o concatenarsi ad altri eventi in sequenze temporali. Più agenti possono partecipare a uno stesso evento con ruoli differenti.

Dalle due definizioni emerge chiaramente una convergenza concettuale: entrambe le classi descrivono entità processuali e relazionali, dotate di durata e contesto, capaci di coinvolgere molteplici agenti e oggetti. È pertanto possibile proporre un allineamento diretto tra `l0:EventOrSituation` e `rico:Event`.

Concetto di attività

Il concetto di attività in ArCo viene modellato tramite l'ontologia OntoPiA, che definisce `l0:Activity` come «the class of activities carried out by an agent». In RiC-O, `rico:Activity` è una sottoclasse di `rico:Event` e viene definita come «the doing of something for a human purpose». Nonostante ArCo non preveda l'attività come sottoclasse di un evento, `l0:Activity` e `arco-core:EventOrSituation` non sono disgiunte. Dunque, poiché concettualmente le due classi esprimono la stessa semantica, possiamo ipotizzare un allineamento fra `l0:Activity` e `rico:Activity`.

Concetto di acquisizione

In ArCo il concetto di acquisizione è modellato da una classe dedicata, `arco-cd:Acquisition`, sottoclasse di `arco-core:EventOrSituationInTime` e, più in generale, di `arco-core:EventOrSituation`. In RiC-O, invece, è presente la classe generica `rico:Event`, senza sottoclassi predefinite (a parte `rico:Activity`). La tipizzazione degli eventi avviene attraverso `rico:EventType`, che consente di associare, sotto forma di *literal*, diverse tipologie di eventi di curatela, quali creazione, acquisizione, trasferimento, descrizione, digitalizzazione, e così via.

[178] Le descrizioni di classi, *object property* e *data property* di OntoPiA qui riportate provengono dalla documentazione ufficiale dell'ontologia, disponibile a link: https://schema.gov.it/lode/extract?url=https://w3id.org/italia/onto/l0.

Ne deriva che non è possibile proporre un allineamento diretto tra `arco-cd:Acquisition` e una classe specifica di RiC-O, ma soltanto un allineamento di tipo gerarchico, in cui `rico:Event` funge da superclasse rispetto a `arco-cd:Acquisition`[179].

Concetto di localizzazione

In RiC-O si individuano cinque classi principali che rappresentano concetti legati alla localizzazione:

- `rico:Place`: «Bounded, named geographic area or region. May have beginning or end date. Scope Note: A jurisdiction is the bounded geographic area within which an Agent has the authority to perform specified activities constrained by rules. Jurisdictions, man-made structures, and natural features are historical entities. A Place thus may have a beginning date and ending date and changing boundaries that result from human or natural events. A Place may be systematically referenced to a location on the earth (geographic coordinates)»;
- `rico:PlaceRelation`: «Connects a Place and at least one Thing when the first is associated with the existence and lifecycle of the second one. The Place is the source of the Relation and the Thing(s) is the target»;
- `rico:PlaceType`: «Broadly, a Place may be a member of three broad categories: jurisdiction, manmade structure, or a natural feature. Each of these three categories can subdivided into narrower categories»;
- `rico:PhysicalLocation`: «A delimitation of the physical territory of a Place»; «Used to describe basic human-readable text such as an address, a cadastral reference, or less precise information found in a record. Use the coordinates datatype property, or the Coordinates class to capture the geographical coordinates of the Place. Use spatial relations (particularly 'has or had location') to capture a relation between two places»;
- `rico:Coordinates`: «Longitudinal and latitudinal information about a Place».

In ArCo, i concetti spaziali sono principalmente modellati attraverso l'ontologia Address (Location) Ontolgy (CLV), che prevede:

[179] La scelta di modellazione di ArCo riflette esigenze specifiche dell'ICCD, che non ha formalizzato eventi ulteriori del *workflow* documentale oltre al concetto di acquisizione. Da un punto di vista concettuale, sarebbe opportuno prevedere ulteriori sottoclassi per rappresentare operazioni quali, ad esempio, la digitalizzazione o lo scarto; tuttavia, tale estensione non è stata finora implementata. La presenza di una classe autonoma per l'acquisizione, d'altro canto, permette di tipizzare in maniera più precisa un evento che, per natura, può assumere forme e caratteristiche molto diverse.

- `clv:Feature`: «This class is used to represent any geographical entity»[180];
- `clv:Geometry`: «This class represents the geometry of a spatial entity».

Accanto a CLV, ArCo utilizza anche elementi dall'ontologia core e dall'ontologia della localizzazione:

- `arco-core:Location`: «Questa classe rappresenta la posizione (e.g. di un 'entità culturale)»[181];
- `a-loc:TimeIndexedTypedLocation`: «Questa classe rappresenta una localizzazione di un bene culturale, inserita in un arco temporale e qualificata in base al ruolo che la localizzazione riveste nei confronti del bene culturale»[182];
- `a-loc:LocationType`: «Questa classe rappresenta il tipo di localizzazione di un bene culturale».

Dall'analisi comparativa emergono quattro possibili allineamenti:

1. **Coordinate e geometrie**. Allineamento gerarchico fra `clv:Geometry` (superclasse) e `rico:Coordinates`, poiché la prima include le coordinate ma rappresenta anche geometrie più complesse;
2. **Localizzazione fisica**. Allineamento diretto fra `clv:Feature` e `rico:PhysicalLocation`, in quanto entrambe descrivono la localizzazione fisica di un luogo, e possono essere collegate a una classe che rappresenta la geometria;
3. **Tipizzazione della localizzazione**. Allineamento gerarchico fra `rico:PlaceType` (superclasse) e `a-loc:LocationType`. Il primo tipizza genericamente qualunque luogo, il secondo tipizza nello specifico la localizzazione di un bene culturale;
4. **Relazioni di localizzazione**. Allineamento gerarchico fra `rico:PlaceRelation` (superclasse) e `a-loc:TimeIndexedTypedLocation`. Entrambe descrivono una relazione n-aria che lega un bene a una localizzazione, ma con alcune differenze: in RiC-O la relazione non è tipizzata

[180] Le descrizioni di classi, *object property* e *data property* di CLV qui riportate provengono dalla documentazione ufficiale dell'ontologia, disponibile a link: https://schema.gov.it/lode/extract?url=https://w3id.org/italia/onto/CLV.

[181] Le descrizioni di classi, *object property* e *data property* dell'ontologia core di ArCo qui riportate provengono dalla documentazione ufficiale dell'ontologia, disponibile a link: https://dati.beniculturali.it/lode/extract?lang=it&url=https://raw.githubusercontent.com/ICCD-MiBACT/ArCo/master/ArCo-release/ontologie/core/core.owl.

[182] Le descrizioni di classi, *object property* e *data property* dell'ontologia della localizzazione di ArCo qui riportate provengono dalla documentazione ufficiale dell'ontologia, disponibile a link: https://w3id.org/arco/ontology/location.

temporalmente e non dipende da `rico:PlaceType`; in ArCo, invece, la temporalità e la connessione al bene culturale sono elementi intrinseci.

In coerenza con l'allineamento delle classi, è possibile allineare anche le proprietà che mettono in relazione un luogo e la sua localizzazione geografica (`rico:hasOrHadPhysicalLocation` e `a-loc:atLocation` e le relative proprietà inverse `rico:isOrWasPhysicalLocationOf` e `a-loc:isLocationIn`).

In generale, si può affermare che RiC-O propone una modellazione più ampia e astratta del concetto di luogo, mentre ArCo lo specifica in funzione della descrizione e della gestione dei beni culturali.

Per definire la collocazione fisica di un documento, ArCo modella la collocazione di edizioni, inventari e documentazione (intesa come documentazione di corredo) e materiale correlato al bene culturale rispettivamente tramite le *data property* `arco-cd:editionLocation`, `arco-cd:inventoryLocation`, `arco-cd:documentationLocation`, `arco-cd:relatedWorkLocation`. Manca, tuttavia, una *data property* che sia in grado di esprimere la collocazione (e.g. locale, scaffale, busta, fascicolo) di un bene archivistico.

In RiC-O non abbiamo *data property* che veicolino questo significato, ma per esprimere questa informazione è possibile orientarsi verso le classi `rico:IdentifierType` e `rico:PlaceType`. In questo caso, nonostante sia un dato rilevante per la rappresentazione di risorse archivistiche, questo tipo di relazione non è allineabile perché è stata modellata in maniera completamente diversa.

Concetto di group

In RiC-O, la classe `rico:Group` si articola nelle due sottoclassi `rico:CorporateBody`, che rappresenta il concetto di organizzazione, e `rico:Family`, relativa al concetto di famiglia. In ArCo, invece, si trovano due classi distinte: `cov:Group` e `arco-core:Organization`.

L'allineamento più immediato sembrerebbe quello tra `arco-core:Organization` e `rico:CorporateBody`. Tuttavia, poiché in ArCo le classi `arco-core:Organization` e `cov:Group` sono dichiarate disgiunte, questo allineamento non risulta percorribile senza violare la coerenza ontologica.

Un'ulteriore differenza riguarda il concetto di famiglia: mentre RiC-O gli dedica una classe specifica (`rico:Family`), in ArCo esso non è formalizzato autonomamente, ma compare unicamente come riferimento testuale nella nota descrittiva associata a `l0:Agent`.

Concetto di lasso temporale

Per i concetti temporali, ArCo utilizza l'Ontologia del Tempo (TI)[183], che presenta le classi: `TI:DayOfTheWeek`, `TI:Duration`, `TI:MonthOfTheYear`, `TI:TemporalEntity`, `TI:TimeIndexedEvent`, `TI:TimeInstant`, `TI:TimeInterval`, `TI:Year`. `TI:TemporalEntity` viene definita semplicemente come "entità temporale" ed è superclasse di tutte le altre, ad eccezione di `TI:TimeIndexedEvent`.

In RiC-O, in relazione alla definizione del tempo, è presente una sola classe, `rico:Date`, che viene definita come «Chronological information associated with an entity that contributes to its identification and contextualization». `rico:Date` viene tipizzata sia tramite la classe `rico:DateType`, che tramite le *object property* che la collegano ad altre entità (come `rico:isEndDateOf`, `rico:isCreationDateOf`, `rico:isLastUpdateDateOf` ecc.). In considerazione di queste premesse, possiamo effettuare un allineamento gerachico fra `TI:TemporalEntity` (superclasse) e `rico:Date`.

Concetto di condizione giuridica del bene

Per rappresentare la condizione giuridica di un bene, in ArCo è prevista la classe `arco-cd:LegalSituation`, sottoclasse di `arco-core:EventOrSituationInTime`, che viene poi tipizzata da `arco-cd:LegalSituationType`, sottoclasse di `arco-core:Type`.

In RiC-O, invece, è presente un'unica classe, `rico:LegalStatus`, sottoclasse di `rico:Type`. La differenza principale consiste nel fatto che in ArCo esiste una classe aggiuntiva per modellare la relazione del bene con la sua condizione giuridica (`arco-cd:LegalSituation`), mentre in RiC-O il concetto è direttamente associato ad altre entità quali `rico:Agent`, `rico:Record`, `rico:RecordPart` e `rico:RecordSet`. Alla luce di ciò, l'allineamento più appropriato può essere effettuato fra le classi di tipo *Type*, ossia tra `arco-cd:LegalSituationType` e `rico:LegalStatus`. Le rispettive proprietà e modalità di collegamento agli altri elementi non sono compatibili, e pertanto non è possibile effettuare un allineamento tra le proprietà stesse.

Concetto di software

In ArCo, il concetto di software è rappresentato dalla classe `arco-dd:Software`, mentre in RiC-O esiste la classe `rico:Mechanism`. Sebbene `arco-dd:Software` sia classificato come *Object* e `rico:Mechanism` come *Agent*, un allineamento risulta comunque possibile. Infatti, la definizione di `l0:Object` in ArCo

[183] Le descrizioni di classi, *object property* e *data property* di TI qui riportate provengono dalla documentazione ufficiale dell'ontologia, disponibile a link: https://w3id.org/italia/onto/TI .

riporta «Any entity that tends to be stable over a more or less long time, both in the physical and social world» ed è sufficientemente ampia da includere anche entità agentive.

In RiC-O, la definizione di `rico:Mechanism` afferma: «A mechanism may have both mechanical and software components or may be exclusively software». Da ciò emerge un possibile allineamento gerarchico in cui `arco-dd:Software` può essere considerato una sottoclasse di `rico:Mechanism`.

Concetto di appartenenza a una collezione

In ArCo, l'appartenenza di un bene culturale a una collezione è modellata tramite `arco-cd:CollectionMembership`, che descrive l'appartenenza di un bene culturale a una collezione in un dato intervallo temporale.

In RiC-O, esistono le classi `rico:RecordResourceToInstantiationRelation` e `rico:RecordResourceToRecordResourceRelation`. Il concetto di `rico:RecordResource` include anche `rico:RecordSet`, che rappresenta raggruppamenti di record come serie o fondi. La relazione fra un documento e la sua serie o fondo di appartenenza viene modellata tramite la property `rico:includesOrIncluded` fra `rico:Record` e `rico:RecordSet`, non fra `rico:Instantiation` e `rico:RecordSet`.

Ne consegue che, se da un lato `rico:RecordResourceToRecordResourceRelation` è una relazione generica che potrebbe essere estesa per rappresentare l'appartenenza, con un possibile allineamento gerarchico che vedrebbe `arco-cd:CollectionMembership` come sottoclasse, dall'altro lato l'allineamento non è fattibile. Questo perché `rico:RecordResourceToRecordResourceRelation` connette due contenuti informativi, mentre `arco-cd:CollectionMembership` collega una o più `cis:CulturalEntity` a una collezione, includendo quindi anche l'istanza fisica.

In sintesi, il concetto di appartenenza di un oggetto culturale a una collezione o serie non trova un allineamento diretto, nemmeno a livello di *object property*, a meno di prediligere un approccio generico in entrambe le ontologie, utilizzando le proprietà `rico:isOrWasPartOf` / `rico:hasOrHadPart` e `arco-core:isPartOf` / `arco-core:hasPart`.

Concetto di consistenza

Per quanto riguarda il concetto di consistenza, è possibile proporre un allineamento gerarchico fra `arco-dd:TechnicalCharacteristic` (superclasse) e `rico:Extent` (sottoclasse). Infatti, `arco-dd:TechnicalCharacteristic` viene definita come una caratteristica tecnica relativa a un bene culturale. Per esempio, può rappresentare una particolare materia di cui è composto, la tecnica con cui è

stato realizzato, la sua forma, il suo colore. Ogni caratteristica tecnica utilizza è definito da un concetto (es.: "terracotta" è una caratteristica tecnica definita dal concetto "materia", relativamente a un bene culturale). Per ogni tipologia di bene culturale possono essere rilevati e registrate caratteristiche tecniche specifiche. `rico:Extent` si riferisce più specificamente alla consistenza di un oggetto, rappresentando una caratteristica fisica o logica quantificabile. In questo senso, `rico:Extent` definisce una caratteristica tecnica più circoscritta rispetto al concetto generale espresso da `arco-dd:TechnicalCharacteristic`. Nel consegue anche l'allineamento gerarchico fra le *object property* `arco-dd:hasTechnicalCharacteristic` e `rico:hasExtent` e le relative inverse `a-dd:isTechnicalCharacteristicOf` e `rico:isExtentOf`. Risulta fattibile anche un allineamento diretto fra le *data property* `rico:physicalOrLogicalExtent` e `arco-archive:extent`, poiché si riferiscono direttamente all'oggetto culturale che descrivono, entrambe in termini quantitativi.

Concetto di rapporto del bene culturale con un'altra opera

In ArCo, la relazione tra un documento e un altro è modellata tramite `arco-cd:RelatedWorkSituation`, che rappresenta il rapporto fra il bene culturale (`cis:CulturalEntity`) in esame e un'altra opera. Tale relazione può descrivere lo stadio di realizzazione del bene culturale in rapporto con l'oggetto che costituisce una fase preparatoria, una fase finale o una derivazione.

In RiC-O, questo concetto viene formalizzato distinguendo esplicitamente le relazioni tra contenuti informativi (`rico:RecordResource`) e istanze fisiche (`rico:Instantiation`):

- `rico:RecordResourceToRecordResourceRelation`: connette due o più `RecordResource`; relazione generica e non orientata;
- `rico:RecordResourceGeneticRelation`: connette due o più `RecordResource` quando esiste un legame genetico, determinato dal processo di sviluppo dei record; relazione generica e non orientata;
- `rico:InstantiationToInstantiationRelation`: connette due o più `Instantiation`; relazione generica e non orientata;
- `rico:DerivationRelation`: connette un'istanza e almeno un'istanza derivata; relazione orientata cronologicamente dalla fonte alla derivata;
- `rico:FunctionalEquivalenceRelation`: collega due o più istanze equivalenti; relazione non orientata;

- `rico:MigrationRelation`: connette un'istanza a un'altra in cui è stata migrata; relazione orientata cronologicamente.

Alla luce di questi elementi e considerando l'allineamento gerarchico già effettuato tra `cis:CulturalEntity`, `rico:Instantiation` e `rico:RecordResource`, è possibile proporre un allineamento gerarchico in cui `arco-cd:RelatedWorkSituation` funge da superclasse di `rico:InstantiationToInstantiationRelation` e `rico:RecordResourceToRecordResourceRelation`.

Concetto di inventario

In ArCo è stato creato un modulo ontologico specifico dedicato alla catalogazione, la Catalogue Ontology (ArCo network)[184], che descrive i concetti relativi al Catalogo Generale Italiano dei Beni Culturali (ICCD-MiBAC). In particolare, le schede di catalogo, ovvero i file XML che registrano tutti i dati raccolti da un catalogatore su un determinato bene culturale, sono rappresentate dalla classe `a-cat:CatalogueRecord`, collegata al bene culturale tramite le *object property* `a-cat:describes` e la relativa inversa `a-cat:isDescribedBy`.

In RiC-O, qualsiasi contenuto informativo è considerato una `rico:RecordResource`, eventualmente tipizzabile tramite la classe `rico:ContentType`. Non esiste quindi una classe dedicata specificamente alla rappresentazione di una descrizione, di un inventario o di un catalogo, come avviene in ArCo. In generale l'*object property* `rico:describesOrDescribed` connette una risorsa `rico:RecordResource` all'entità che la descrive (`rico:Thing`). Di conseguenza, l'allineamento delle *object property* può essere effettuato solo se si stabilisce un allineamento gerarchico tra le classi, considerando `rico:Record` come superclass e `a-cat:CatalogueRecord` come sottoclasse.

Per riassumere informazioni di carattere storico relative ad un'entità, in RiC-O viene utilizzata la *data property* `rico:history` e in ArCo due *data property* più specifiche a seconda dell'oggetto descritto: `arco-cd:historicalBiographicalInformation` e `arco-cd:historicalInformation`. Ne segue un allineamento gerarchico, in cui le due proprietà di ArCo sono subproperties di `rico:history`.

Per quanto riguarda l'origine delle informazioni utilizzate per descrivere o identificare entità o relazioni, in RiC-O viene modellata la fonte relativa a una relazione tramite la *data property* `rico:relationSource`, mentre in ArCo è presente la *data property* più generale `arco-`

[184] https://w3id.org/arco/ontology/catalogue.

core:informationSource, che rappresenta la fonte da cui vengono ricavate le informazioni per documentare qualsiasi entità owl:Thing.

Dal momento che in RiC-O viene considerata solo la fonte di una relazione, è possibile proporre un allineamento gerarchico fra le due proprietà, considerando arco-core:informationSource come superclasse e rico:relationSource come sottoclasse. Questo allineamento mantiene la coerenza concettuale, collegando la tracciabilità delle informazioni alle entità e alle relazioni in entrambe le ontologie.

Concetto di soggetto

Il concetto di soggetto (*subject*) si riferisce all'entità o all'oggetto a cui un contenuto informativo, un documento o un bene culturale è riferito o relativo. Le proprietà in esame sono rispettivamente arco-cd:hasSubject e rico:hasOrHadSubject. In RiC-O, rico:hasOrHadSubject ha come domain rico:RecordResource e come range rico:Thing, mentre in ArCo arco-cd:hasSubject ha come domain arco:Thing e come range arco:Subject.

Poiché ArCo specifica il concetto di rico:Subject, mentre RiC-O ammette qualsiasi rico:Thing come range, è possibile proporre un allineamento gerarchico tra le proprietà, considerando rico:hasOrHadSubject come superclasse e arco-cd:hasSubject come sottoclasse. Lo stesso principio vale per le rispettive proprietà inverse.

Concetto di professione

Le classi rico:OccupationType e arco-cd:Profession, entrambe rappresentanti il concetto di professione, possono essere allineate gerarchicamente. Infatti, rico:OccupationType include sia l'occupazione di singoli individui sia quella di gruppi, mentre arco-cd:Profession si riferisce esclusivamente alla professione di una persona. Di conseguenza, un allineamento gerarchico consente di preservare questa differenza di granularità senza perdere la coerenza concettuale.

Per collegare un'entità alla sua occupazione o professione, ArCo utilizza la proprietà arco-cd:hasProfession, mentre RiC-O impiega rico:hasOrHadOccupationType. In questo caso, l'allineamento è diretto e non gerarchico, a differenza di quanto avviene per le rispettive classi, perché il domain di entrambe le proprietà è limitato a rico:Person. Sebbene rico:OccupationType includa anche l'occupazione di un rico:Group, questa caratteristica non si riflette nelle *object property*, che non

contemplano il domain dei gruppi[185]. Allo stesso modo, si possono allineare anche le proprietà inverse: `arco-cd:isProfessionOf` e `rico:isOrWasOccupationTypeOf`.

Concetto di tipologia

Le classi arco-core:Type e rico:Type sono perfettamente allineabili, così come le *object property* che li relazionano alle entità che tipizzano: rico:isOrWasTypeOf[186] e arco-core:isTypeOf e le relative proprietà inverse rico:hasOrHadType e arco-core:hasType.

In sintesi, a valle dell'analisi, sono stati individuati allineamenti fra 20 classi, 40 *object property* e 7 *data property*. L'allineamento ha permesso di disambiguare e allineare direttamente classi e relazioni rappresentanti concetti fondamentali del dominio archivistico. Inoltre, tramite l'allineamento gerarchico effettuato in alcune casistiche particolari, è stato possibile anche definirle e tipizzarle ulteriormente.

Tuttavia, rimangono privi di allineamento alcuni concetti importanti, per via di una modellazione fin troppo differente fra le due ontologie. Ad esempio, non è stato possibile allineare i concetti di gruppo e organizzazione, per via della disgiunzione presente in ArCo. Non è stato percorribile neanche l'allineamento di una parte delle classi rappresentanti relazioni n-arie, ossia relazioni non binarie generalmente codificate in classi rappresentanti relazioni fra più entità (persone, organizzazioni, date, luoghi, ecc.).

In conclusione, l'allineamento di ArCo a RiC-O rappresenta un passo significativo verso una gestione più efficiente e integrata del patrimonio archivistico italiano. Attraverso l'interoperabilità dei dati, la coerenza semantica, la facilitazione della ricerca e l'ottimizzazione delle operazioni, questa sinergia tra le due ontologie offre una base verso la realizzazione di un dataset culturale veramente interconnesso, riflettendo un impegno verso una *knowledge base* condivisa più ricca e accessibile. Nonostante le limitazioni tecniche e concettuali ancora esistenti, il lavoro di allineamento tra ArCo e RiC-O, sottolinea anche l'importanza di procedere verso un sistema di descrizione del patrimonio culturale che sia pienamente inclusivo delle risorse native digitali, nell'ottica di un patrimonio culturale sempre più orientato verso le nuove tecnologie.

[185] Si tratta, con ogni probabilità, di un errore di modellazione per il quale è stata già fatta segnalazione al team di sviluppo di RiC-O.

[186] Fino alla versione 1.0 di RiC-O, questa *property* e la sua inversa apparivano nella forma `rico:hasOrHadCategory` e `rico:isOrWasCategoryOf`. Dal momento che nella versione 1.1 è stato modificato solo il nome dell'IRI, si è proceduto all'allineamento con la denominazione più accurata.

5.2.2 RiC-O e ArchOnto

L'analisi comparativa fra RiC-O e ArchOnto (modello basato su CIDOC CRM) prende avvio dal caso di studio della descrizione archivistica della serie "Registos de Baptismos" della Parrocchia di Aldoar, fornita dall'Archivio Distrettuale di Porto (Arquivo Distrital do Porto)[187]. I dati dei registri sono stati rappresentati con entrambi i modelli, così da valutare direttamente la loro capacità di descrivere lo stesso insieme informativo. Questa procedura ha permesso di mettere in evidenza punti di forza e limiti di ciascun approccio, facendo emergere analogie e differenze in termini di espressività, struttura e interpretabilità.

"Registos de Baptismos" è una serie appartenente al fondo "Paróquia de Aldoar" (con codice di riferimento PT/ADPRT/PRQ/PPRT01/001), costituita da 17 registri contenenti atti di battesimi effettuati nella parrocchia di Aldoar dal 1° febbraio 1644 al 30 marzo 1911. Nella descrizione archivistica originaria, all'interno della serie ogni libro è stato identificato come unità archivistica, e la descrizione si spinge fino al singolo atto di battesimo, considerato come unità documentaria. La descrizione originale è stata elaborata secondo lo standard ISAD(G) ed è consultabile nella piattaforma DigitArq della Direção-Geral do Livro, dos Arquivos e das Bibliotecas (DGLAB)[188]. Per finalità di ricerca, in vista della necessità di conversione dei dati, l'Archivio ha fornito una copia della descrizione in formato EAD XML.

Il caso di studio si è rivelato particolarmente adatto poiché era già stato rappresentato secondo il modello ArchOnto nell'ambito del progetto EPISA (I. Koch et al. 2020, 2023). L'integrazione con un progetto già avviato ha garantito continuità e messo a disposizione un dataset ben documentato, sul quale è stato possibile confrontare le prestazioni dei due modelli. Questo ha permesso di concentrare l'analisi, da un lato, sull'implementazione in RiC-O e, dall'altro, sugli sviluppi introdotti in ArchOnto rispetto alla sua versione iniziale, insieme alle aree ancora suscettibili di miglioramento. Inoltre, la natura del dataset ha permesso di sperimentare con una ricca rete di persone, eventi e attività coinvolte nella creazione, nella gestione e nella conservazione nei documenti archivistici. Gli atti di battesimo, infatti, contengono informazioni sul documento stesso e sulla serie di appartenenza, nonché dati relativi ai bambini battezzati, ai genitori, ai nonni, alle madrine e padrini, e alle date di nascita e battesimo. Di conseguenza,

[187] Questa ricerca è stata resa possibile grazie al periodo di *visiting* svolto tra marzo e ottobre 2024 presso la Facoltà di Ingegneria dell'Università di Porto, in collaborazione con Inês Koch e le professoresse Carla Teixeira Lopes e Cristina Ribeiro. Gli esiti di questo lavoro sono stati pubblicati nell'articolo Giagnolini, Lucia, Inês Koch, Francesca Tomasi, and Carla Teixeira Lopes. "Comparative Insights into Semantic Archival Modelling: Evaluating RiC-O and ArchOnto Representation Capabilities." *Journal of Documentation* 81, no. 4 (August 14, 2025): 1003-31. https://doi.org/10.1108/JD-12-2024-0310.

[188] https://digitarq.arquivos.pt/documentDetails/72374bcc699b48a99399c4104865f07d.

tali atti forniscono dettagli sull'evento di nascita, sull'attività di battesimo e sugli individui che vi hanno partecipato. Questa complessità costituisce un'opportunità ideale per valutare la capacità di ciascun modello di gestire dati interconnessi e relazioni contestuali e offre spunti sia sull'efficacia nella rappresentazione del singolo documento, sia sulla capacità di descrivere il più ampio contesto archivistico.

Sulla base di queste considerazioni, le attività si sono articolate in tre fasi (sintetizzate nella Figura 5.6):

1) Nella prima fase del lavoro, abbiamo esaminato il dataset per identificare gli scenari più rilevanti dal punto di vista della rappresentazione. Abbiamo individuato quattro esigenze di rappresentazione specifiche: il documento archivistico, l'evento di nascita, l'attività di battesimo e l'attestazione della *provenance* dei dati. Questi quattro aspetti hanno permesso una rappresentazione completa degli individui, delle loro relazioni, degli eventi e delle attività cui hanno preso parte, nonché del contesto archivistico di riferimento. Gli scenari sono stati raffigurati attraverso mappe basate su RiC-O e ArchOnto in modo che ciascuno di essi fosse descritto da entrambi i modelli, possibilmente anche attraverso diverse opzioni di rappresentazione[189]. Questo ha richiesto uno studio attento delle classi e proprietà di entrambe le ontologie: per RiC-O, trattandosi della prima applicazione al dataset; per ArchOnto poiché, anche se il dataset era già stato rappresentato con una prima versione del modello, sono state valutate nuove opzioni rese possibili dalla sua evoluzione.
2) È seguita un'analisi approfondita del file EAD XML originale contenente i registri di battesimo della parrocchia di Aldoar. L'attività ha comportato l'estrazione e l'isolamento di informazioni rilevanti da stringhe descrittive di testo, come il nome del bambino, la data di nascita, i nomi dei genitori, dei nonni, della madrina e del padrino, nonché altri dettagli rilevanti.
3) Da ultimo, per ottenere due dataset comparabili in LOD, il processo ha previsto la migrazione dei dati da EAD XML verso RiC-O attraverso lo strumento RiC-O Converter[190].

Figura 5.6. Fasi del processo di comparazione fra RiC-O e ArchOnto e arricchimento semantico.

[189] È importante notare che i file OWL delle descrizioni effettuate tramite ArchOnto sono disponibili nel *repository* dedicato al progetto EPISA (Pires et al. 2023) ma riflettono dati rappresentati con una versione obsoleta del modello; pertanto, la rappresentazione OWL disponibile non è pienamente conforme agli schemi che sono stati disegnati per massimizzare la portata rappresentativa del modello nella sua versione più recente.
[190] https://archivesnationalesfr.github.io/rico-converter/en/.

In questo capitolo l'attenzione è rivolta alla comparazione tra i due modelli attraverso l'analisi dei quattro scenari di rappresentazione individuati (punto 1): l'evento di nascita, l'attività di battesimo, il documento archivistico e l'attestazione della *provenance* dei dati. Il processo e le metodologie adottate per l'arricchimento semantico e la migrazione dei dati (corrispondenti ai punti 2 e 3) sono descritti nel capitolo 5.3.1.

Evento di nascita

La Figura 5.7 illustra le classi e le proprietà necessarie per la rappresentazione di un'istanza di evento di nascita all'interno di ArchOnto. Per quanto riguarda RiC-O, sono state individuate due principali possibilità di modellazione, presentate rispettivamente nelle Figure 5.8 e 5.9[191].

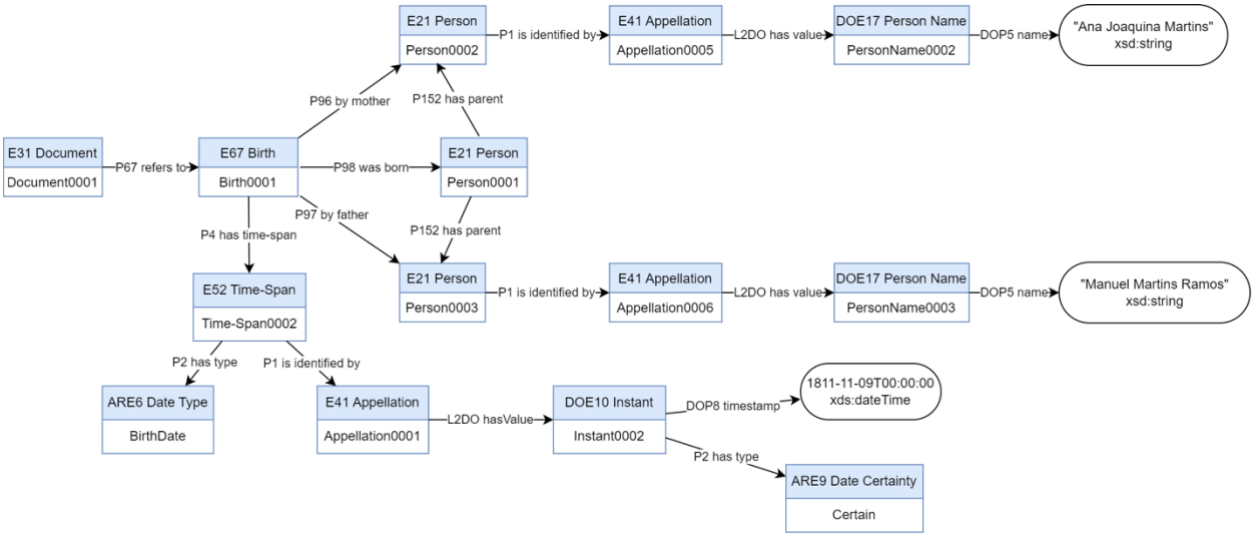


Figura 5.7. Rappresentazione della nascita mediante ArchOnto.

[191] Tutte le rappresentazioni proposte nel presente capitolo sono consultabili anche sotto forma di *snippet* XML/RDF, riportati nelle tabelle in appendice dell'articolo *Comparative insights into semantic archival modelling: evaluating RiC-O and ArchOnto representation capabilities* (Giagnolini, Koch, et al. 2025), dove sono affiancati per facilitarne il confronto.

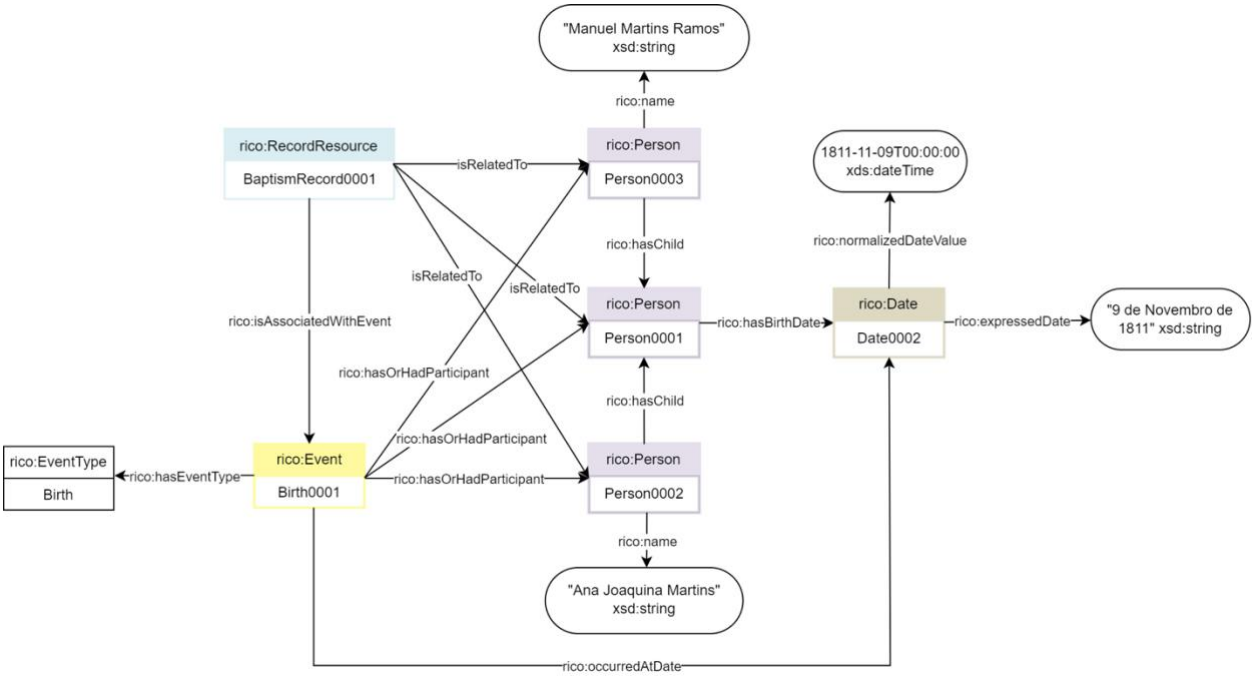


Figura 5.8. Prima opzione per rappresentare un evento di nascita mediante RiC-O.

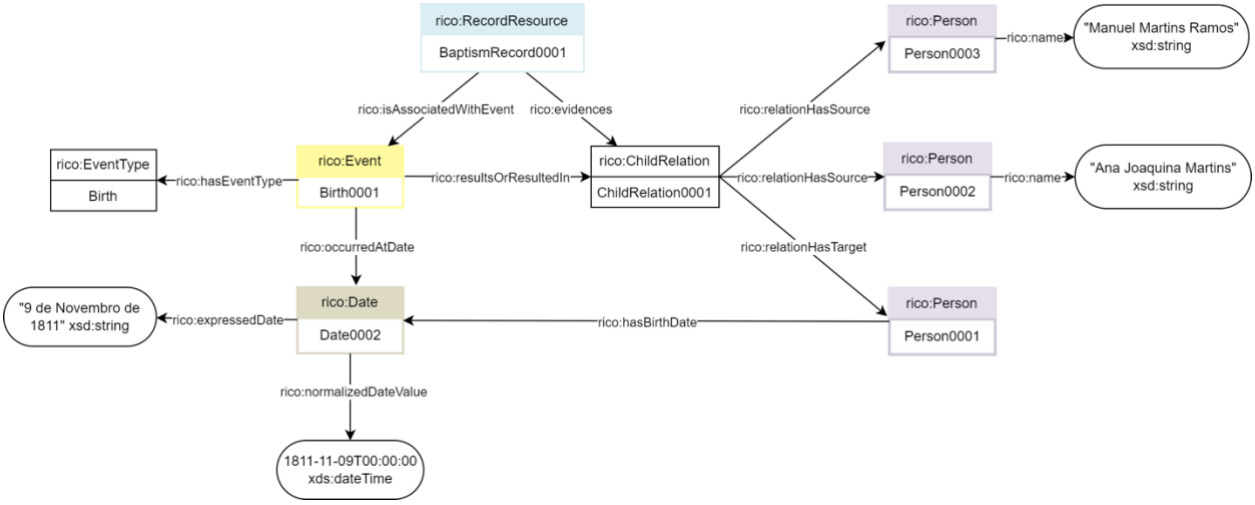


Figura 5.9. Seconda opzione per rappresentare un evento di nascita mediante RiC-O.

Il primo aspetto da analizzare riguarda la rappresentazione delle date. In ArchOnto (e, di conseguenza, in CIDOC CRM[192]), la modalità di rappresentazione della data è unica e risulta poco immediata. Come si può osservare nella Figura 5.7, la data deve essere rappresentata mediante un'istanza della classe `E52 Time-Span`, tipizzata da un'istanza di `ARE6 Date Type` e identificata da un'istanza di `E41 Appellation`,

[192] Nel seguito dell'analisi, le considerazioni formulate per ArchOnto sono da intendersi valide anche per CIDOC CRM; dove necessario, eventuali differenze sono evidenziate.

la quale, a sua volta, deve essere associata a un valore `DOE10 Instant` che consente di determinare la data normalizzata e il relativo grado di certezza.

RiC-O, al contrario, offre una notevole flessibilità nell'espressione delle date associate a una risorsa, spaziando dall'impiego di semplici *data property*, come `rico:Date`, fino alla rappresentazione mediante la classe `rico:Date`. In quest'ultimo caso, come mostrano le Figure 5.8 e 5.9[193], è possibile tipizzare la data attraverso la *object property* che la collega alla risorsa, come `rico:hasBirthDate`. Inoltre, è possibile specificare ulteriori caratteristiche tramite *data property* associate direttamente alla classe `rico:Date`, come `rico:DateQualifier` e `rico:normalizedDateValue`.

Rispetto ad ArchOnto, RiC-O consente una rappresentazione delle date più intuitiva. Il principale meccanismo per definire il tipo di data è rappresentato dall'utilizzo di *object property* esplicite, quali `rico:hasBirthDate`, `rico:hasCreationDate`, `rico:hasDateOfOccurrence` e `rico:hasEndingDate`. Si segnala, inoltre, l'esistenza della classe `rico:DateType` anche in RiC-O; tuttavia, la relativa *scope note* chiarisce che questa classe «should not be confused with the date relations defined in RiC-CM to connect a date entity and any other entity (such as RiC-R069 *is beginning date of*)» (International Council on Archives, Expert Group on Archival Description 2023, 48).

Di conseguenza, sebbene il ventaglio di *object property* sia piuttosto ampio, l'impossibilità di utilizzare `rico:DateType` per tipizzare la data rappresenta una criticità. Definire il tipo di data per fenomeni che esulano parzialmente dai contesti archivistici tradizionali può infatti comportare una rappresentazione non del tutto adeguata della relazione (ad esempio, ricorrendo alla più generica `rico:isDateAssociatedWith`), oppure portare all'adozione di soluzioni non canoniche, come l'impiego della classe generica `rico:Type` o della *data property* `rico:type`.

Un'altra situazione interessante da analizzare riguarda la rappresentazione della relazione tra genitori e figlio. RiC-O consente di descriverla in due modalità: in modo diretto, attraverso la *object property* `rico:hasChild` (e la sua inversa `rico:isChildOf`), che collegano le istanze della classe `rico:Person` rappresentanti i genitori e il figlio; oppure in modo più articolato, attraverso l'aggiunta della classe `rico:ChildRelation`, che associa i genitori come *source* e il figlio come *target* della relazione. Tuttavia, considerando che l'inclusione di `rico:ChildRelation` complica la modellazione senza apportare benefici significativi dal punto di vista semantico, i dati sono stati modellati secondo lo schema definito

[193] Si segnala che, per i diagrammi rappresentativi delle classi e delle proprietà di RiC-O, sono stati utilizzati i colori della palette definita nella documentazione ufficiale di RiC-CM (International Council on Archives, Expert Group on Archival Description 2023).

nella Figura 5.8. Va inoltre osservato che RiC-O non prevede un meccanismo diretto per esplicitare il ruolo di "madre" o "padre", che vengono invece rappresentati genericamente come "genitori". Qualora fosse necessario, è comunque possibile specificare ulteriormente tali ruoli identificando il `rico:DemographicGroup` di appartenenza o esplicitando il genere o il sesso biologico di ciascun soggetto. La semantica della relazione, espressa con entrambi gli approcci, risulta sufficiente per attestare la nascita di un figlio e collegarla alla risorsa documentaria che la menziona. È tuttavia possibile raffinare ulteriormente questa rappresentazione identificando la nascita come un vero e proprio evento, tramite la classe `rico:Event`, collegata direttamente alle istanze delle persone (Figura 5.8) oppure alla relazione `rico:ChildRelation` (Figura 5.9). In ogni caso, il figlio viene direttamente associato a una data di nascita tramite la *object property* `rico:hasBirthDate`; la stessa data rappresenta naturalmente il momento in cui si è verificato l'evento di nascita.

In ArchOnto, la rappresentazione dell'evento di nascita appare come una crasi delle modalità previste da RiC-O (Figura 5.7). La classe `E67 Birth` è specificamente impiegata per attestare l'evento e viene collegata alle istanze delle persone mediante *object property* esplicite: `P96 by mother` e `P97 by father` per i genitori, e `P89 was born` per il figlio. Tra i genitori e il figlio può inoltre essere utilizzata la proprietà `P152 has parent`. La principale differenza consiste nel fatto che, in ArchOnto, nessuna data può essere associata direttamente all'istanza della persona, come avviene in RiC-O, ma deve necessariamente passare attraverso un nodo intermedio. In questo caso specifico, ciò rende imprescindibile la presenza dell'istanza `E67 Birth` per descrivere in modo completo la nascita di un individuo.

Con queste premesse, è possibile ipotizzare anche un'analisi finalizzata all'identificazione dei fratelli e sorelle del soggetto (sebbene non sia stata rappresentata nei dati, né nei diagrammi). In RiC-O, ciò può avvenire direttamente tramite la proprietà oggetto `rico:hasSibling`, oppure tramite la classe `rico:SiblingRelation`. In ArchOnto, invece, non esiste una modalità diretta per definire questa relazione: l'unica possibilità consiste nell'identificare tutte le istanze di `E67 Birth` collegate a entrambi i genitori.

Attività di battesimo

L'attività di battesimo costituisce l'elemento centrale dei record archivistici analizzati. Per rappresentare adeguatamente tale attività, è necessario considerare le persone che vi hanno preso parte, i rispettivi ruoli ricoperti e la data in cui ha avuto luogo. La Figura 5.10 illustra le classi e le proprietà necessarie per la rappresentazione di un'istanza di battesimo in ArchOnto. Per quanto riguarda RiC-O, sono state selezionate due principali possibilità di modellazione, presentate nelle Figure 5.11 e 5.12.

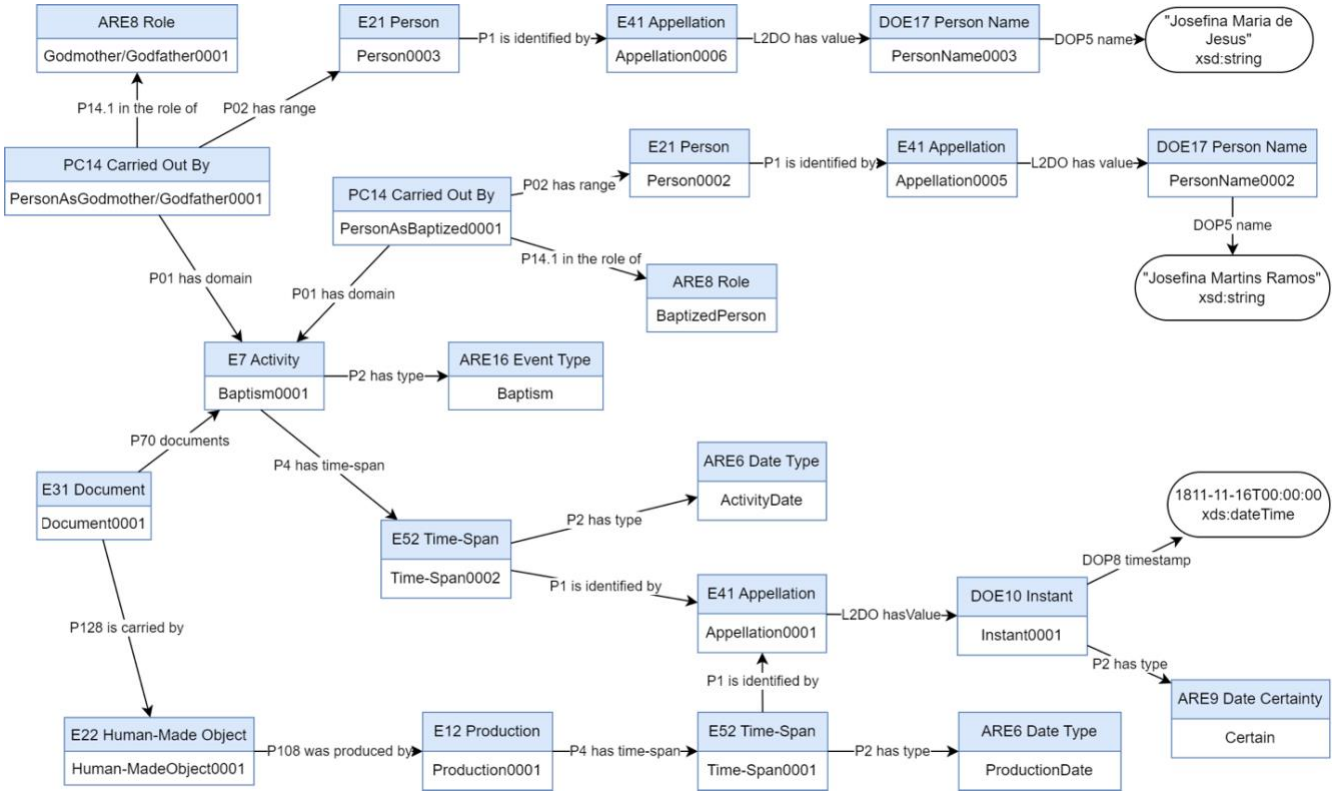


Figura 5.10. Rappresentazione dell'attività di battesimo mediante ArchOnto.

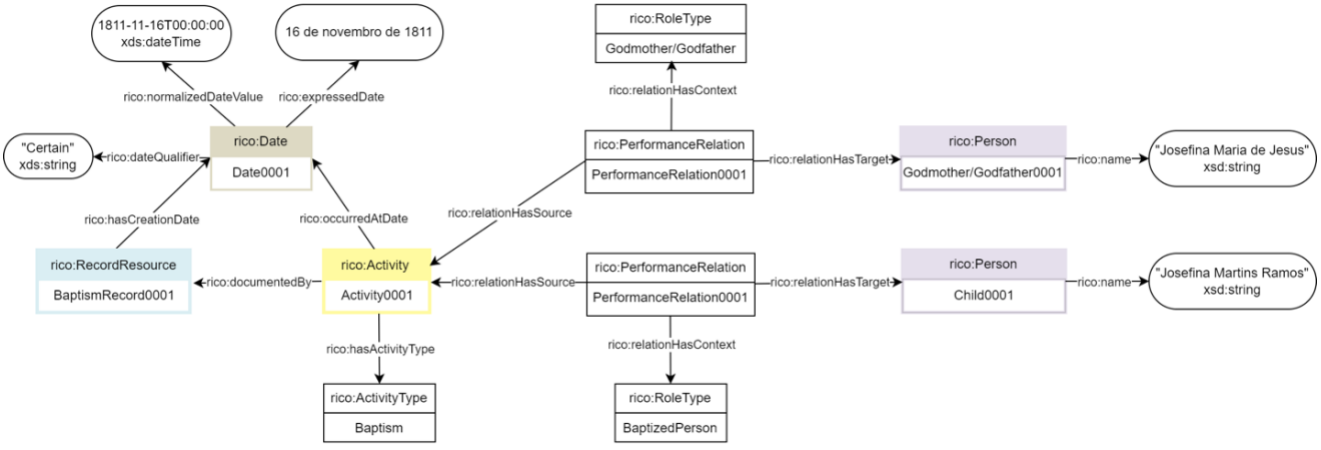


Figura 5.11. Prima possibilità di rappresentazione dell'attività di battesimo mediante RiC-O.

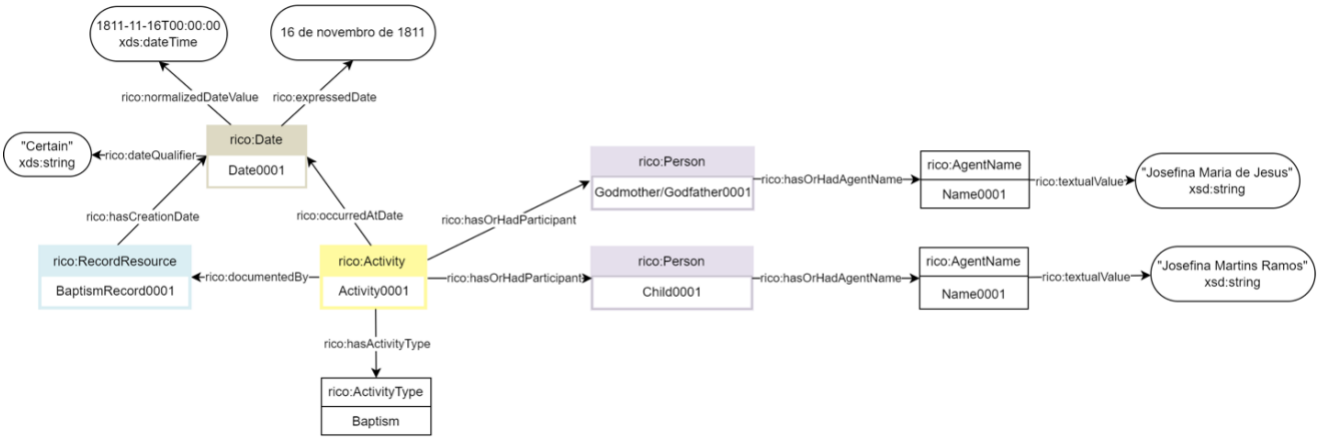


Figura 5.12. Seconda possibilità di rappresentazione dell'attività di battesimo mediante RiC-O.

In ArchOnto, come illustrato nella Figura 5.10, l'attività di battesimo è rappresentata attraverso la classe `E7 Activity`, il cui tipo è specificato tramite `ARE16 Event Type`. Tale attività coinvolge diversi attori definiti mediante la classe `E21 Person`. Per rappresentare questi attori, è necessario considerare il ruolo da essi ricoperto nell'ambito dell'attività in questione, attraverso una relazione n-aria. Questa rappresentazione consente di indicare che una persona (`E21 Person`) ha ricoperto un determinato ruolo (`ARE8 Role`) nell'attività di battesimo (`E7 Activity`), ad esempio il ruolo di padrino o madrina. Qualora il ruolo non sia noto, è comunque possibile affermare in modo generico che la persona (`E21 Person`) ha partecipato all'attività (`E7 Activity`) utilizzando la *object property* `P11 had participant`.

RiC-O si dimostra piuttosto flessibile nella descrizione dell'attività di battesimo. La rappresentazione illustrata nella Figura 5.11 include l'impiego della classe `rico:PerformanceRelation`, che consente di collegare la classe `rico:Person`, ovvero il padrino, al ruolo da questi ricoperto nell'attività, identificata come il contesto della relazione. Tale rappresentazione non risulta immediatamente intuitiva, in quanto RiC-O non prevede una *object property* che espliciti il concetto di "ha ruolo".

Un approccio più diretto, come mostrato nella Figura 5.12, consiste nello stabilire una relazione immediata tra l'attività e i suoi partecipanti. Analogamente a quanto previsto da CIDOC CRM, l'utilizzo della proprietà `rico:hasOrHadParticipant` appare più adeguato in situazioni in cui il ruolo non sia noto oppure, se lo è, laddove si preferisca una modellazione meno complessa. Tuttavia, questo implica che sia necessario fare riferimento ad altre informazioni contestuali per comprendere il ruolo dei soggetti coinvolti nell'attività (ad esempio al campo descrittivo `rico:scopeAndContent`). Per questa ragione, si è scelto di rappresentare i dati seguendo la mappa concettuale illustrata nella Figura 5.11.

Si potrebbe considerare anche l'utilizzo della classe `rico:Position`, ma, secondo quanto indicato nella relativa *scope note*, sembra riferirsi più esplicitamente al contesto lavorativo: «Position is commonly defined in a Mandate, often called a position description or job description. The Mandate may specify the work to be performed (Activity) as well as the competencies for performing the Activity» (International Council on Archives, Expert Group on Archival Description 2023, 30).

Un ulteriore aspetto da considerare è il modo in cui le due ontologie assegnano il nome alla persona. In ArchOnto, la connessione tra la classe `E21 Person` e la stringa che ne definisce il nome in formato *literal* è necessariamente mediata da due ulteriori classi: `E41 Appellation` e `DOE17 Person Name`, quest'ultima creata per contemplare diverse possibili forme del nome. Grazie a tale categorizzazione, la rappresentazione acquisisce un livello maggiore di granularità, evitando di collocare tutti gli elementi testuali all'interno della medesima classe. Di conseguenza, anche l'elaborazione delle interrogazioni all'interno della base di dati risulterà più precisa e mirata. Come si evince dalla Figura 5.12, RiC-O può rappresentare la relazione in modo analogo mediante l'utilizzo della classe `rico:AgentName`, riducendo di un passaggio il processo rispetto a quanto previsto in ArchOnto. In alternativa, è possibile collegare direttamente la stringa del nome alla persona mediante la *data property* `rico:name`. Poiché nel contesto del caso di studio non è necessario fornire ulteriori dettagli sui nomi degli agenti, si è scelto di rappresentare i dati in RiC-O adottando questa seconda opzione.

Infine, come emerge da entrambe le rappresentazioni, va sottolineato che la data di registrazione dell'atto di battesimo coincide con la data del battesimo stesso, poiché la pratica di registrazione avveniva generalmente in concomitanza al battesimo.

Descrizione della risorsa archivistica

La Figura 5.13 illustra le classi e le proprietà necessarie per la rappresentazione di un'istanza di documento archivistico in ArchOnto. Per quanto riguarda RiC-O, il risultato è presentato nella Figura 5.14.

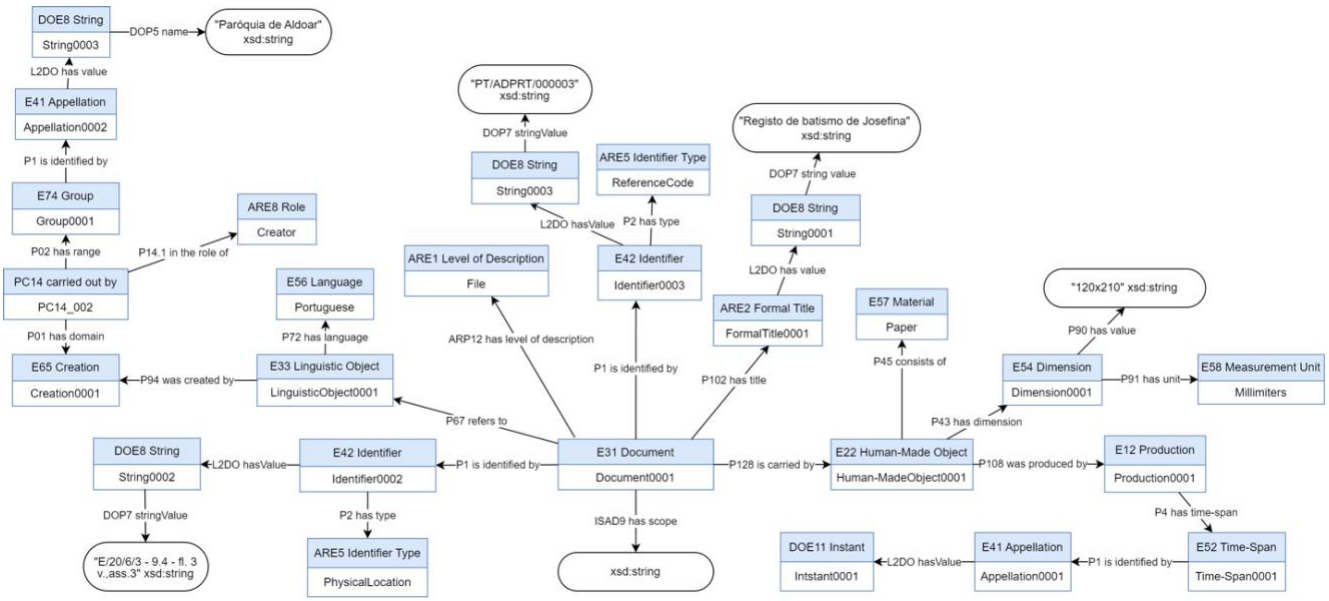


Figura 5.13. Rappresentazione della risorsa archivistica e dei suoi contesti mediante ArchOnto.

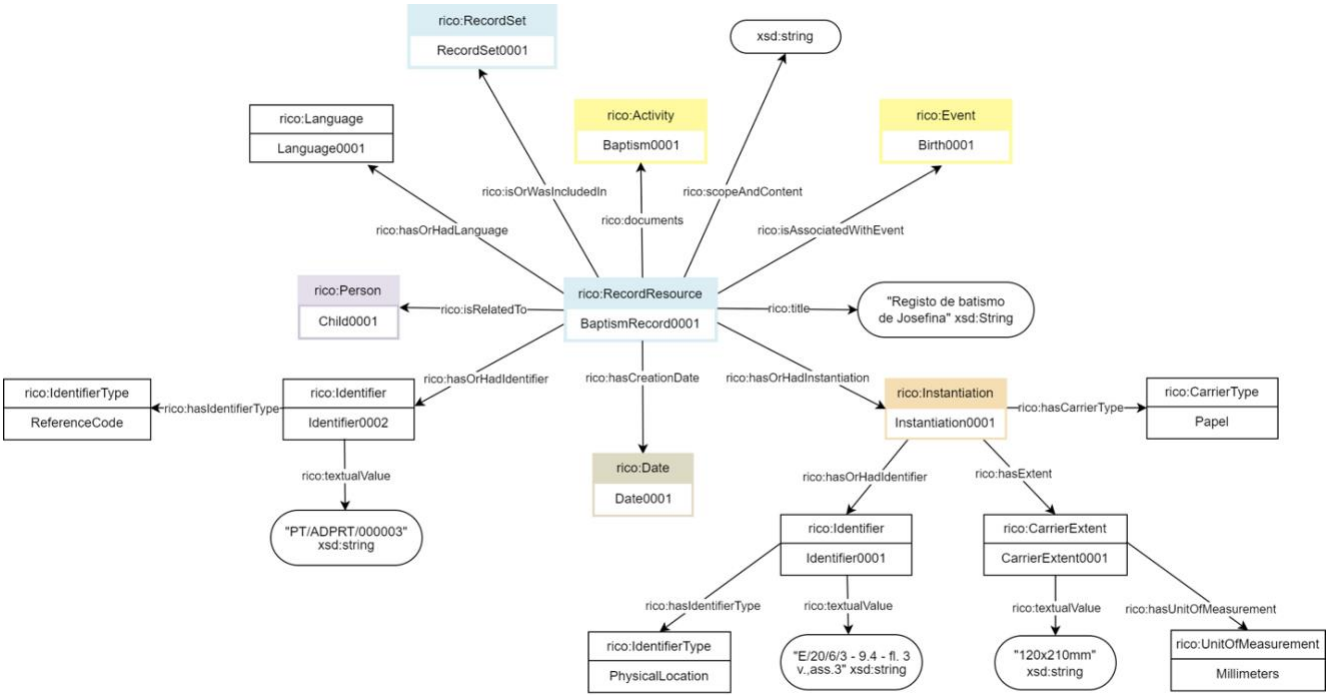


Figura 5.14. Possibile rappresentazione della risorsa archivistica e dei suoi contesti mediante RiC-O.

La rappresentazione della risorsa archivistica in ArchOnto è sviluppata attraverso l'impiego di tre classi: `E31 Document`, `E33 Linguistic Object` ed `E22 Human-Made Object`, come illustrato nella Figura 5.13. La prima, `E31 Document`, costituisce il nodo centrale della rappresentazione e connette il contenuto linguistico-testuale del documento (`E33 Linguistic Object`) alla sua realizzazione in un oggetto fisico (`E22 Human-Made Object`). Alla classe `E31 Document` sono collegate le principali coordinate

identificative del documento archivistico, ossia gli identificativi, il livello di descrizione e il titolo formale.

La rappresentazione della risorsa archivistica in RiC-O, come mostrato nella Figura 5.14, si articola invece su due livelli: da un lato il contenuto informativo, rappresentato da `rico:RecordResource`, e dall'altro la sua o le sue materializzazioni fisiche, rappresentate da `rico:Instantiation`. Di conseguenza, tutte le caratteristiche del documento relative al suo contenuto informativo (come la lingua o l'attività documentata) sono collegate alla `rico:RecordResource`, mentre tutte le caratteristiche strettamente fisiche (come la localizzazione materiale o l'estensione del supporto) sono collegate alla `rico:Instantiation`.

La rappresentazione dei registri di battesimo esemplifica una delle principali motivazioni che hanno portato all'introduzione della classe `rico:RecordResource` come superclasse di `rico:Record` e `rico:RecordPart`, ossia la difficoltà potenziale nel categorizzare con precisione un oggetto archivistico come record o come parte di record (Clavaud e Wildi 2021, 8). Infatti, mentre i registri di battesimo possiedono un contenuto informativo autonomo e indipendente che ne giustificherebbe la rappresentazione come `rico:Record`, essi fanno anche parte, a livello fisico, di un volume che li raccoglie in maniera continua, rendendoli potenzialmente rappresentabili come `rico:RecordPart`. Per evitare di operare scelte nette e discutibili da un lato o dall'altro, si è pertanto optato per rappresentare i singoli registri di battesimo come `rico:RecordResource`.

Il registro di battesimo è inoltre collegato a istanze di classi derivate dall'estrazione di informazioni dall'elemento *scope and content* secondo ISAD(G), come il neonato o l'evento di nascita, nelle modalità esplicitate al capitolo 5.2.1. Per garantire chiarezza e accessibilità, si è preferito mantenere anche l'integrità dell'elemento *scope and content*, conservandolo come stringa all'interno della *data property* `rico:scopeAndContent`.

Il primo elemento da considerare nel confronto tra le due rappresentazioni riguarda l'uso delle classi per identificare il documento archivistico (sia come oggetto che come informazione). Con l'impiego di tre classi (`E22 Human-Made Object`, `E33 Linguistic Object` e `E31 Document`) ArchOnto propone una struttura più articolata. La differenza semantica tra le ultime due classi è sottile: mentre entrambe si riferiscono al contenuto informativo, `E31 Document` è un oggetto informativo astratto che può esistere in diversi formati (testuali, visivi, audiovisivi), mentre `E33 Linguistic Object` si riferisce specificamente a contenuti espressi mediante linguaggio. Tale categorizzazione dettagliata consente di

cogliere con precisione la natura e le relazioni tra i documenti, ma comporta anche un certo livello di complessità che richiede uno sforzo notevole in termini di comprensione e implementazione.

La più semplice bipartizione tra `rico:RecordResource` e `rico:Instantiation` proposta da RiC-O può agevolare una comprensione e una gestione più intuitive dei documenti, semplicemente distinguendo il contenuto intellettuale dal supporto fisico. Poiché questa modellazione rappresenta una novità rispetto agli standard precedenti (Clavaud e Wildi 2021, 8), e richiede ancora un processo di familiarizzazione da parte di archivisti e ricercatori, la sua semplicità può rivelarsi un fattore chiave: un modello più diretto consente infatti una comprensione più immediata e un'applicazione più agevole ed efficace.

CIDOC CRM, in linea teorica, permetterebbe un collegamento diretto tra `E33 Linguistic Object` e `E22 Human-Made Object`, rendendo la rappresentazione maggiormente comparabile e allineata a quella di RiC-O. In tale scenario, le informazioni associate alla classe `E31 Document` devono però essere distribuite tra queste due classi. Ciononostante, pur essendo tecnicamente possibile, tale approccio escluderebbe i documenti archivistici che non possono essere qualificati come `E33 Linguistic Object` (come fotografie, disegni, video o registrazioni audio). D'altro canto, omettere la classe `E33 Linguistic Object` mantenendo soltanto `E31 Document` impedirebbe di indicare la lingua del documento, che in CIDOC CRM può essere associata solo alla classe `E33 Linguistic Object`.

Un altro aspetto interessante riguarda la gestione dell'elemento ISAD(G) *level of description*. Si tratta infatti di una delle estensioni apportate dal modello ArchOnto, non previso in CIDOC CRM in quanto concepito primariamente per la descrizione di oggetti (I. Koch et al. 2020) ArchOnto risponde a tale esigenza introducendo la classe `ARE1 Level of Description`, collegata al documento tramite la proprietà oggetto `ARP12 has level of description`.
RiC-O esprime il medesimo concetto in modo diverso, identificando innanzitutto una tripartizione della classe `rico:RecordResource` nelle tre sottoclassi `rico:Record`, `rico:RecordPart` e `rico:RecordSet`. Facendo riferimento ai livelli di descrizione proposti da ISAD(G), il record è generalmente inteso come "unità documentaria". Tutti i livelli che, virtualmente o di fatto, raggruppano più record sono considerati record set. Ogni record set può essere ulteriormente tipizzato attraverso l'identificazione di un `rico:RecordSetType`, attraverso cui può essere associato ad un vocabolario controllato di tipologie (ad es. fondo, subfondo, serie, ecc.).

Approfondendo ulteriormente, si osserva che ArchOnto introduce un elemento non previsto né da CIDOC CRM né dalla prima versione di RiC-O: la distinzione tra titolo originale e titolo attribuito. ArchOnto estende infatti la classe `E35 Title` di CIDOC CRM creando due sottoclassi specifiche, `ARE2`

`Formal Title` e `ARE3 Supplied Title`. In RiC-O 1.0, invece, non esisteva una classe analoga e l'unica modalità per caratterizzare i titoli era l'associazione della classe `rico:Title` con la *data property* `rico:type` o con un'istanza della classe `rico:Type`. A fronte di questa lacuna, la comunità ha richiesto l'introduzione di un meccanismo più esplicito per la tipizzazione dei titoli, che è stato quindi inserito nella versione 1.1 di RiC-O sotto forma della nuova classe `rico:TitleType`. Questo aspetto può sembrare marginale, ma non lo è affatto se consideriamo che molte risorse archivistiche non dispongono di una denominazione originale e vengono denominate dall'archivista. In un contesto in cui la rappresentazione del contenuto informativo del documento è distinta da quella della sua forma materiale, diviene sempre più importante indicare l'origine del titolo, che potrebbe anche differire tra le due classi.

Provenance

Nel contesto dei LOD, l'attestazione della provenienza dei dati rappresenta un aspetto cruciale per garantirne l'affidabilità, la tracciabilità e la responsabilità (Tomasi 2017). Fondamentalmente, un'ontologia dovrebbe sempre essere sufficientemente espressiva da consentire la registrazione di informazioni essenziali sulla provenienza anche attraverso una semplice tripla RDF, come l'identificazione dell'autore di una risorsa. Tuttavia, nella maggior parte dei casi, tale rappresentazione si rivela insufficiente. Sebbene le triple RDF tradizionali siano adeguate alla codifica di fatti basilari, risultano carenti nel rappresentare la complessità del contesto informativo che accompagna i dati, necessario per una piena attestazione della loro *provenance* (Sikos e Philp 2020). Una corretta rappresentazione della *provenance* implica infatti non solo la modellazione dei dati in sé, ma anche delle circostanze della loro generazione e modifica, includendo agenti coinvolti, tempi, luoghi, metodi di raccolta e le fonti o evidenze che ne supportano la validità.

Nel caso di studio qui considerato, l'estrazione di entità dai campi testuali, l'attribuzione di relazioni tra di esse e la successiva conversione del formato rappresentano un vero e proprio atto interpretativo del contenuto originario (Giagnolini et al. 2024). Di conseguenza, il nuovo inventario espresso in LOD deve essere esplicitamente identificato come il risultato di un'interpretazione successiva, distinta dall'attività descrittiva che ha generato l'inventario originario in formato EAD. I file risultanti dalla conversione dovrebbero pertanto essere accompagnati da una serie di triple aggiuntive che attestino esplicitamente la loro provenienza, le modalità di produzione e la responsabilità autoriale (Tomasi 2023). Per affrontare questo problema, le strategie più comuni includono l'uso della reificazione RDF, delle relazioni n-arie, di RDF-star o dei *named graphs* (Sikos e Philp 2020; Rupp et al. 2022).

ArchOnto, pur non avendo ancora adottato un approccio univoco per la rappresentazione della provenienza, si fonda su CIDOC CRM, il che ha consentito di esplorare modelli compatibili già esistenti. Tra questi, CRMdig[194] si è rivelato una soluzione promettente: si tratta di un'estensione di CIDOC CRM (attualmente alla versione 3.2.1) concepita per modellare i metadati relativi alle fasi e ai metodi di produzione dei prodotti digitali derivati da processi di digitalizzazione e modellazione, inclusi oggetti 2D e 3D (FORTH e CRM SIG 2016). In tal senso, è stato avviato uno studio per valutare la fattibilità dell'integrazione di CRMdig all'interno di ArchOnto.

D'altra parte, anche EGAD, nel quadro di sviluppo del modello *Records in Contexts* (RiC), riconosce formalmente l'importanza della documentazione della provenienza. Come evidenziato nella documentazione ufficiale di RiC-CM, la descrizione archivistica dovrebbe articolarsi secondo almeno quattro livelli contestuali: l'identificazione dell'ente conservatore, l'individuazione della posizione responsabile dell'elaborazione e descrizione dei documenti, la documentazione del record descrittivo stesso, e l'attestazione delle fonti probatorie su cui si fondano le affermazioni contenute nella descrizione (International Council on Archives, Expert Group on Archival Description 2023). In tal senso, RiC-O consente di rendere espliciti questi livelli mediante un uso sistematico di relazioni n-arie. Questo approccio deriva dalla scelta di reificare le relazioni attraverso la classe `rico:Relation` e le sue numerose sottoclassi, come `rico:AuthorshipRelation`, `rico:AccumulationRelation` o `rico:MigrationRelation`, progettate per rappresentare in modo granulare le molteplici relazioni emergenti nei contesti documentari e curatoriali.

Tuttavia, la necessità di sviluppare soluzioni capaci di rappresentare la provenienza anche di affermazioni complesse è già emersa all'interno della comunità di sviluppo RiC, tanto da essere stata inserita tra gli obiettivi prioritari delle prossime evoluzioni di RiC-O (ICA-EGAD (International Council on Archives - Expert Group on Archival Description) 2024). In questo quadro, le Figure 5.15 e 5.16 illustrano due possibili modalità di rappresentazione della *provenance* rispettivamente tramite ArchOnto e RiC-O.

[194] https://cidoc-crm.org/crmdig.

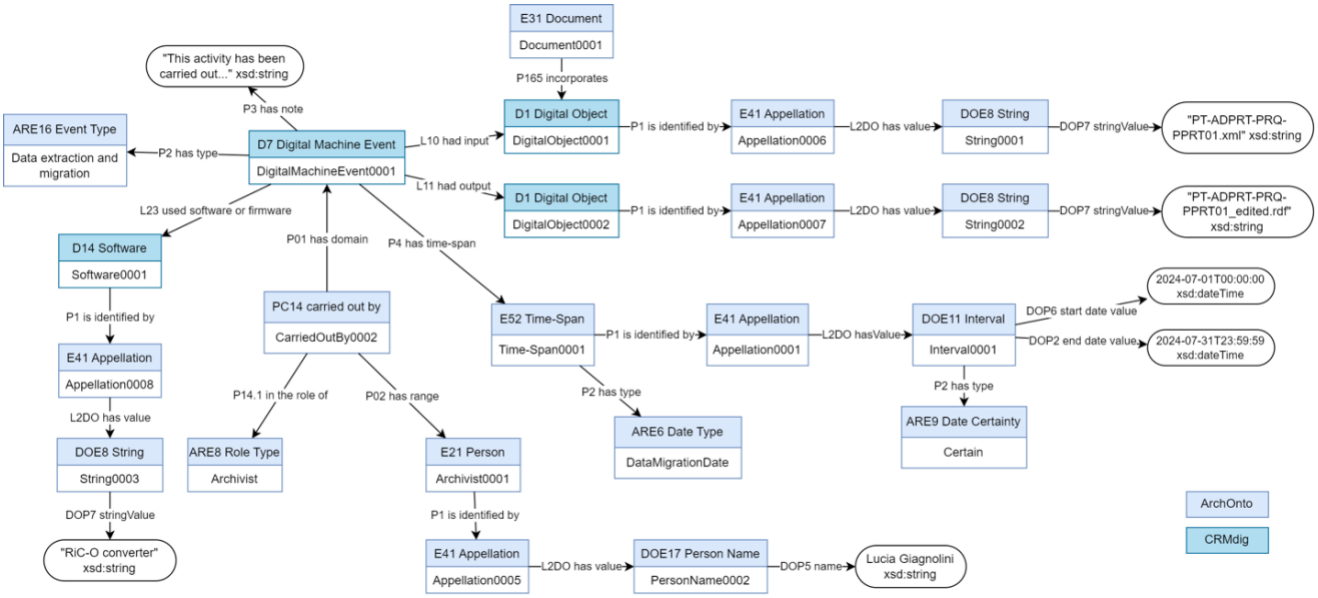


Figura 5.15. Rappresentazione dell'estrazione e della conversione dei dati in ArchOnto e CRMdig.

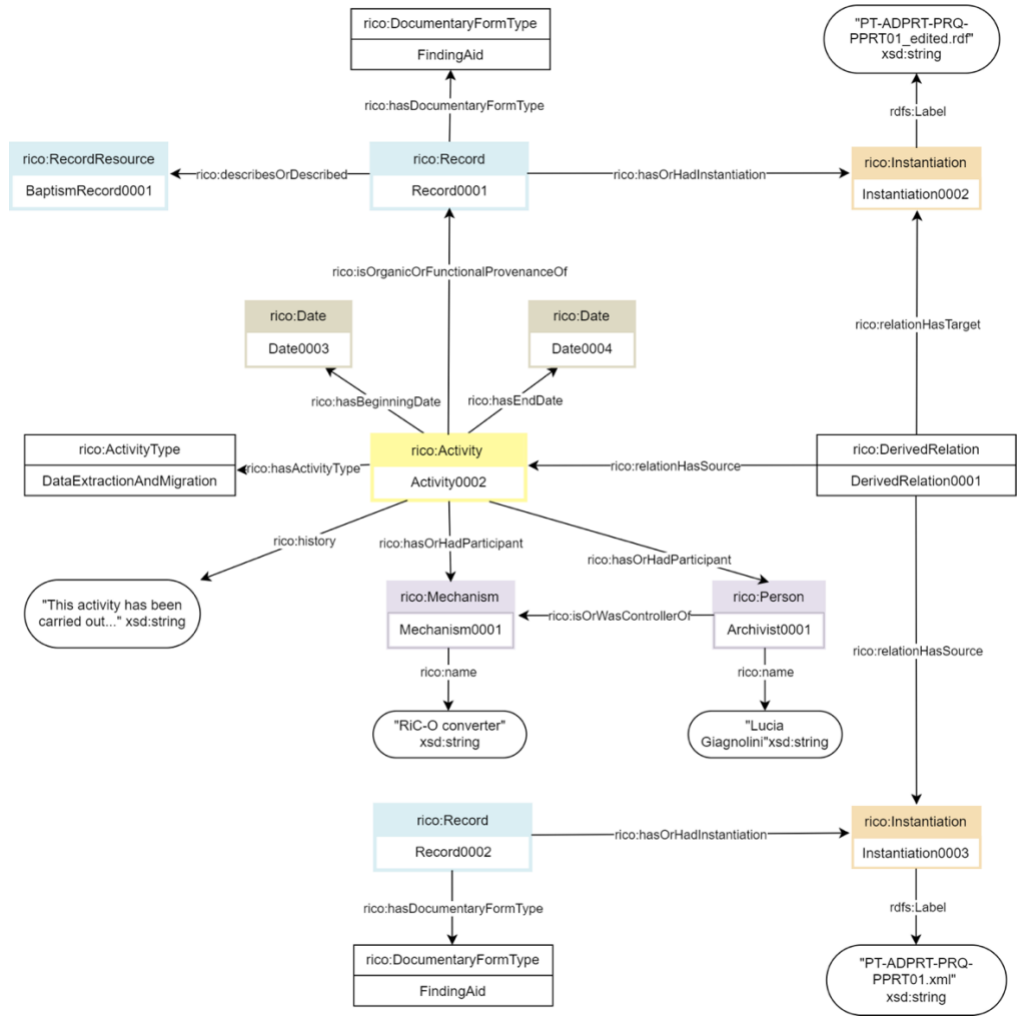


Figura 5.16. Possibile rappresentazione dell'estrazione e della conversione dei dati in RiC-O.

La rappresentazione della *provenance* in ArchOnto, attraverso l'integrazione di CRMdig, si concentra principalmente sull'attestazione dell'evento di estrazione e migrazione dei dati da un inventario in formato EAD verso una rappresentazione RDF[195]. Il punto di partenza è stato l'utilizzo della classe `D7 Digital Machine Event`, alla quale è stato associato il tipo di evento mediante la classe `ARE16 Event Type` di ArchOnto. Secondo la definizione di CRMdig, la classe `D7 Digital Machine Event` rappresenta eventi che avvengono su dispositivi fisico-digitali, avviati intenzionalmente da un agente umano, i cui risultati includono la creazione di una nuova istanza della classe `D1 Digital Object` (FORTH e CRM SIG 2016). A partire da questa modellazione è stato possibile rappresentare, mediante una relazione n-aria, la responsabilità dell'archivista nell'attività di conversione svolta attraverso l'uso di RiC-O Converter. Per modellare questa relazione n-aria, è stata individuata una classe intermedia (`PC14 Carried Out By`) in grado di collegare tutti gli elementi coinvolti. In questo modo si è potuto esprimere che una persona (`E21 Person`) ha ricoperto il ruolo di archivista (`ARE8 Role`) nell'ambito dell'evento (`D7 Digital Machine Event`). L'utilizzo del RiC-O Converter è stato invece rappresentato attraverso la proprietà `L23 used software or firmware`, che collega l'evento digitale al software utilizzato (`D14 Software`). La trasformazione dell'inventario archivistico da EAD a RDF è stata esplicitata attraverso le proprietà `L10 had input` e `L11 had output`, che mettono in relazione l'evento con le istanze di `D1 Digital Object` corrispondenti, rispettivamente, al file EAD originale e al documento RDF generato. Per completare la descrizione dell'evento, è stata utilizzata la proprietà `P3 has note` di ArchOnto.

L'integrazione di CRMdig in ArchOnto presenta diversi vantaggi, in particolare nella capacità di rappresentare con coerenza i processi di conversione tra formati digitali, riflettendo accuratamente l'ambiente in cui tali attività hanno luogo. Un elemento di forza è la possibilità di associare l'evento digitale all'agente umano a cui esso è attribuibile, rappresentando l'intenzionalità dell'azione e l'interazione con il software. Tuttavia, è importante notare che CRMdig, non essendo ancora completamente stabilizzato, potrebbe presentare problematiche in termini di implementazione e interoperabilità. Nonostante ciò, la sua integrazione in ArchOnto rappresenta un passaggio importante nello sviluppo del modello.

Per quanto riguarda la rappresentazione con RiC-O, in linea con le specifiche di RiC-CM, gli inventari possono essere modellati come istanze della classe `rico:Record` (ICA-EGAD (International Council on

[195] Si noti che la rappresentazione in ArchOnto con CRMdig riguarda esclusivamente le operazioni di estrazione e conversione realizzate nel contesto della presente ricerca, escludendo quelle precedentemente effettuate per il progetto EPISA. Questa scelta è stata adottata deliberatamente, per evitare possibili inesattezze nella ricostruzione di processi condotti da altri autori.

Archives - Expert Group on Archival Description) 2024, 94). In questo caso, sia l'inventario in EAD che la sua versione RDF sono rappresentati come `rico:Record` aventi come `rico:DocumentaryFormType` il valore "FindingAid". Come visibile dalla Figura 5.16, il collegamento tra le due versioni è espresso tramite la classe `rico:DerivationRelation`, che connette un record a quello dal quale deriva. Il contesto di questa relazione è costituito dall'attività di estrazione e migrazione dei dati, considerata anche come la "provenienza organica" dell'inventario in RDF. Tale attività, caratterizzata da una data di inizio e una di fine, coinvolge due partecipanti: l'archivista (in qualità di supervisore) e il software RiC-O Converter. Una descrizione testuale dell'attività è fornita tramite la proprietà `rico:history`.

Sebbene questa modellazione soddisfi l'obiettivo di attestare la *provenance*, spiccano alcune criticità legate alla complessità e a possibili ridondanze. Secondo le specifiche di RiC-O, la `rico:DerivationRelation` può collegare solo `rico:Instantiation`, il ché complica la rappresentazione quando si intende esprimere una derivazione tra soli contenuti informativi. Estendere l'ambito di applicazione di tale classe per includere anche i record potrebbe rendere il modello più flessibile e facilmente applicabile nella pratica.

Inoltre, la relazione tra archivista e software è modellata tramite la proprietà `rico:isOrWasControllerOf`, che tuttavia non riflette adeguatamente la natura strumentale del rapporto. Sarebbe più opportuno introdurre una proprietà che espliciti l'uso del software da parte dell'archivista, analogamente a quanto previsto in CRMdig, per restituire in modo più chiaro l'interazione uomo-macchina nel contesto delle attività archivistiche.

Un'ulteriore osservazione riguarda la definizione dell'attività di estrazione e conversione come contesto della relazione di derivazione tra inventari. La proprietà `rico:HasOrganicOrFunctionalProvenance` descrive correttamente l'attività come fonte della creazione del nuovo inventario, mentre la proprietà `rico:relationHasContext` risulta semanticamente debole, poiché l'attività in questione non è semplicemente un contesto, ma il fattore generativo della relazione. L'utilizzo di `rico:relationHasSource`, in questo caso, risulterebbe più appropriato.

Le due modellazioni proposte costituiscono un primo tentativo di affrontare la complessa sfida della tracciabilità nella conversione dei formati, ma necessitano di ulteriori sviluppi. Considerando che la migrazione da inventari XML verso RDF potrebbe presentarsi frequentemente negli istituti archivistici, se le tendenze attuali e le indicazioni dell'ICA continueranno in questa direzione, emerge la necessità di modelli più solidi e precisi.

Un aspetto positivo è rappresentato dalla consapevolezza, all'interno delle comunità legate a CIDOC CRM e RiC-O, dell'urgenza di rappresentare situazioni complesse e, a tal fine, entrambe le comunità hanno avviato gruppi di lavoro specifici per il potenziamento dei modelli. Oltre a considerare strategie avanzate come l'uso di *named graphs* o *rdf-star*, una direzione promettente potrebbe essere l'integrazione o l'allineamento con ontologie specifiche per la rappresentazione della *provenance*, come PROV-O (Belhajjame et al. 2013) o HiCO (Tomasi 2017; Daquino e Tomasi 2015), già largamente applicate in diversi domini. Ciò permetterebbe una terminologia più precisa e faciliterebbe l'interoperabilità, contribuendo allo sviluppo di *framework* più esaustivi e adeguati alle esigenze evolutive delle pratiche archivistiche digitali.

L'analisi finora condotta ha permesso di individuare le modalità attraverso cui due modelli, ArchOnto e RiC-O, rappresentano i record archivistici, le relazioni e i contesti, evidenziandone le caratteristiche in termini di flessibilità, complessità, grado di astrazione, capacità di esprimere la provenienza, nonché il supporto garantito dalle rispettive comunità di utenti e sviluppatori. Ciascuno degli scenari esaminati ha messo in luce differenze e analogie significative, dalle quali sono emerse riflessioni più generali sintetizzate nella Tabella 5.7.

ArchOnto	**RiC-O**
Rappresentazione meno flessibile	Rappresentazione più flessibile
Rappresentazione più macchinosa	Rappresentazione più lineare
Maggiore integrazione con altri domini GLAM	Minore integrazione con altri domini GLAM
La base (CIDOC CRM) è sviluppata specificamente per il dominio dei musei	Sviluppato specificamente per il dominio archivistico
La base (CIDOC CRM) è un modello consolidato	Modello e ontologia recenti
Incompletezza nell'espressione della provenienza	Incompletezza nell'espressione della provenienza
Comunità di utenti e sviluppatori attive	Comunità di utenti e sviluppatori attive

Tabella 5.7. Comparazione delle caratteristiche generali di ArchOnto e RiC-O.

Nel complesso, i risultati di questo studio si allineano a quanto osservato da Oliveira et al. (2024), i quali hanno confrontato CIDOC CRM e RiC-O partendo dalle definizioni di classi e proprietà. In particolare, ArchOnto beneficia dell'integrazione con un modello ampiamente consolidato e utilizzato nel settore GLAM, ovvero CIDOC CRM, adottato come standard da ISO con l'introduzione della norma ISO 21127:2014. Tale modello fornisce un *framework* concettuale ricco e articolato, adatto a rappresentare

entità e relazioni complesse in contesti multidisciplinari. Tuttavia, il livello di astrazione necessario per garantire l'interoperabilità tra domini diversi comporta una minore specializzazione rispetto alle esigenze archivistiche specifiche, lacuna che le estensioni di ArchOnto mirano a colmare. Al contrario, RiC-O, pur essendo un modello ancora in fase di stabilizzazione e non ancora integrato con altri domini del GLAM, risulta concepito appositamente per il contesto archivistico, rispondendo in modo più diretto alle sue esigenze.

La minore complessità strutturale di RiC-O si traduce in una rappresentazione più agile e versatile rispetto a quella richiesta da CIDOC CRM, rendendolo particolarmente adatto all'impiego operativo negli archivi. Infatti, il confronto basato su un caso di studio concreto, a complemento delle analisi di Oliveira et al. (2024), ha reso possibile esplorare le *property chains* necessarie in CIDOC CRM per rappresentare concetti che, in RiC-O, possono essere espressi con un'unica tripla. Un esempio emblematico è la rappresentazione della data di nascita, che in CIDOC CRM richiede l'impiego di più classi e proprietà, mentre in RiC-O è direttamente esplicitabile.

Entrambi i modelli, tuttavia, presentano alcune limitazioni comuni, in particolare nella rappresentazione della *provenance*. Nonostante ciò, il fatto che siano supportati da comunità attive di sviluppatori e utenti ne garantisce la continua evoluzione e il miglioramento. RiC-O dimostra una netta efficacia nei contesti archivistici grazie alla sua applicabilità immediata e alla capacità di adattarsi a una vasta gamma di scenari, dalle attività di catalogazione più semplici a processi complessi di gestione dei dati. Questa flessibilità, tuttavia, può comportare un rischio di frammentazione: istituzioni diverse potrebbero adottare strategie di modellazione divergenti per concetti simili, compromettendo l'interoperabilità. Per superare la frammentazione e il problema diffuso dei data silos, diventa dunque necessario un maggiore sforzo verso la convergenza su modelli e pattern di rappresentazione comuni.

CIDOC CRM, d'altro canto, offre una struttura concettuale robusta e trasversale, in grado di rappresentare una vasta gamma di domini culturali oltre a quello archivistico. Questa caratteristica lo rende uno strumento privilegiato per progetti interdisciplinari, anche grazie all'estensione LRMoo, che ne ha favorito l'integrazione nel dominio bibliotecario. Tuttavia, la sua complessità rappresenta una sfida concreta, soprattutto per quelle istituzioni che non dispongono delle risorse, delle competenze o del tempo necessario a padroneggiarne la curva di apprendimento.

Le prospettive future dovrebbero orientarsi verso un allineamento tra RiC-O e CIDOC CRM all'interno del paradigma LOD e dei principi FAIR. Tale convergenza, oltre a essere auspicabile, è funzionale agli obiettivi a lungo termine delineati da EGAD (Clavaud 2023), ovvero la costruzione di descrizioni

archivistiche sempre più integrate e interoperabili. In questa direzione, l'allineamento semantico tra i modelli potrà contribuire a colmare le distanze tra domini diversi, preservando al contempo la specificità dei contesti archivistici e culturali.

Questo studio intende inoltre sottolineare l'importanza di convertire gli inventari archivistici da EAD/XML a LOD, al fine di migliorarne la reperibilità, l'accessibilità, l'interoperabilità e la riusabilità, in linea con i principi FAIR. In tale prospettiva, è fondamentale andare oltre la mera conversione tecnica e avviare vere e proprie operazioni di estrazione dell'informazione, volte a strutturare entità e relazioni latenti nei campi testuali. Una simile strategia permette di arricchire semanticamente i dati e di valorizzare pienamente la natura interconnessa dei LOD, migliorando al contempo la capacità di scoperta e il riuso in contesti trasversali. Il capitolo successivo approfondisce queste dinamiche, soffermandosi in particolare sulle strategie di recupero e trasformazione del patrimonio informativo pregresso e sulle sfide connesse alla migrazione semantica degli inventari tradizionali verso modelli a semantica esplicita.

5.3 Migrazioni semantiche di inventari tradizionali

Il recupero del patrimonio informativo pregresso, ossia dei dati e delle descrizioni già codificati secondo standard e sistemi precedenti, rappresenta uno dei nodi critici nella transizione verso nuovi paradigmi descrittivi (Di Marcantonio 2023, 7). L'obiettivo di creare bacini di conoscenza comprensibili alle macchine, capaci di favorire un accesso integrato e coerente alle risorse culturali, non si limita all'adozione di ontologie condivise, ma richiede anche strumenti e strategie per integrare e valorizzare il patrimonio già esistente.

La migrazione di dataset culturali verso il paradigma LOD si concretizza nella trasposizione di dati rappresentati con modelli a semantica implicita verso modelli a semantica esplicita. Questa trasformazione ha dimostrato, attraverso un serie casi di studio, un potenziale trasformativo considerevole sul piano dell'interoperabilità semantica, della valorizzazione informativa e della sostenibilità a lungo termine delle risorse digitali. Esperienze come WarSampo, BiographySampo (Hyvönen et al. 2021), l'archivio fotografico di Federico Zeri (Daquino et al. 2016) e i progetti di Europeana nell'ambito dell'aggregazione semantica (Charles e Isaac 2020) hanno dimostrato come la strutturazione formale dei dati attraverso modelli ontologici consenta una migliore integrazione interistituzionale e un incremento sostanziale delle possibilità di riuso e analisi automatizzata (Hawkins 2022).

Nonostante le potenzialità, l'applicazione dei LOD alle descrizioni archivistiche esistenti presenta sfide significative: la conversione dei dati in formato strutturato richiede infatti un notevole lavoro di arricchimento e normalizzazione, oltre allo sviluppo di strumenti adeguati alla ricerca. Qing Zou e Eun G. Park sottolineano inoltre che l'efficacia dei LOD dipende dalla preesistente disponibilità di descrizioni archivistiche standardizzate e ben strutturate, evidenziando due criticità particolarmente diffuse: da un lato, la frammentarietà delle informazioni sulla provenienza e sul contesto archivistico; dall'altro, la prevalenza di descrizioni in forma testuale non strutturata (Zou e Park 2024, 815).

Il Web Semantico, infatti, non ha sconvolto completamente l'approccio tradizionale delle istituzioni alla descrizione, ma ha enfatizzato la necessità di adottare una semantica esplicita e granulare per consentire un'interoperabilità basata sull'impiego di modelli concettuali, facilitando così la comprensione e il riuso dei dati. Come evidenziato da Jenny Bunn, «work is already being undertaken in some institutions to convert the information they already hold into the more atomised format of such assertions, using for example the RDF standard (often serialised as either XML or JSON) to ensure the wider interoperability and tractability of the information so atomised» (2021, 3).

In questo scenario, si impone una riflessione critica sulle pratiche descrittive tradizionali adottate dagli archivi e sulle modalità con cui la componente testuale, da sempre centrale nella pratica archivistica, possa essere riformulata in termini semantici per potenziarne il valore computazionale. Campi testuali quali nota biografica, storia archivistica e criteri di riordinamento racchiudono un patrimonio informativo ricco ma spesso non valorizzato, poiché trattato come semplice testo non strutturato. Il potenziale inespresso di questi contenuti testuali è evidente: pur mantenendo la loro unitarietà in un campo descrittivo dedicato, potrebbero essere strutturati ed esplicitati attraverso nuove triple RDF. Ogni nuova tripla diventerebbe portatrice di una componente informativa specifica presente nel testo aggregato, come attestazioni di istituzioni, persone, eventi e coordinate spazio-temporali. L'estrapolazione della specifica semantica del dato preesistente, attraverso la creazione di triple che attestano relazioni più o meno esplicite nel testo, consentirebbe un arricchimento significativo della *knowledge base* contestuale.

La risposta risiede nell'applicazione sistematica di tecniche di *Knowledge Extraction* (KE). La KE comprende varie attività che mirano all'estrazione automatica di informazioni, sia in dimensioni verticali (ad esempio tassonomie, classi, tipi, entità denominate) sia orizzontali (relazioni), combinando metodologie provenienti dall'elaborazione del linguaggio naturale (NLP), dall'intelligenza artificiale simbolica (*Symbolic Artifical Intelligence*) e dal Machine Learning (Giagnolini, Schimmenti, et al. 2025). Applicando queste metodologie, una fonte di informazioni non strutturate, come il testo semplice, può essere convertita in formati strutturati e leggibili da una macchina. Solo attraverso questo approccio

sistematico sarà possibile realizzare il pieno potenziale della migrazione semantica, trasformando ogni nuova asserzione espressa in forma di tripla in generatrice di inferenza e di nuova informazione, arricchendo progressivamente la rete semantica e trasformandola in informazione classificata e computazionalmente elaborabile.

Il tentativo di «stabilire se e in che misura le tecniche e le tecnologie di gestione del testo possano potenziare i nostri strumenti nel rispetto del contesto, arricchendoli di appigli informativi» (Valacchi 2023, 7), si traduce nel comprendere come impiegare gli strumenti ad oggi disponibili per acquisire informazione strutturata dalle descrizioni archivistiche. A questo scopo, è necessario chiarire gli step del processo di estrazione dell'informazione, ovvero elaborare un modello di workflow che sia capace di contemplare tanto l'esigenza di definire il tipo di analisi da delegare allo strumento, quanto la necessità di valutare degli esiti della sua applicazione. L'approccio che Giagnolini, Tomasi e Bonora (2024) hanno proposto per l'implementazione del processo è articolato nei seguenti punti:

1. Selezionare il tipo di atto interpretativo che si delega allo strumento (ad esempio, analisi morfosintattica, lessicale o semantica) a seconda dei contenuti da analizzare.
2. Individuare le tecnologie e le rispettive implementazioni in funzione del tipo di atto interpretativo atteso (ad esempio, da tecniche di *Natural Language Processing* (NLP) elementari al *Deep Learning* (DL)).
3. Definire il modello di valutazione dell'esito e della qualità dell'atto interpretativo automatico, dove per qualità si intende «la possibilità di attingere a dati ragionevolmente affidabili, perché parte di un contesto che li giustifica e li spiega» (Valacchi 2023, 10). Gli output dell'atto interpretativo dovranno, infatti, essere vagliati e selezionati da un esperto di dominio per essere ritenuti validi.
4. Individuare il modello di rappresentazione e sedimentazione della conoscenza estratta nell'ottica di una struttura semanticamente controllata e interoperabile dal punto di vista dell'accesso al dato (ad esempio, *Entity-Relationship model* in SQL; RDF in prospettiva LOD).
5. Modellare i criteri e le modalità di acquisizione della conoscenza estratta in funzione della capacità espressiva del relativo modello descrittivo (ad esempio, Dublin Core, RiC-O, SAN LOD) e dei criteri redazionali. A questo scopo andrà definito un modello di attestazione della *provenance* del dato che espliciti il tipo di atto interpretativo, lo strumento e il processo utilizzato per ottenerlo, le metriche di valutazione (ad esempio, *recall* e *precision*) e il riferimento al supervisore scientifico (ossia l'attribuzione di responsabilità).
6. Valutare le strategie per mettere in relazione il dato analizzato nel sistema nativo e la serie di triple esito dell'atto interpretativo.

7. Modellare l'interazione utente-sistema in termini di processo operativo e di interfacce, ovvero individuare strategie di *information visualization* che consentano al contenuto informativo estratto di essere adeguatamente presentato e gestito.
8. Valutare potenziali modalità di rinforzo dello strumento esterno per migliorarne le prestazioni (ad esempio, producendo dataset annotati e sfruttando *Pre-trained Models* (PTLMs) o *Large Language Models* (LLMs)).

In questa formulazione, il processo è sufficientemente astratto da poter essere applicato a una varietà di contesti e obiettivi di KE, operando su più livelli (dall'analisi lessicale superficiale all'interpretazione semantica del testo). Ciò implica che l'implementazione specifica può variare in funzione di diversi fattori: la struttura e lo stile linguistico del documento di riferimento, il livello di granularità semantica desiderato, i requisiti ontologici specifici del dominio e l'obiettivo finale del grafo di conoscenza da estrarre. I campi testuali più estesi degli strumenti di corredo archivistici tradizionali comprendono solitamente la nota biografica, la storia archivistica, il sistema di ordinamento e la nota di ambito e contenuto. Per testare le potenzialità dell'approccio metodologico delineato, sono stati individuati due casi di studio[196]:

1. L'inventario archivistico dei registri di battesimo della Parrocchia di Aldoar, approfondito nel capitolo 5.3.1, con un focus sull'analisi del campo di ambito e contenuto dei singoli record;
2. L'inventario archivistico del fondo personale di Andrea Costa, esaminato nel capitolo 5.3.2, con analisi dedicata alla nota biografica del soggetto produttore.

Sebbene questi casi di studio siano correlati a contesti analogici, la riflessione metodologica qui sviluppata è estendibile agli archivi digitali, che costituiscono il focus primario della presente ricerca. La scarsità di sperimentazioni in letteratura e la limitata disponibilità di casi già analizzati in tale ambito rendono opportuno trarre suggestioni dal mondo analogico per inquadrare problemi e soluzioni applicabili anche ai materiali *born-digital*. Va inoltre sottolineato che certe descrizioni testuali, in particolare la nota biografica e la storia archivistica, rimarranno presenti negli strumenti di corredo indipendentemente dalla natura del fondo documentario, garantendo così una continuità metodologica nell'applicazione delle tecniche di estrazione della conoscenza. L'obiettivo di superare i silos informativi

[196] I casi di studio illustrati di seguito provengono dalle esperienze di ricerca pubblicate nei seguenti articoli: Giagnolini, Lucia, Inês Koch, Francesca Tomasi, e Carla Teixeira Lopes. 2025. "Comparative Insights into Semantic Archival Modelling: Evaluating RiC-O and ArchOnto Representation Capabilities." *Journal of Documentation* 81, n. 4 (14 agosto): 1003-1031. https://doi.org/10.1108/JD-12-2024-0310; Giagnolini, Lucia, Andrea Schimmenti, Paolo Bonora, and Francesca Tomasi. 2025. "Expliciting Contexts: Semantic Knowledge Extraction from Traditional Archival Descriptions". *Umanistica Digitale* 9 (20):115-44. https://doi.org/10.6092/issn.2532-8816/21229.

attraverso l'estrazione e la formalizzazione della conoscenza rimane infatti valido per qualsiasi tipologia di fonti: l'approccio qui delineato, proprio per il suo carattere astratto e modulare, si presta a essere adattato a diverse realtà documentarie.

Questi due casi di studio, scelti per la loro diversa natura, consentono di verificare l'applicabilità del workflow proposto a differenti tipologie di testi descrittivi, mostrando come sia possibile derivare nuova conoscenza strutturata e favorire un arricchimento dei dati archivistici. Al tempo stesso, essi rappresentano un banco di prova concreto per la migrazione semantica degli inventari tradizionali e per il recupero del patrimonio informativo pregresso, dimostrando come l'estrazione e la formalizzazione della conoscenza latente nei testi possano costituire un passaggio chiave nel processo di integrazione dei dati archivistici nel paradigma dei LOD.

5.3.1 Approccio rule-based

In questo capitolo viene illustrata la metodologia *rule-based* (Waltl et al. 2018) adottata per la migrazione semantica dell'inventario archivistico dei registri di battesimo della Parrocchia di Aldoar (Porto, Portogallo), condotta nell'ambito del progetto presentato al capitolo 5.2.2. L'obiettivo principale della migrazione è stato l'arricchimento semantico delle descrizioni archivistiche e dei relativi contesti, mediante la destrutturazione delle informazioni testuali presenti nel campo *scope and content* dei singoli record e la loro rielaborazione in RDF, al fine di popolare due grafi di conoscenza distinti: uno modellato secondo ArchOnto, l'altro secondo RiC-O.

Per quanto riguarda la rappresentazione tramite ArchOnto, il processo di estrazione era già stato condotto nell'ambito del più ampio progetto EPISA, basandosi su dati già migrati in formato OWL anziché sull'XML EAD originale. L'estrazione ha interessato informazioni provenienti da diversi strumenti di corredo e dai vari elementi ISAD(G), sfruttando tecniche di *Natural Language Processing* (NLP) (Varagnolo et al. 2023). In particolare, il processo adottato da EPISA si articola in tre fasi:

1. un classificatore viene applicato alla rappresentazione in OWL;
2. per ciascun campo testuale classificato e collegato a un documento (`E31 Document`), il testo e il relativo codice di riferimento del documento vengono inviati a un processo di estrazione delle informazioni;
3. il processo di estrazione identifica un insieme di relazioni che, mediante le regole di mapping in OWL, rappresentano le informazioni estratte secondo il modello CIDOC CRM, ricollegandole collegandole al documento da cui provengono nella *knowledge base* finale (Varagnolo et al. 2022).

L'approccio seguito per RiC-O, invece, si distingue per il diverso punto di partenza: non si basa sulla classificazione testuale di informazioni già convertite in OWL, ma sulla migrazione a partire da XML EAD**.** Anche in questo caso, il processo è articolato in tre fasi[197]:

1. identificazione degli scenari di rappresentazione più rilevanti (documento archivistico, evento di nascita, battesimo e attestazione della *provenance*) e loro modellazione tramite mappe RiC-O e ArchOnto (v. cap. 5.2.2);
2. analisi del file EAD XML dei registri di battesimo della parrocchia di Aldoar, con estrazione delle informazioni chiave dai testi descrittivi, quali nomi, date e relazioni familiari;
3. migrazione dei dati da XML a RiC-O tramite RiC-O Converter, al fine di ottenere due dataset comparabili in LOD.

Durante la seconda fase dello studio è stata condotta un'analisi approfondita del file XML EAD originale, contenente i registri di battesimo della parrocchia di Aldoar (cfr. Listing 1 per un esempio di file EAD XML). Secondo il *workflow* proposto da Giagnolini, Bonora e Tomasi (2024) l'obiettivo riscontrabile per lo step 1 è quello di individuare specifici elementi strutturati nelle descrizioni; si è quindi proceduto con l'isolamento delle informazioni pertinenti dalle stringhe descrittive di testo, quali il nome del bambino, la data di nascita, i nomi dei genitori, dei nonni, della madrina e del padrino, nonché altri dettagli.

```
<c level5"item" id5"000001">
  <did>
    <langmaterial>Por (português)</langmaterial>
    <origination label5"RecipientAddress">Lugar de Vilarinha, Freguesia de Aldoar
    </origination>
    <physdesc>
      <dimensions>120x210 mm; papel</dimensions>
    </physdesc>
    <physloc>E/20/6/3 - 9.2 - fl. 1, ass. 1</physloc>
    <repository>Arquivo Distrital do Porto</repository>
    <unitdate label5"UnitDates" type5"inclusive" certainty5"True/True" normal5"
    1707-07-20/1707-07-20">1707-07-20/1707-07-20</unitdate>
    <unitid countrycode5"PT" repositorycode5"PT-ADPRT">000001</unitid>
    <unittitle type5"Formal">Registo de batismo de Maria </unittitle>
```

[197] Le risorse generate attraverso il processo comprendono: immagini rappresentanti i quattro scenari individuati (evento di nascita, attività di battesimo, descrizione del record e documentazione della *provenance*) modellati secondo RiC-O e ArchOnto (Figure 5.7-5.16); lo script Python per l'estrazione dei dati; il file XML modificato per la conversione del formato, comprensivo delle nuove informazioni estratte; file XSLT del RiC-O Converter aggiornato; file RDF contenente il dataset conforme a RiC-O. Questi materiali, insieme al file XML EAD originale fornito dall'Archivio Distrettuale di Porto, sono disponibili nel *repository* ufficiale di INESC TEC (Giagnolini e Koch 2024), al fine di offrire una visione completa delle metodologie e degli strumenti impiegati, garantendo trasparenza e riproducibilità per ulteriori ricerche.

```
  </did>
  <scopecontent>
    <p>Pais: Manuel Antonio e Ana da Silva Data de nascimento: não mencionado</p>
  </scopecontent>
  <odd label5“OriginalNumbering”>
    <p>M2</p>
  </odd>
</c>
```

Listato 5.1. Registrazione di battesimo di Maria da Silva in XML EAD.

Il processo di estrazione ha previsto l'uso combinato di espressioni regolari per individuare scenari specifici all'interno delle stringhe, della libreria lxml[198] per la manipolazione dei tag XML e della libreria datetime[199] per la gestione e la formattazione delle date (step 2 del *workflow*). Si tratta di un'operazione di *pattern matching*, ossia di riconoscimento automatico di pattern testuali predefiniti nelle stringhe descrittive sulla base di regole predefinite. Le informazioni estratte sono state poi riorganizzate in formato XML (si veda il Listato 5.2 per un esempio di XML arricchito), arricchendo i file iniziali con nuovi tag (ad esempio: *child*, *mother*, *father*, *birth date* e altri). Pur essendo al di fuori del dominio EAD, questi tag hanno fornito una base strutturata e semanticamente chiara per il successivo processo di conversione. La strutturazione in elementi XML distinti ha inoltre facilitato le operazioni di controllo e validazione dell'esito dell'estrazione, effettuate manualmente (step 3 del *workflow*).

```
<c level5“item” id5“000001”>
  <did>
    <langmaterial>Por (português)</langmaterial>
    <origination label5“RecipientAddress”>Lugar de Vilarinha, Freguesia de Aldoar
    </origination>
    <physdesc>
      <dimensions>120x210 mm; papel</dimensions>
    </physdesc>
    <physloc>E/20/6/3 - 9.2 - fl. 1, ass. 1</physloc>
    <repository>Arquivo Distrital do Porto</repository>
    <unitdate label5“UnitDates” type5“inclusive” certainty5“True/True” normal5”
    1707-07-20/1707-07-20”>1707-07-20/1707-07-20</unitdate>
    <unitid countrycode5“PT” repositorycode5“PT-ADPRT”>000001</unitid>
    <unittitle type5“Formal”>Registo de batismo de Maria</unittitle>
  </did>
  <scopecontent>
    <p>Pais: Manuel Antonio e Ana da Silva Data de nascimento: não mencionado</p>
  </scopecontent>
  <odd label5“OriginalNumbering”>
    <p>M2</p>
  </odd>
  <child>Maria da Silva</child>
  <birth_date>não mencionado</birth_date>
```

[198] https://lxml.de/.

[199] https://docs.python.org/3/library/datetime.html.

```
  <father>Manuel Antonio</father>
  <mother>Ana da Silva</mother>
</c>
```

Listato 5.2. Registrazione di battesimo di Maria da Silva in XML con tag aggiuntivi derivati dal processo di knowledge extraction.

L'uso delle espressioni regolari ha consentito di ottenere rapidamente risultati affidabili, permettendo di concentrarsi direttamente sui requisiti di rappresentazione, obiettivo principale dello studio. Tuttavia, ciò è stato possibile solo grazie all'elevata omogeneità dei testi; per progetti di più ampia scala, caratterizzati da una maggiore varietà di tipi di stringhe, strutture e stili testuali, è necessario ricorrere a tecniche di NLP o a strumenti basati su AI (Babaei Giglou et al. 2023; Giagnolini, Schimmenti, et al. 2025), come verrà approfondito nel capitolo successivo.

Nella terza fase, per la conversione verso RDF, è stato selezionato lo strumento *open source* messo a disposizione dagli Archives Nationales de France, ovvero RiC-O Converter[200] (Clavaud et al. 2023; Francart 2024). Si tratta di un applicativo a riga di comando in Java, basato prevalentemente su fogli di stile XSLT, che consente la conversione di file XML in formato EAD ed EAC in dataset RDF conformi al modello RiC-O. La scelta di questo strumento ha permesso di concretizzare lo step 4 del workflow: l'individuazione di un modello di rappresentazione e sedimentazione della conoscenza estratta in una struttura semanticamente controllata e interoperabile[201].

Seguendo le linee guida indicate nella documentazione, il convertitore è stato adattato per rispecchiare le specificità del dataset in esame, al fine di ottenerne la rappresentazione in RDF. Le principali modifiche hanno riguardato il file *ead2rico.xslt*, contenuto nella cartella dedicata alla conversione degli EAD e responsabile della logica di trasformazione. Gli interventi hanno riguardato principalmente: l'inserimento del mapping di tag non previsti dal convertitore ma presenti nel file EAD originale, come *phylsoc* e *bioghist*; l'inserimento del mapping dei tag generati durante la procedura di estrazione dei dati, come *father* e *mother*; la creazione e modifica di mapping in grado di rispondere alle esigenze di rappresentazione del nostro caso di studio. Un esempio riguarda la gestione delle date: nell'impostazione di default di RiC-O Converter, esse vengono tradotte di default come *data property* (`rico:date`, `rico:beginningDate`, `rico:endDate`). Per rispondere ai requisiti del progetto, le date sono state invece modellate come istanze della classe `rico:Date` e arricchite attraverso l'uso di *object property* come

[200] https://archivesnationalesfr.github.io/rico-converter/en/.

[201] Non è stato possibile implementare gli step 5, 6 e 7 (acquisizione della conoscenza e integrazione nel sistema nativo, e interazione utente-sistema) poiché la collaborazione con gli Archivi Nazionali portoghesi si è limitata alla fornitura dei file XML EAD, senza includere il reinserimento delle informazioni estratte nel sistema DigitArq.

`rico:isBirthDateOf`, `rico:isCreationDateOf`, `rico:isBeginningDateOf` e `rico:isEndDateOf`. Infine, tutti i riferimenti specifici agli Archives Nationales de France, relativi in particolare alla gestione delle informazioni su autorità, URI e vocabolari sono stati rimossi.

RiC-O Converter si è dimostrato uno strumento fondamentale per ottenere una rappresentazione RDF conforme a RiC-O e in linea con i requisiti di rappresentazione delineati al capitolo 5.2.2., consultabile e scaricabile dal *repository* INESC TEC dedicato al progetto[202] (Giagnolini e Koch 2024). La natura *open source* di RiC-O Converter, insieme alla relativa facilità di utilizzo e personalizzazione, rappresenta un avanzamento degno di nota per la comunità archivistica che si confronta con la migrazione dei dati verso modelli a semantica esplicita. Tuttavia, senza un'adeguata fase di customizzazione e raffinamento dei dati in sede di estrazione, il rischio è quello di ottenere una conversione meccanica "uno a uno" da XML a RDF, inadatta nel restituire la ricchezza delle descrizioni e, dunque, continuando a trattenere la valorizzazione dei contesti. Si conferma dunque di grande importanza il processo di esplicitazione semantica delle stringhe testuali, così da garantire una rappresentazione realmente espressiva e coerente con le potenzialità dei modelli ontologici.

5.3.2 Approccio generativo

In questo capitolo viene presentata una metodologia per l'estrazione di conoscenza dalle note biografiche contenute negli inventari archivistici. L'analisi si è concentrata sul caso di studio del Fondo Andrea Costa, sviluppato nell'ambito del progetto di testing del workflow descritto al capitolo 5.3. L'obiettivo è stato verificare l'efficacia di diverse tecniche, partendo da strumenti NLP tradizionali fino ad arrivare a soluzioni basate su Large Language Models (LLM), per l'estrazione di conoscenza strutturata da testi biografici (non strutturati).

Le note biografiche presenti negli strumenti di corredo archivistici presentano un grado di complessità superiore rispetto alle note di ambito e contenuto analizzate nel capitolo precedente. A differenza delle descrizioni dei registri parrocchiali, caratterizzate da schemi regolari che hanno reso efficace un approccio *rule-based*, esse seguono convenzioni narrative più articolate, che comportano la necessità di riconoscere entità nominate, ricostruire relazioni temporali complesse, individuare eventi interconnessi e collocarli in contesti storici con diversi livelli di dettaglio.

Pur presentando questa complessità, le note biografiche archivistiche risultano più regolari e prevedibili delle biografie letterarie, ad esempio, nelle quali prevalgono intrecci narrativi, interpretazioni soggettive

[202] https://rdm.inesctec.pt/dataset/cs-2024-009.

e soluzioni stilistiche idiosincratiche. Ciò consente di evitare i livelli più onerosi di *parsing* semantico tipici dell'analisi di testi letterari, pur richiedendo strumenti che vadano oltre il semplice *pattern matching* o l'applicazione di regole. Per questo motivo, si è avviata una *proof of concept* basata sull'applicazione di tecnologie generaliste *open source* per l'estrazione di informazioni strutturate dal primo paragrafo della nota biografica del fondo archivistico di Andrea Costa[203].

Per l'analisi iniziale sono stati individuati i seguenti strumenti (fasi n. 1 e n. 2 del workflow):

1. Tint[204], tool per analisi linguistica, tra cui *part-of-speech tagging*, *parsing* delle dipendenze sintattiche e il riconoscimento di entità nominate (NER), con modelli specificamente addestrati su testi in lingua italiana (Palmero Aprosio e Moretti 2019);
2. FRED[205], strumento di lettura automatica in grado di rappresentare il discorso. È indipendente dal dominio e concepito per essere utilizzato come middleware, in particolare per attività di Estrazione della Conoscenza (Gangemi et al. 2017);
3. modelli linguistici pre-addestrati (PTLM), e in particolare LLM, che rappresentano strumenti preziosi per qualsiasi attività di analisi testuale basata sul significato di KE. Per questa fase è stato selezionato, tra i LLM disponibili, GPT-4o[206], utilizzato tramite l'interfaccia ChatGPT.

Attraverso l'applicazione di tali strumenti al primo paragrafo della nota biografica di Andrea Costa, si è inteso valutare la loro idoneità per l'attività di KE.

L'applicazione di Tint all'analisi del primo paragrafo della nota biografica ha permesso di identificare organizzazioni, luoghi e nomi di persona tramite NER (Figura 5.17), nonché le dipendenze sintattiche presenti nel testo. Inoltre, lo strumento ha automaticamente classificato le parti del discorso (Figura 5.18)[207].

[203] http://www.san.beniculturali.it/web/san/dettaglio-soggetto-produttore?id=66756.
[204] https://github.com/dhfbk/tint.
[205] http://wit.istc.cnr.it/stlab-tools/fred/.
[206] Si veda la system card del modello: https://openai.com/it-IT/index/gpt-4o-system-card/.
[207] Per accedere ai risultati completi dell'analisi effettuata con Tint, si veda Giagnolini, Lucia e Paolo Bonora. "Refining Context: Extracting Structured Information for Archival Context Enrichment. Results Of The Analysis Performed With Tint". Figshare, 31 gennaio 2024. https://doi.org/10.6084/m9.figshare.25119116.v4.

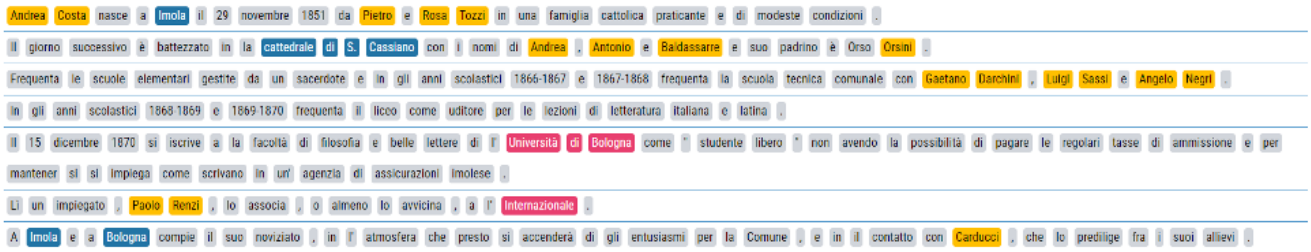


Figura 5.17. Entità riconosciute nel testo e relativo classificazione.

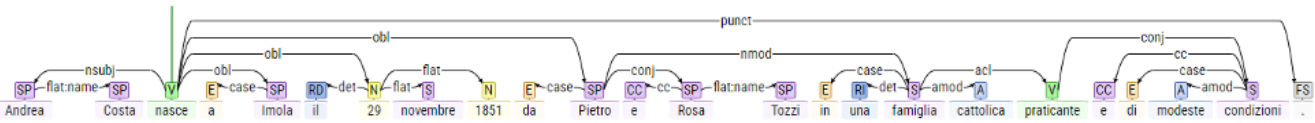


Figura 5.18. Dipendenze sintattiche relative alla prima frase.

Tint si dimostra efficace nell'identificazione della struttura morfologica del testo e delle entità denominate di base. Attraverso metodi di estrazione *rule-based* aggiuntivi è possibile generare triple relative a eventi specifici (ad esempio, estrarre le informazioni di nascita individuando verbi come *nascere* e i loro argomenti corrispondenti). Tuttavia, tale approccio, come molte *pipeline* NLP tradizionali, presenta alcune limitazioni:

- sebbene robusto nell'analisi linguistica di base, fornisce una comprensione semantica limitata;
- la necessità di creare manualmente regole specifiche per ogni tipologia di informazione da estrarre riduce la scalabilità;
- l'output richiede un'ampia validazione da parte di esperti, limitando fortemente i benefici dell'automazione;
- gli approcci basati su regole incontrano difficoltà nel gestire variazioni linguistiche e contestuali.

Tint offre dunque il vantaggio di una rapida applicabilità all'analisi linguistica di base, con uno sforzo di configurazione minimo; la sua utilità rimane tuttavia circoscritta alle fasi preliminari di elaborazione del testo e al riconoscimento elementare delle entità. Tale limitazione non è peculiare di Tint, ma caratterizza in generale gli approcci NLP tradizionali fondati in larga misura su regole linguistiche esplicite e sul *pattern matching* (Shiri et al. 2024).

FRED, invece, annota i frame semantici presenti nel testo. Per la frase, tradotta in inglese, «Andrea Costa was born in Imola on November 29, 1851, to Pietro and Rosa Tozzi in a practicing Catholic family of modest means», FRED è in grado di identificare e rappresentare diverse relazioni semantiche. Inoltre,

esegue operazioni di NER ed *Entity Linking* con DBpedia. L'applicazione di FRED ha prodotto una rappresentazione a grafo unificata e formalizzata dei fatti e dei concetti espressi dal testo in linguaggio naturale, come, ad esempio, l'interpretazione delle condizioni di nascita di Andrea Costa (Figura 5.19).

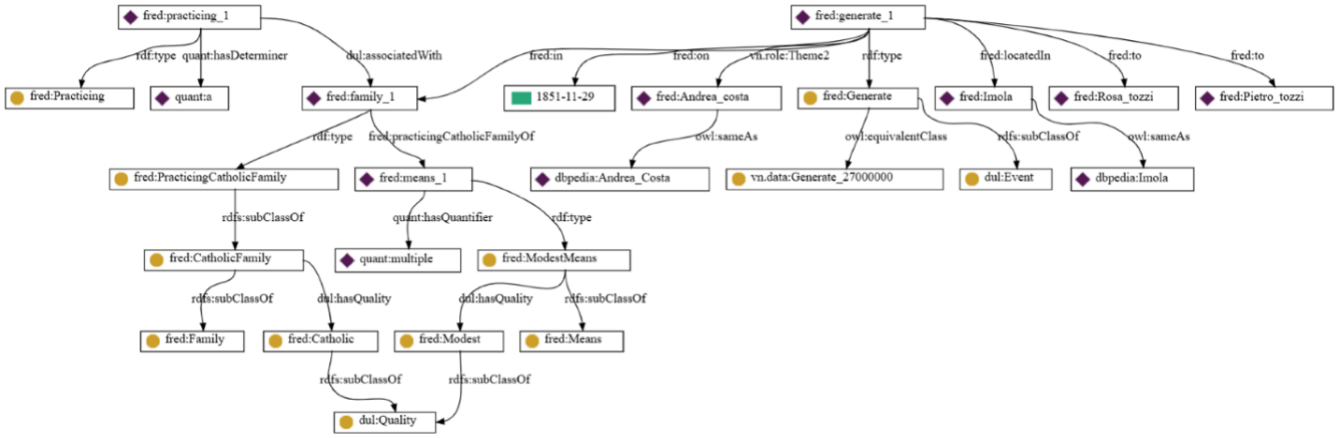


Figura 5.19. Grafo dell'interpretazione delle condizioni di nascita elaborato da FRED (sulla base del testo tradotto in inglese).

L'evento centrale (`fred:Generate`) è connesso a informazioni temporali (`fred:on 1851-11-29`), localizzazione (`fred:locatedIn fred:Imola`), soggetto (`fred:Andrea_costa`, *collegato a* `dbpedia:Andrea_Costa`), genitori (relazioni `fred:to` con `fred:Pietro_tozzi` e `fred:Rosa_tozzi`). Il contesto familiare è rappresentato mediante una struttura gerarchica complessa: *fred:family_1* è classificata come `fred:PracticingCatholicFamily`. Tale classificazione è ulteriormente articolata attraverso relazioni di sottoclasse: `fred:CatholicFamily` - `fred:Family`; le proprietà `fred:Catholic` e `fred:Modest` sono collegate tramite relazioni di qualità (`quality relationships`). La condizione economica (`ModestMeans`) è espressa mediante opportuni quantificatori.

Tuttavia, utilizzando la medesima frase nell'originale italiano si ottengono risultati differenti: le classi non vengono identificate correttamente, il grafo risulta maggiormente distribuito in senso orizzontale e compaiono numerosi nodi intermedi superflui (Figura 5.20).

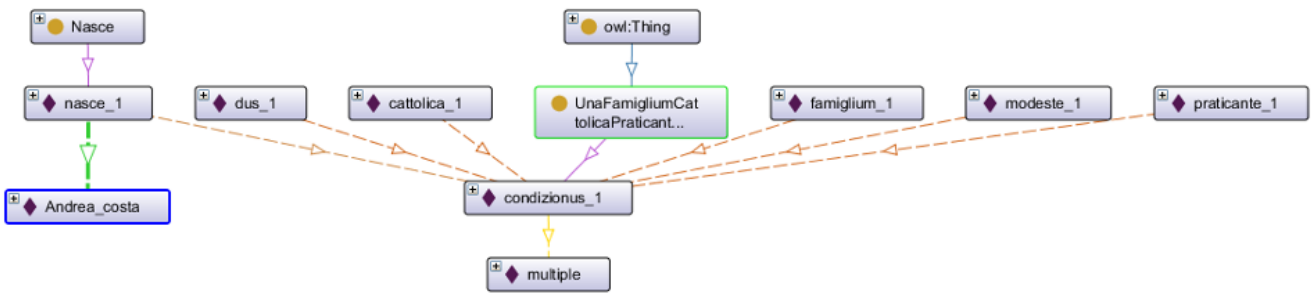


Figura 5.20. Grafo dell'interpretazione delle condizioni di nascita elaborato da FRED con il testo originale in italiano.

Sebbene FRED dimostri un'elevata qualità rispetto a diversi tipi di input e si riveli promettente nell'elaborazione in lingua inglese, la sua caratteristica produzione di morfologie grafiche altamente variabili comporta che le strutture risultanti necessitino generalmente di un impiego diretto, piuttosto che di un allineamento con ontologie standard.

L'ultimo tool testato, GPT-4, è stato impiegato per esplorare l'emergere di informazioni attraverso *prompt* concepiti secondo strategie differenti e finalizzati a specifici obiettivi. A titolo di esempio, il *prompt* «Elenca le entità che emergono dal test» ha prodotto i seguenti risultati:

- **luogo di nascita**: Imola (29 novembre 1851);
- **genitori**: Pietro e Rosa Tozzi;
- **contesto familiare**: Famiglia cattolica praticante di condizioni modeste;
- **atti religiosi**: Battezzato nella cattedrale di S. Cassiano con i nomi Andrea, Antonio e Baldassarre; padrino Orso Orsini;
- **formazione scolastica**: Scuola primaria gestita da un sacerdote; frequentò la scuola tecnica municipale con Gaetano Darchini, Luigi Sassi e Angelo Negri; seguì come uditore le lezioni di letteratura italiana e latina al liceo;
- **percorso universitario**: Iscritto, in qualità di "studente libero", alla Facoltà di Filosofia e Belle Arti dell'Università di Bologna il 15 dicembre 1870;
- **situazione economica**: Impossibilitato a sostenere le ordinarie tasse universitarie; si mantenne lavorando come scrivano presso un'agenzia di assicurazioni a Imola;
- **influenze politiche**: Entrò in contatto con l'Internazionale grazie a Paolo Renzi, impiegato presso la stessa agenzia di assicurazioni;
- **contesto storico-politico**: Periodo di noviziato a Imola e a Bologna in un clima di entusiasmo per la Comune; rapporti con Carducci, che lo annoverò tra i suoi allievi prediletti.

GPT-4o, essendo ampiamente testato nel campo della *Natural Language Understanding* comprendente attività quali il *question answering*, la disambiguazione e il recupero dell'informazione, ha fornito una risposta articolata e all'altezza delle aspettative (Polley et al. 2021). Tuttavia, la trasformazione di tali informazioni, di elevata qualità, in un grafo di conoscenza equivalente non risulta un processo immediato. In tale prospettiva, sono stati condotti ulteriori esperimenti richiedendo al modello di annotare il testo

mediante una semplice sintassi *Entity-Relationship* (ER), quale quella offerta da Mermaid[208], che ha prodotto un risultato incoraggiante (Figura 5.21).

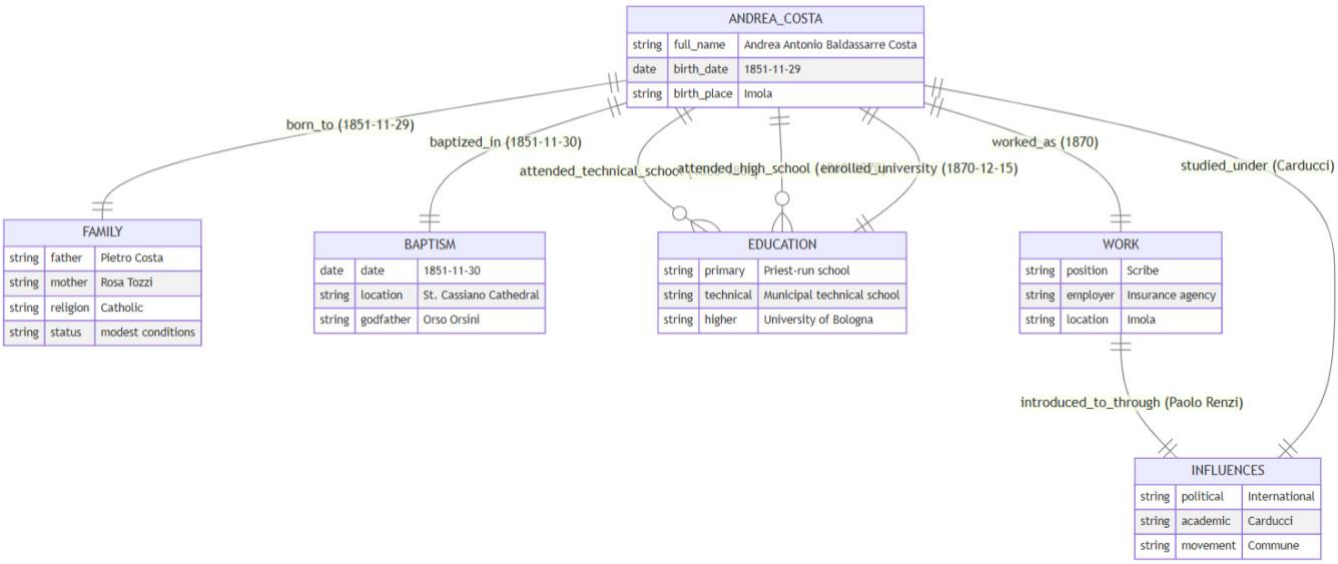


Figura 5.21. Output di GPT-4o relativo alle condizioni di nascita di Andrea Costa (visualizzazione creata da mermaid.live).

Nel complesso, Tint, FRED e GPT-4o hanno restituito opzioni valide per la strutturazione delle informazioni, ma ogni output deve essere sottoposto a validazione mediante supervisione (passo n. 3 del *workflow*). Tale fase di validazione è cruciale, poiché i sistemi di estrazione automatica, nonostante la loro sofisticazione, non possono sostituire completamente la competenza umana nell'interpretazione dei contenuti delle descrizioni archivistiche. La qualità delle informazioni estratte deve essere valutata non solo in termini di accuratezza fattuale, ma anche per la loro rilevanza contestuale e coerenza semantica all'interno del più ampio quadro archivistico.

Nel complesso, i risultati forniti da GPT-4 si sono rivelati particolarmente promettenti. Gli output ottenuti tramite GPT-4o hanno fornito un insieme di informazioni che va oltre la semplice identificazione di entità canoniche. In generale, i modelli basati su architettura transformer hanno mostrato solide performance nella cattura di relazioni semantiche complesse e dipendenze contestuali, mediante apprendimento basato su esempi e processi di ragionamento induttivo (Brown et al. 2020; Wei, Wang, et al. 2022). Inoltre, poiché gli LLM hanno evidenziato quelle che solitamente vengono definite *emergent abilities*, ossia la capacità di generalizzare e fornire risultati anche in presenza di compiti e input precedentemente non visti (Wei, Tay, et al. 2022), possono essere istruiti anche senza un addestramento mirato per nuove attività.

[208] Mermaid è un *markdown* basato su JavaScript per la definizione di diagrammi: https://mermaid.js.org/.

Tuttavia, l'adozione di un approccio di estrazione aperta, in cui l'LLM è libero di descrivere qualsiasi relazione riscontri nel testo, risulta considerevolmente più semplice rispetto all'adesione a uno schema prestabilito. Pur eccellendo in scenari di *few-shot learning* e nella gestione di linguaggi specifici di dominio (Brown et al. 2020), tali modelli presentano limitazioni rilevanti per il nostro compito, in particolare nel mantenere una struttura di output coerente e nell'aderire a schemi rigidi di KE. Gli LLM, in generale, soffrono della mancanza di meccanismi interni di controllo strutturato, con una variabilità intrinseca nei loro output generativi.

In aggiunta, le preoccupazioni etiche legate all'utilizzo di modelli non *open source* sono rilevanti, spaziando dalla mancanza di trasparenza sui dati e processi di addestramento alla limitata responsabilità per bias ed errori. Affidarsi a una specifica funzionalità di un'azienda privata limita la riproducibilità nell'ambito accademico, poiché i ricercatori potrebbero non essere in grado di verificare o sviluppare i risultati pubblicati a causa di modifiche nelle API, nei modelli di *pricing* o nelle politiche di accesso. Questa dipendenza da soluzioni proprietarie pone sfide significative per la sostenibilità della ricerca a lungo termine e per la validazione scientifica.

Date queste premesse e in considerazione dei risultati promettenti, per esplorare più approfonditamente l'impiego degli LLM nella KE da testi archivistici non strutturati e affrontare le criticità emerse è stata sviluppata una pipeline basata su Llama, una famiglia di modelli *open source* dalle prestazioni comparabili a GPT-4o[209]. La pipeline, sviluppata in collaborazione con Andrea Schimmenti, è stata progettata per convertire testi archivistici non strutturati in dati strutturati semanticamente interoperabili. La pipeline, descritta in dettaglio nella sezione 4 dell'articolo *Expliciting contexts: Semantic knowledge extraction from traditional archival descriptions* (Giagnolini, Schimmenti, et al. 2025), è illustrata dalla Figura 5.22 e può essere sintetizzata come segue[210]:

[209] In particolare, Llama-3.3-70B-Instruct ottiene 86.0 in MMLU e 92.1 in IFEval, *benchmark* rispettivamente dedicati alla comprensione del linguaggio naturale e alla capacità di seguire istruzioni. La *model card* è disponibile al link: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

[210] Il codice completo sviluppato per la *proof of concept* è disponibile nel *repository* dedicato al link: https://github.com/aschimmenti/expliciting-context.

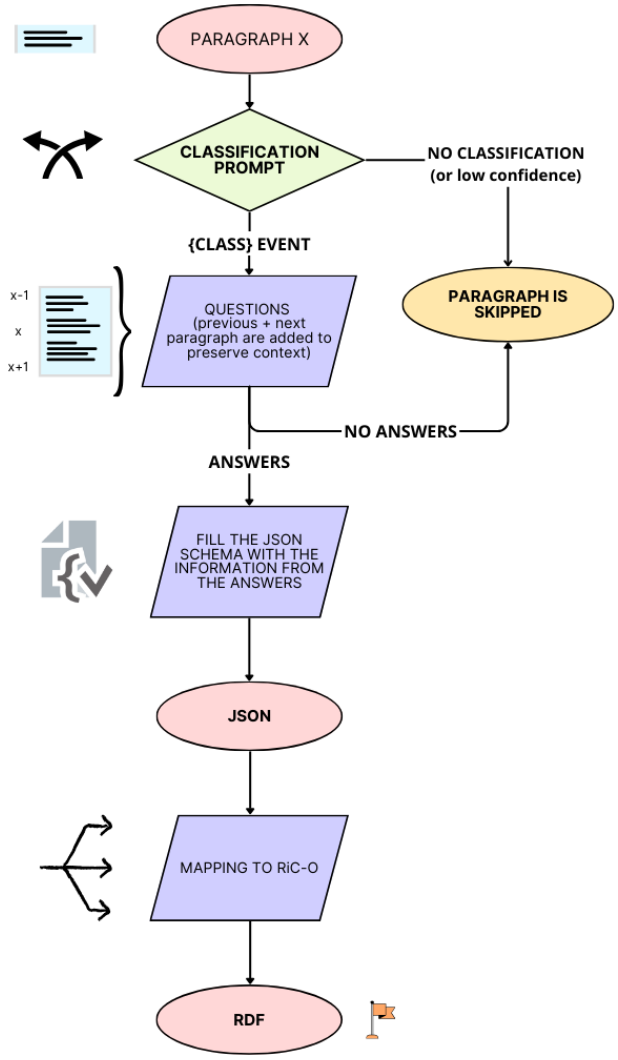


Figura 5.22. Diagramma di flusso della pipeline implementata.

1. **Classificazione degli eventi.** La prima fase consiste nell'identificazione automatica di eventi biografici presenti nel testo. A tal fine, il modello riceve un *prompt* costruito per classificare ogni segmento di testo in una o più tra sette categorie predefinite: nascita, morte, istruzione, impiego, relazioni personali e professionali, eventi politici e creazione documentaria. Tali classi sono ispirate al modello ontologico DOLCE+DnS (Borgo et al. 2022), in cui ogni evento è interpretato come una "situazione" in cui partecipano entità con ruoli distinti, in un contesto spazio-temporale specifico. Il modello restituisce una lista JSON contenente le classi individuate per ogni porzione di testo, ciascuna accompagnata da un livello di confidenza della classificazione (dove 0 equivale al minimo e 1 al massimo della certezza dell'assegnazione). Solo le classificazioni con un livello superiore a 0.7 vengono ammesse alla fase successiva. L'adozione di questo approccio ibrido (tra

classificazione chiusa e generazione aperta) consente di coniugare flessibilità analitica e controllo ontologico.

2. **Estrazione delle informazioni.** Per ciascuna classe individuata, il modello è incaricato di estrarre le informazioni salienti tramite un sistema a domande guidate (*prompt* strutturati): si richiedono esplicitamente le entità coinvolte, i ruoli, le date, i luoghi e altre proprietà contestuali. Viene fornito anche il contesto testuale (paragrafi precedenti e successivi) per aumentare la capacità disambiguante del modello. Questa fase, mantenuta separata dalla successiva per motivi di controllo, si configura come un atto interpretativo supervisionato del testo.
3. **Rappresentazione strutturata.** Una volta estratte, le informazioni vengono trasformate in un output strutturato conforme a uno schema semantico definito per ogni tipo di evento. Lo schema, ancora una volta ispirato a DOLCE+DnS e implementato in formato JSON, viene infine allineato con l'ontologia RiC-O (Records in Context Ontology) e tradotto in RDF/Turtle.

Per esemplificare in modo concreto il funzionamento della *pipeline*, riportiamo l'output generato a partire dalla prima frase della nota biografica di Andrea Costa:

«Nasce a Imola il 29 novembre 1851 da Pietro e Rosa Tozzi in una famiglia cattolica praticante e di modeste condizioni».

Questa singola frase, attraverso le tre fasi della pipeline, è stata tradotta in un insieme di triple RDF espresse RiC-O che formalizzano l'evento di nascita, con i soggetti coinvolti (Andrea Costa e i genitori), la data, il luogo e le relazioni di parentela (Listato 5.3).

```
@prefix rico: <https://www.ica.org/standards/RiC/ontology#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.org/event/Birth_Andrea_Costa> a rico:Event ;
    rico:hasOrHadParticipant <http://example.org/person/Andrea_Costa>,
        <http://example.org/person/Pietro_Costa>,
        <http://example.org/person/Rosa_Tozzi> ;
    rico:name "Birth of Andrea Costa" ;
    rico:occurredAtDate <http://example.org/date/1851-11-29> .

<http://example.org/family/Andrea_Costa_Family> a rico:Family ;
    rico:hasOrHadMember <http://example.org/person/Andrea_Costa>,
        <http://example.org/person/Pietro_Costa>,
        <http://example.org/person/Rosa_Tozzi> ;
    rico:name "Andrea Costa Family" .

<http://example.org/relation/birth_place_Andrea_Costa> a rico:PlaceRelation ;
    rico:date "1851-11-29"^^xsd:date ;
    rico:placeRelationType "birthPlace" ;
    rico:relationHasSource <http://example.org/place/Imola> ;
    rico:relationHasTarget <http://example.org/person/Andrea_Costa> .
```

```
<http://example.org/relation/family_Pietro_Costa_Andrea_Costa> a
rico:FamilyRelation ;
    rico:familyRelationType "parent-child" ;
    rico:hasBeginningDate "1851-11-29"^^xsd:date ;
    rico:relationHasSource <http://example.org/person/Pietro_Costa> ;
    rico:relationHasTarget <http://example.org/person/Andrea_Costa> .

<http://example.org/relation/family_Rosa_Tozzi_Andrea_Costa> a
rico:FamilyRelation ;
    rico:familyRelationType "parent-child" ;
    rico:hasBeginningDate "1851-11-29"^^xsd:date ;
    rico:relationHasSource <http://example.org/person/Rosa_Tozzi> ;
    rico:relationHasTarget <http://example.org/person/Andrea_Costa> .

<http://example.org/date/1851-11-29> a rico:Date ;
    rico:normalizedDateValue "1851-11-29"^^xsd:date .

<http://example.org/place/Imola> a rico:Place ;
    rico:name "Imola" .

<http://example.org/person/Pietro_Costa> a rico:Person ;
    rico:hasChild <http://example.org/person/Andrea_Costa> ;
    rico:name "Pietro Costa" .

<http://example.org/person/Rosa_Tozzi> a rico:Person ;
    rico:hasChild <http://example.org/person/Andrea_Costa> ;
    rico:name "Rosa Tozzi" .

<http://example.org/person/Andrea_Costa> a rico:Person ;
    rico:hasParent <http://example.org/person/Pietro_Costa>,
        <http://example.org/person/Rosa_Tozzi> ;
    rico:name "Andrea Costa" .
```

Listato 5.3. Rappresentazione RDF dell'evento di nascita di Andrea Costa secondo RiC-O che formalizza soggetti, data, luogo e relazioni a partire dalla descrizione testuale.

Per validare i risultati, è stato adottato un *framework* di valutazione manuale su tre livelli:

1. **Strutturale**: verifica quanto l'output rispetti lo schema previsto, controllando sia le prestazioni complessive sia quelle relative a ciascun tipo di evento. Gli aspetti considerati includono la conformità allo schema (verifica che tutti gli eventi generati rispettino il formato JSON e i requisiti predefiniti per ciascuna classificazione) e la coerenza informativa (assicura che informazioni simili siano rappresentate uniformemente tra eventi diversi). Ad esempio, l'evento di nascita dovrebbe essere associato a campi "date", "place" e "parents". Se nel JSON generato manca uno di questi campi, considerato fondamentale per l'esplicitazione della semantica del testo, viene considerato un errore strutturale. Nel test, il 90,2% degli output è risultato conforme allo schema predefinito.

2. **Informativo**: valuta quanto accuratamente il sistema riesca a estrarre le informazioni dal testo, utilizzando metriche quantitative come precisione (*precision*: misura la proporzione di informazioni correttamente estratte rispetto al totale), richiamo (*recall*: la capacità del modello di identificare tutte le informazioni effettivamente presenti nel testo) e F1-score (media di precisione e richiamo). Per misurare questi valori si utilizza la cosiddetta matrice di confusione, che confronta le informazioni presenti nel testo con quelle riportate nell'output. Un'informazione correttamente individuata e riportata viene considerata un "vero positivo" (*true positive*), mentre un'informazione assente sia nel testo che nell'output rappresenta un "vero negativo" (true negative). Se invece l'output contiene un'informazione che non è presente nel testo o la riporta in modo errato, si parla di "falso positivo" (*false positive*), mentre un'informazione presente nel testo ma non riportata nell'output è classificata come "falso negativo" (*false negative*). Per esempio, se il testo indica che Andrea Costa ha frequentato l'Università di Bologna, ma il modello riporta "Università di Milano", questo costituisce un falso positivo. Se invece l'informazione sull'università non viene riportata affatto, si tratta di un falso negativo. Nel complesso, i risultati ottenuti sono ottimi: la precisione è risultata pari a 0,947, il richiamo a 0,982 e l'F1-score a 0,964. Alcune difficoltà marginali si sono manifestate nelle classi istruzione e impiego, soprattutto nella corretta identificazione di date e ruoli multipli.
3. **Interpretativo**: valuta la fedeltà semantica e contestuale delle informazioni estratte. In pratica, ciascun evento generato dal modello è stato confrontato con il testo originale per verificare che le relazioni tra le entità fossero corrette, che i ruoli fossero attribuiti in modo appropriato e che il contesto temporale e spaziale fosse preservato. Ad esempio, nel caso di un evento di relazione, se il testo afferma che Andrea Costa collaborava con Jules Guesde come giornalista, l'output deve riportare correttamente il ruolo di giornalista, e mantenere il contesto della collaborazione. Se invece l'output indica soltanto "alleati politici" senza precisare il tipo di relazione, il punteggio assegnato alla correttezza dell'interpretazione viene ridotto. Allo stesso modo, per un evento legato ad un impiego professionale, se Andrea Costa lavorava contemporaneamente per due organizzazioni diverse, l'output deve distinguere chiaramente i ruoli e le date di ciascuna posizione. La valutazione è stata effettuata su una scala da 1 a 10, dove il punteggio massimo indica che tutte le informazioni erano state preservate con piena accuratezza. Gli errori sono stati penalizzati in base alla gravità: dati critici mancanti comportano una riduzione di un punto, la mancanza di sotto-eventi viene penalizzata di mezzo punto per elemento, gli errori di annotazione determinano una decurtazione di 0,8 punti e le informazioni completamente errate o inventate ("allucinate") comportano una penalità di un punto. Considerando tutti gli eventi valutati, il

punteggio medio complessivo è risultato pari a 8,8 su 10, con leggere variazioni tra le diverse classi di eventi a seconda della complessità e della precisione dell'output.

La pipeline sviluppata si è dunque dimostrata promettente nel generare strutture semantiche accurate e riutilizzabili a partire da testi archivistici biografici, rappresentando un passo apprezzabile verso l'automazione dei processi di arricchimento contestuale. Dal punto di vista metodologico, l'applicazione di strumenti automatizzati di KE ai metadati testuali dei documenti archivistici offre benefici evidenti, senza però sostituire la valutazione critica umana. L'esperimento con modelli LLM deve essere considerato una funzione complementare e di supporto alla documentazione, in grado di potenziare l'interpretazione dei dati.

I risultati evidenziano sia le potenzialità promettenti della *pipeline* sia le aree che richiedono ulteriori sviluppi. Per un'implementazione più robusta sarà necessario condurre una validazione su corpora diversificati, migliorare la disambiguazione di ruoli e date e sviluppare interfacce *user-friendly* che consentano agli archivisti di intervenire nei processi di validazione. Questi strumenti dovranno integrarsi perfettamente con i sistemi di gestione archivistica esistenti e supportare workflow collaborativi di annotazione e validazione. Per garantire un'adozione ampia e sostenibile, è essenziale sviluppare soluzioni accessibili ed efficienti, preferibilmente basate su LLM *open source*, manutenibili e implementabili da istituzioni con diversi livelli di risorse tecniche. Lo sviluppo di componenti modulari e riutilizzabili, ottimizzati anche per risorse computazionali limitate, sarà cruciale soprattutto per realtà più piccole o centri di ricerca. L'ecosistema dovrà essere sufficientemente flessibile da adattarsi a diverse esigenze istituzionali, mantenendo al contempo elevati standard di accuratezza e affidabilità nella KE.

L'integrazione degli outputs nei sistemi informativi archivistici dipenderà anche dalla capacità di adattare i modelli descrittivi esistenti, come SAN LOD, a una maggiore espressività semantica. In mancanza di questa adattabilità, i dati estratti dovranno essere gestiti come grafi stand-off, collegati ma separati dalle descrizioni originali.

Nell'ottica dell'estensione dei sistemi informativi italiani verso ArCo, diventa quindi interessante proseguire il ragionamento sia sul piano dell'allineamento, sia su quello dell'adattamento della *pipeline* alla semantica di ArCo4Archives, così da garantire continuità e interoperabilità tra diversi modelli e, allo stesso tempo, ampliare la capacità espressiva delle descrizioni archivistiche.

6. Modellazione del digitale d'autore

L'analisi dei modelli e delle ontologie di dominio archivistico condotta nel capitolo precedente ha evidenziato la crescente convergenza verso approcci basati sui LOD come riferimento per la descrizione archivistica a livello internazionale. Tali approcci, centrali anche nel più ampio contesto dei beni culturali, offrono un vantaggio decisivo per la rappresentazione del digitale d'autore grazie alla flessibilità dei modelli. Il nativo digitale, infatti, non presenta continuità o stabilità negli oggetti descritti: ogni archivio combina formati eterogenei, dispositivi di produzione differenti, stratificazioni temporali complesse e metadati tecnici estremamente variabili. Questa natura richiede soluzioni descrittive adattabili, che gli standard tradizionali, concepiti per documenti relativamente uniformi, non sono in grado di garantire. La sfida consiste nel trasformare la complessità fenomenologica del digitale d'autore da ostacolo descrittivo a risorsa informativa.

Per finalizzare la risposta alla RQ2, ossia come rappresentare il digitale d'autore restituendone la complessità e le relazioni contestuali, il presente capitolo sviluppa BoDi (Born-Digital Ontology), un'estensione specializzata di RiC-O dedicata alla rappresentazione del digitale d'autore. La trattazione si articola attraverso l'identificazione di requisiti specifici per la rappresentazione (cap. 6.1) e lo sviluppo del modulo BoDi (cap. 6.2) e la sua validazione (cap. 6.3).

6.1 Metodologia e approccio alla ricerca

La convergenza tra stato dell'arte, analisi fenomenologica e riflessioni metodologiche, espresse nei capitoli 1-4, ha messo in evidenza criticità specifiche nella rappresentazione del nativo digitale d'autore. Da questa analisi sono stati individuati cinque requisiti per la rappresentazione dei materiali in considerazione della loro identità, struttura, provenienza e contesti. Tali requisiti, pur rappresentando solo una parte della complessità del tema, costituiscono un punto di partenza per la modellazione, anche nella prospettiva di integrazione con i modelli descrittivi consolidati, e possono essere definiti come segue:

Requisito 1. La rappresentazione del file nell'ambiente digitale.
Come delineato nel capitolo 4.1, la distinzione tra caratteristiche intrinseche ed estrinseche dei documenti, consolidata nella tradizione paleografica e diplomatica per l'analisi dei documenti cartacei, conserva la sua validità concettuale nel contesto *born-digital* pur richiedendo una ridefinizione che tenga conto delle peculiarità dei documenti digitali. La natura del digitale è intrinsecamente stratificata secondo

la lettura di Thibodeu (2002): ogni oggetto digitale manifesta una stratificazione di livelli di astrazione che comprende il livello fisico (memorizzazione in sequenze di bit codificate su supporti hardware), logico (struttura dei dati nel *file system*) e concettuale (contenuto informativo). Come evidenziato da Trace, «born-digital records are forged in the alliance of the user, the computer hardware, the operating system software, and the application software» (2011, 24), rendendo indispensabile una curatela consapevole dei legami tra materialità, organizzazione logica e significato informativo. L'archivio Evangelisti documenta questa complessità attraverso oltre quaranta tipologie di file diverse distribuite su supporti diversi, ciascuno con configurazioni organizzative differenziate che riflettono diverse funzioni d'uso e vincoli tecnologici.

Il modello deve consentire la rappresentazione esplicita della natura di un file distinguendo tra dimensione fisica, logica e contenutistica, supportando la diversità tipologica senza privilegiare o escludere specifici formati.

Requisito 2. Rappresentazione dell'integrità dei materiali

La descrizione archivistica dei materiali nativi digitali può essere considerata non solo come strumento di accesso e mediazione, ma anche come documentazione per valutare e garantire l'integrità delle collezioni (InterPARES 1 Authenticity Task Force 2002; MacNeil 2005; Forstrom 2009). Nel digitale, l'integrità non può essere data per scontata ma deve essere ciclicamente verificata e documentata attraverso specifiche procedure. La descrizione assume il valore di un'attestazione dello stato dei materiali nel momento in cui viene redatta, e questo principio si traduce nella necessità di integrare nella descrizione, in particolare, il dato dell'hash dei documenti, ovvero il risultato di un'operazione crittografica che genera un'impronta digitale del file. Il valore così ottenuto, se ricalcolato in momenti successivi e confrontato con quello originario, consente di verificare l'integrità del documento: anche la minima alterazione del file, infatti, produrrebbe una modifica dell'hash. In questo senso, l'hash rappresenta un elemento tecnico imprescindibile per attestare e preservare l'autenticità dei materiali digitali nel tempo, e dunque fondamentale è la rappresentazione di questa specifica informativa. Fondamentale è documentare anche l'algoritmo con cui l'hash è stato calcolato, poiché solo l'indicazione esplicita del metodo consente di garantire la ripetibilità della verifica e di preservarne il valore probatorio in prospettiva futura. Algoritmi differenti, infatti, producono output di lunghezza e formato diversi: MD5 genera hash di 128 bit (32 caratteri esadecimali), SHA-1 produce hash di 160 bit (40 caratteri), mentre SHA-256 e SHA-512, appartenenti alla famiglia SHA-2, generano rispettivamente hash di 256 bit (64 caratteri) e 512 bit (128 caratteri) (Dang 2012).

Il modello deve supportare la rappresentazione dei codici hash e dell'algoritmo con cui sono calcolati, come elementi identificativi del documento e della sua integrità nel tempo.

Requisito 3. Rappresentazione dei metadati nativi.

Come analizzato nel capitolo 4.2, i file nativi digitali sono accompagnati fin dalla loro creazione da metadati generati automaticamente, che si accumulano e si aggiornano lungo l'intero ciclo di vita del documento. Questa situazione introduce una complessità inedita rispetto alla tradizione analogica, poiché alla curatela descrittiva si affianca una componente di informazioni prodotte da sistemi, applicazioni e utenti nelle fasi precedenti. Ne risulta un insieme eterogeneo di dati, derivato dall'analisi di caratteristiche differenti dello stesso oggetto digitale.

I metadati nativi comprendono diverse tipologie, tra cui metadati di sistema (o *file system*) e metadati *embedded*, variabili a seconda dell'ambiente tecnologico in cui vengono generati, gestiti e interpretati. Come osserva Langdon, «the profession also needs to analyse what information users require that is not currently captured in descriptive standards», evidenziando la necessità di considerare anche informazioni non ancora standardizzate. La varietà e la granularità dei metadati nativi sono infatti spesso imprevedibili a priori: non è possibile sapere in anticipo quali informazioni veicolino né se possano risultare rilevanti per gli utenti, considerato che le esplorazioni e la ricerca in questo ambito sono ancora agli albori. Di conseguenza, piuttosto che modellare campi predefiniti per i metadati, è opportuno orientarsi verso un approccio flessibile e agnostico rispetto al tipo e alla granularità delle informazioni.

Il modello deve garantire l'integrazione sistematica dei metadati nativi, documentandone valore e provenienza, senza vincoli a priori sui tipi o sulla granularità delle informazioni raccolte.

Requisito 4. Rappresentazione della *provenance*.

Il principio di provenienza, fondamento dell'archivistica tradizionale, assume nel contesto digitale nuove sfaccettature alla luce di una complessa stratificazione di dipendenze tecnologiche (Bak 2024), come illustrato nel capitolo 4.3. La storia del materiale non si limita al creatore originario né all'ordinamento dei file, ma coinvolge l'intera catena di attori e strumenti che hanno contribuito alla sua creazione, gestione e fruizione, comprendendo creatori, archivisti, software, hardware e contesti operativi (Nesmith 2002; Bak 2024). La *provenance* digitale richiede dunque una rappresentazione multidimensionale (Alfier 2017) che includa: la dimensione tecnica, relativa agli strumenti utilizzati per la creazione dei documenti, alle versioni software e alle caratteristiche operative degli ambienti; la dimensione processuale, che documenta tutte le operazioni di trasformazione, estrazione, migrazione e normalizzazione dei materiali e dei dati (Gorini e Giagnolini 2025); e la dimensione curatoriale, che

registra le decisioni e le responsabilità nella gestione dei materiali. Questa prospettiva permette di comprendere pienamente la complessa storia dei documenti digitali, garantendo la tracciabilità e la trasparenza delle operazioni che li hanno prodotti, gestiti e conservati (Tomasi 2022).

Il modello deve supportare la rappresentazione della *provenance* tecnica, processuale e interpretativa, documentando la storia e le trasformazioni dei documenti digitali e dei relativi metadati.

Requisito 5. Rappresentazione dei contesti.

Nel capitolo 4.4 si evidenzia come *provenance* e contesti archivistici rappresentino concetti fondamentali e strettamente interrelati nella disciplina. Il concetto di contesto viene esteso oltre la *provenance*, includendo relazioni e circostanze ampie, come quelle culturali e sociali tipiche del materiale digitale nativo. Nesmith parla di un «pronounced contextual turn», definendolo come un movimento verso una più profonda comprensione del ruolo della conoscenza contestuale dei documenti nel lavoro archivistico (Nesmith 2002). Dilley descrive il contesto archivistico come un'articolazione riguardante un insieme di connessioni e disconnessioni ritenute rilevanti per un determinato soggetto, situato socialmente e storicamente, e per uno scopo specifico (Dilley 2002). In questa prospettiva, i contesti archivistici possono aprirsi ad altri domini, integrando informazioni bibliografiche, riferimenti a risorse esterne e relazioni con dati provenienti da differenti ambiti disciplinari, ampliando la comprensione dei documenti e valorizzandone le potenzialità informative a seconda dei contenuti dell'archivio (Tomasi 2022).

Il modello deve consentire la rappresentazione di multipli contesti, consentendo di collegare i documenti alle molteplici relazioni, verticali e orizzontali, che ne arricchiscono il significato, garantendone una comprensione articolata e contestualizzata.

L'implementazione di questi sei requisiti è finalizzata alla generazione di grafi semantici che consentano di muoversi verso quello che Valacchi (2024) definisce un "inventario aumentato": un ecosistema informativo che espanda i confini tradizionali dell'archivio per amplificare i contesti interpretativi e le possibilità di accesso.

6.2 Il modulo Born-Digital (BoDi)

La ricognizione delle ontologie esistenti ha permesso di delineare un panorama articolato di modelli testimoniano la vitalità della riflessione metodologica sul dominio archivistico. Le esperienze analizzate nel capitolo 5.1 offrono contributi significativi: ArCo rappresenta il più ampio *knowledge graph* italiano per il patrimonio culturale e presenta un'architettura modulare sofisticata, sebbene il modulo archives, ancora in fase evolutiva attraverso l'iniziativa ArCo4Archives, non abbia finora raggiunto una stabilità

implementativa piena e mantenga un orientamento prevalentemente nazionale che potrebbe limitarne l'integrazione nell'ecosistema internazionale.

ArchOnto ha dimostrato concretamente la propria efficacia attraverso l'esperienza portoghese, fornendo evidenze empiriche sulla fattibilità della transizione semantica e sulla scalabilità dei modelli ontologici; tuttavia, la sua fondazione su CIDOC CRM, pur garantendo solidità concettuale e interoperabilità con il dominio museale, lo posiziona a una certa distanza dalle specificità della tradizione descrittiva archivistica.

PREMIS, dal canto suo, costituisce uno standard consolidato e imprescindibile per la gestione dei metadati di preservazione, ma la sua vocazione primaria verso gli aspetti tecnico-conservativi lo rende più adatto a un ruolo di integrazione mirata piuttosto che di *framework* di riferimento complessivo.

È in questo quadro che si colloca la scelta di RiC-O come *framework* di base per la rappresentazione del digitale d'autore. RiC-O permette di rappresentare reti complesse e multidirezionali di relazioni tra entità archivistiche, agenti e contesti, rispecchiando più adeguatamente la natura stratificata e interconnessa della documentazione contemporanea attraverso la capacità di gestire molteplici livelli di aggregazione documentaria e contesti stratificati.

A queste caratteristiche si aggiunge il fatto che RiC-O si sta progressivamente affermando come nuovo standard di riferimento internazionale. Sviluppato dall'ICA attraverso un processo lungo e partecipato, il modello mira a sostituire gli standard esistenti beneficiando di una governance consolidata e di una comunità internazionale sempre più ampia che ne garantisce la manutenzione e l'evoluzione. Questo processo di standardizzazione rappresenta un elemento di continuità rispetto alla tradizione archivistica, pur introducendo un cambio di paradigma nella concezione della descrizione archivistica, che da struttura gerarchica si riconfigura come ecosistema relazionale.

La validità di questo impianto teorico per la rappresentazione del digitale ha trovato conferma nello studio condotto da Haklae Kim (2023) sul caso della crisi finanziaria coreana del 1997, che ha applicato RiC-O per la caratterizzazione di record digitali all'interno di un *knowledge graph*, dimostrando come il modello fornisca una base adeguata a rappresentare le relazioni tra entità archivistiche eterogenee e per esplorare i record nei loro appropriati contesti. La ricerca ha evidenziato la capacità di RiC-O di adattarsi alle caratteristiche del digitale e la sua flessibilità nell'essere integrato con altri vocabolari senza compromettere la coerenza complessiva del modello. Tuttavia, l'approfondimento delle specificità del digitale d'autore e l'individuazione dei requisiti esposti nel capitolo 6.1 hanno evidenziato come RiC-O, pur rappresentando una solida impalcatura concettuale, non risulti autonomamente sufficiente a coprire

l'intero spettro delle esigenze descrittive che caratterizzano questo particolare dominio. Del resto, la stessa documentazione di EGAD riconosce questa necessità, affermando che «RiC is designed, though, with the expectation that it may be extended using cross-domain ontologies specifically addressing technical data needed for the management of instantiations» (International Council on Archives, Expert Group on Archival Description 2025b). Questa apertura programmatica costituisce sia un riconoscimento dei limiti del modello, sia un invito alla sua estensione attraverso l'integrazione con ontologie specializzate.

In risposta a questa esigenza, si è proceduto allo sviluppo di BoDi (Born-Digital Ontology)[211], un'estensione specializzata di RiC-O dedicata alla rappresentazione del digitale d'autore. La scelta di design, laddove possibile, ha privilegiato il riutilizzo di ontologie consolidate, in particolare PREMIS, PROV-O e LRMoo. A queste ontologie si affiancano classi e proprietà sviluppate ad hoc, necessarie per rispondere a esigenze specifiche del digitale d'autore che non trovano copertura nei modelli esistenti. Il modello architetturale adottato prevede che RiC-O mantenga il suo ruolo di *framework* di riferimento, utilizzabile nella sua interezza, mentre PREMIS, PROV-O e LRMoo vengano impiegati selettivamente, limitatamente alle classi e proprietà direttamente importate e rilevanti per il contesto applicativo. Questa strategia permette di mantenere interoperabilità con standard consolidati limitando lo sviluppo di nuovi componenti alle sole specificità del dominio non altrimenti modellabili.

Ciascun requisito individuato al capitolo 6.1 si è tradotto in specifiche scelte di modellazione, che vengono di seguito analizzate articolando per ognuna le classi e le proprietà impiegate.

Modellazione Requisito 1. Rappresentazione delle caratteristiche intrinseche ed estrinseche nel born-digital

La stratificazione di livelli di astrazione propria del digitale (Thibodeau 2002) trova realizzazione in BoDi attraverso una modellazione che distingue livello contenutistico, logico e fisico per la rappresentazione di una risorsa[212].

[211] BoDi è reso disponibile all'IRI http://w3id.org/bodi#, mentre la documentazione tecnica completa, comprensiva di diagrammi, esempi d'uso e linee guida implementative, è consultabile nel *repository* GitHub dedicato al progetto: https://github.com/LuciaGiagnolini12/bodi.

[212] Per la modellazione di questo requisito è stata presa in considerazione anche un'ontologia esterna al dominio archivistico, la NEPOMUK Information Element Ontology (NIE). Cfr. https://www.semanticdesktop.org/ontologies/2007/01/19/nie/ Sebbene NIE condivida con BoDi l'obiettivo di modellare risorse native digitali (in particolare file e metadati tecnici) le differenze negli scopi della rappresentazione rendono NIE inadatto come *framework* per la descrizione archivistica del digitale d'autore. NIE, infatti, è stata progettata per supportare il *Personal Information Management* (PIM) in ambiente desktop, con

Il contenuto è modellato attraverso la classe `rico:RecordResource`. Infatti, la descrizione che RiC-O fornisce di `rico:RecordResourse`, pur facendo eco alla più canonica definizione di archivio (Penzo Doria 2022), sottolinea in particolare il suo carattere informativo: «Information produced or acquired and retained by an Agent in the course of life or work activity».

RiC-O distingue tre specializzazioni di `rico:RecordResource` che BoDi adotta direttamente:

- `rico:Record`, per le singole unità documentarie (ad esempio il contenuto intellettuale espresso in un file DOCX);
- `rico:RecordSet`, per le aggregazioni di risorse, come una cartella o un corpus di testi;
- `rico:RecordPart`, per le componenti interne a un documento, ad esempio un'immagine o una tabella incluse nel file.

RiC-O permette di modellare anche la classica gerarchia archivistica, per cui un `rico:RecorSet`, rappresentante una cartella, è correlato al file che contiene tramite la *object property* `rico:includesOrIncluded` e la relativa inversa `rico:isOrWasIncludedIn`. Le caratteristiche di una `rico:RecordResource`, come data di creazione e di modifica, identificativo o autore, possono essere specificate tramite il riutilizzo di tutte le classi e le proprietà già previste da RiC-O.

Ad esempio, osservando il documento unicamente dal punto di vista contenutistico, ipotizzando un file "Tortuga.docx", che contiene il primo capitolo del romanzo Tortuga (Mondadori, 2009) di Valerio Evangelisti, salvato nella cartella "Romanzi" potrebbe essere descritto come mostrato nel Listato 6.1.

```
# Record che rappresenta il contenuto intellettuale del primo capitolo di
Tortuga
:record_tortuga a rico:Record ;
    rico:hasOrHadTitle :title_tortuga ;
    rico:hasAuthor :author_evangelisti ;
    rico:hasCreationDate :date_2007_03_21 ;
    rico:hasModificationDate :date_2009_01_15 ;
    rico:isOrWasIncludedIn :recordset_romanzi .

# Titolo del record
:title_tortuga a rico:Title ;
    rdfs:label "Tortuga. Capitolo 1" .
```

l'obiettivo di aiutare l'utente nelle sue attività quotidiane attraverso l'integrazione semantica di applicazioni personali e l'estrazione automatica di metadati da fonti eterogenee. BoDi, invece, si colloca nel dominio della descrizione archivistica, dove l'obiettivo non è facilitare la produttività personale, ma documentare contesti di produzione, relazioni archivistiche e stratificazioni interpretative attraverso standard internazionali. Questa divergenza negli obiettivi si riflette nell'architettura dei due modelli: la distinzione binaria proposta da NIE tra DataObject e InformationElement, utile per astrarre l'interpretazione dalla rappresentazione nel contesto PIM, risulta strutturalmente incompatibile con la distinzione tripartita individuata come requisito modellistico: i domini e codomini delle proprietà NIE non permettono di mappare coerentemente i livelli fisico, logico e concettuale richiesti, rendendo problematico anche qualsiasi tentativo di allineamento.

```
# Autore del record
:author_evangelisti a rico:Person ;
    rico:name "Evangelisti, Valerio" .

# Data di creazione del contenuto
:date_2007_03_21 a rico:Date ;
    rico:normalizedDateValue "2007-03-21"^^xsd:date .

# Data di ultima modifica del contenuto
:date_2009_01_15 a rico:Date ;
    rico:normalizedDateValue "2009-01-15"^^xsd:date .

# RecordSet che aggrega il record nella cartella "Romanzi"
:recordset_romanzi a rico:RecordSet ;
    rico:hasOrHadTitle :title_romanzi ;
    rico:includesOrIncluded :record_tortuga .

# Titolo del RecordSet
:title_romanzi a rico:Title ;
    rdfs:label "Romanzi" .
```

Listato 6.1. Modellazione espressa in Turtle dell'istanza "Tortuga.docx" nel contesto del file system.

Accanto al contenuto, ogni documento digitale ha una propria manifestazione, che RiC-O individua come `rico:Instantiation`, classe con cui è possibile rappresentare il file come entità logica distinta dalla `rico:RecordResource` (ma collegata ad essa tramite le *object property* `rico:isOrWasInstantiationOf` e `rico:hasOrHasInstantiation`).

La definizione stessa di `rico:Instantiation` ne esemplifica una possibile adozione del contesto digitale:

> «A Record or Record Part must have been instantiated at least once, though this instantiation may no longer exist at the moment of description. […] A RecordResource may have many Instantiations simultaneously (for instance, a record printed and saved in the same time as DOCX and PDF/A would have 3 concurrent instantiations) or through time (for example, copy of a record). Depending on the context, a new instantiation may be seen as a new or as the same record resource».

BoDi arricchisce `rico:Instantiation` con proprietà che catturano le caratteristiche più tecniche:

- `bodi:hierarchyDepth`, che misura la profondità del file nella gerarchia delle cartelle, dove 0 rappresenta il livello radice della gerarchia e *n* indica la profondità della posizione del file o della cartella rispetto a tale livello radice;
- la posizione relativa di un file o di una cartella, oltre alla gerarchia esprimibile anche fra diverse `rico:Insantiantion`, viene documentata attraverso l'integrazione della *object property*

`prov:atLocation`, la quale collega una `rico:Instantiation` a una `prov:Location`. Questo approccio nasce dalla constatazione che RiC-O non include proprietà o classi specifiche per rappresentare la localizzazione nel senso di *file path*, mentre modello PROV definisce `prov:Location` come «a non-geographic place such as a directory, row, or column», rendendolo adatto al caso. Per facilitare questa integrazione, è stato proposto un allineamento di `rico:Instantiation` come sottoclasse di `prov:Entity`;

- le proprietà `bodi:hasTechnicalDescription` e `bodi:isTechnicalDescriptionOf` consentono di associare una `rico:Instantiation` a un testo descrittivo generale del file, espresso attraverso la classe `bodi:TechnicalDescription`. Anche in questo caso, è possibile definire la tipologia della descrizione tramite `bodi:TechnicalDescriptionType`, specificando se sia stata prodotta manualmente dall'archivista o automaticamente da un LLM;
- come illustrato nella sezione successiva, è inoltre prevista la correlazione di una `rico:Instantiation` con i metadati tecnici del file mediante la *object property* `bodi:hasTechnicalMetadata` e la sua inversa `bodi:isTechnicalMetadataOf`, creando così un legame strutturato tra l'entità logica e i suoi metadati di carattere tecnico.

Riprendendo l'esempio precedente, il file "Tortuga.docx" può essere rappresentato nel contesto del *file system* come da Listato 6.2 (lasciando la modellazione dettagliata dei metadati tecnici alla sezione dedicata).

```
# Istanziazione del file Tortuga.docx con caratteristiche logico-tecniche
:inst_tortuga_docx a rico:Instantiation ;
    rico:instantiates :record_tortuga ;
    bodi:hierarchyDepth 2 ;
    prov:atLocation :loc_tortuga_docx ;

    bodi:hasTechnicalDescription :desc_tortuga_docx ;
    bodi:hasTechnicalMetadata :metadata_tortuga_filesize .

# Localizzazione del file nel file system
:loc_tortuga_docx a prov:Location ;
    rdfs:label "/Desktop/Romanzi/Tortuga.docx" .

# Descrizione tecnica generata automaticamente da IA
:desc_tortuga_docx a bodi:TechnicalDescription ;
    bodi:TechnicalDescriptionType "AIgenerated" ;
    rdfs:comment "Documento DOCX creato con MS Word 2007 il 21 marzo 2007 da
Valerio Evangelisti." .

# Metadato tecnico: dimensione del file in byte
:metadata_tortuga_filesize a bodi:TechnicalMetadata ;
    rdf:value "245760"^^xsd:integer ;
    bodi:hasTechnicalMetadataType :type_filesize .
```

```
# Tipologia del metadato tecnico
:type_filesize a bodi:TechnicalMetadataType ;
    rdfs:label "FileSize" ;
```

Listato 6.2. Modellazione espressa in Turtle dell'istanza "Tortuga.docx" nel contesto del file system.

Infine, ogni documento digitale ha anche una traccia materiale sul supporto di memorizzazione. Questo livello è espresso da BoDi con la classe `bodi:FileClusterPosition`, che descrive l'allocazione fisica dei bit sul supporto. La correlazione con il livello logico `rico:Instantiation` è assicurata dalle proprietà `rico:hasFileClusterPosition` e `rico:isFileClusterPositionOf`[213].

In questo modo, un unico documento come può essere visto contemporaneamente come un contenuto informativo (`rico:RecordResource`), un oggetto logico-informatico (`rico:Instantiation`) e traccia fisica sul supporto (`bodi:FileClusterPosition`) (Figura 6.1).

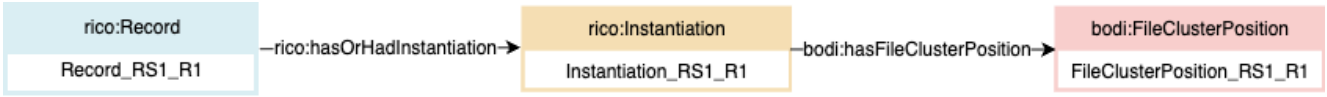


Figura 6.1. Modellazione grafica dei tre livelli di astrazione (contenutistico, logico, fisico).

Modellazione Requisito 2. Rappresentazione dell'integrità dei materiali

La verifica dell'integrità costituisce un requisito fondamentale nella gestione del digitale, permettendo di accertare che i contenuti non abbiano subito alterazioni, intenzionali o accidentali, nel corso del tempo. La modellazione di questo aspetto in BoDi si realizza attraverso l'integrazione della classe `premis:Fixity`, componente centrale del modello PREMIS per la documentazione delle caratteristiche di integrità degli oggetti digitali, che viene correlata alla `rico:Instantiation` sulla base delle *object property* `bodi:hasHashCode` e la relativa inversa `bodi:isHashCodeOf`. `Rico:Instantiation` aventi lo stesso codice hash sono collegate fra loro tramite la *object property* `bodi:HasSameHashCodeAs`.

La scelta di riutilizzare `premis:Fixity` risponde all'esigenza di adottare un vocabolario consolidato e ampiamente riconosciuto nel dominio della preservazione digitale. Un'istanza di `premis:Fixity` rappresenta un valore di *checksum* o hash crittografico calcolato su un oggetto digitale: la scelta di

[213] La presente ricerca non si propone di approfondire la dimensione forense del digitale d'autore, che richiederebbe competenze specifiche e strumenti metodologici non ricompresi nell'ambito dello studio, il cui focus è principalmente l'analisi e la modellazione dei livelli logico e contenutistico. L'introduzione della classe `bodi:FileClusterPosition`, insieme alle *data property* `bodi:startSector` e `bodi:endSector`, intende fornire una rappresentazione minimale dell'aspetto forense, finalizzata a dare la possibilità di una rappresentazione di base per questa dimensione e fornire un punto di partenza per possibili estensioni future.

collegare `premis:Fixity` a `rico:Instantiation` piuttosto che a `rico:RecordResource` merita una riflessione, poiché introduce una tensione concettuale: l'hash viene calcolato sul contenuto del file; quindi, sembrerebbe riferirsi alla dimensione informativa del record. Tuttavia, questo contenuto è quello di una specifica manifestazione, dunque una `rico:Instantiation`. Infatti, due istanziazioni della medesima risorsa archivistica (ad esempio, un documento in formato PDF e lo stesso documento in formato DOCX) condividono lo stesso contenuto intellettuale (`rico:RecordResource`) ma possiedono sequenze di bit differenti e, conseguentemente, hash diversi. La distinzione si applica anche a casi più sottili: due file DOCX con testo identico ma metadati nativi differenti (quali autore, data di ultima modifica, numero di revisione) genererebbero hash completamente diversi, poiché tali metadati sono parte integrante della struttura binaria del file. L'hash cattura quindi l'identità binaria di una specifica istanziazione e, laddove ci siano ricorrenze identiche, BoDi introduce inoltre la proprietà simmetrica `bodi:hasSameHashCodeAs`, che permette di documentare esplicitamente le relazioni tra `rico:Instantiation` che condividono lo stesso valore di hash. Questa proprietà supporta scenari applicativi rilevanti nella gestione archivistica digitale, quali l'identificazione automatica di duplicati, la verifica dell'esito di operazioni di migrazione o copia, e la documentazione di relazioni di derivazione tra istanziazioni. Due istanziazioni con hash identico sono, dal punto di vista del contenuto binario, sostanzialmente equivalenti, indipendentemente dal fatto che risiedano su supporti differenti, abbiano nomi di file diversi o si collochino in contesti organizzativi distinti. Dal punto di vista dell'efficienza dell'estrazione dell'informazione, inoltre, la *property* `bodi:hasSameHashCodeAs` funge da *shortcut property*, consentendo l'identificazione diretta delle `rico:Instantiation` senza la necessità di attraversare il nodo intermedio `premis:Fixity`. La materializzazione esplicita di tale relazione semplifica la formulazione dei pattern SPARQL, eliminando la relazione intermedia e riducendo il numero di operazioni di *join* richieste, ottenendo un incremento dell'efficienza computazionale nelle interrogazioni del grafo.

La generazione di un hash viene modellata come un processo documentabile attraverso una `rico:Activity` di generazione opportunamente documentata, come evidenziato nella sezione dedicata alla modellazione del Requisito 4.

Riprendendo nuovamente l'esempio del file "Tortuga.docx", la rappresentazione dell'hash code di un file può essere quindi modellata come mostrato nel Listato 6.3 e nella Figura 6.2.

```
# Istanziazione del file Tortuga.docx
:inst_tortuga_docx a rico:Instantiation ;
    rico:instantiates :record_tortuga ;
```

```
    bodi:hierarchyDepth 2 ;
    prov:atLocation :loc_tortuga_docx ;
    bodi:hasHashCode :fixity_tortuga_001 .
    bodi:hasSameHashCodeAs :inst_tortuga_docx .

# Istanza premis:Fixity con valore hash e riferimenti alle istanziazioni
:fixity_tortuga_001 a premis:Fixity ;
    rdf:value
"a3d5c8f2e9b1a7c4d6e8f0a2b4c6d8e0f2a4b6c8d0e2f4a6b8c0d2e4f6a8b0c2”;
    bodi:isHashCodeOf :inst_tortuga_docx ;
    bodi:isHashCodeOf :inst_tortuga_backup ;
    bodi:generatedBy :activity_hash_generation_001 .

# Attività di generazione hash
:activity_hash_generation_001 a rico:Activity ;
    rico:title “Generazione hash SHA-256 per Tortuga.docx”;
    bodi:hasGenerated :fixity_tortuga_001 ;
    rico:isOrWasPerformedBy :algorithm_sha256 ;
    rico:occurredAtDate :date_hash_generation .

# Algoritmo utilizzato
:algorithm_sha256 a bodi:Algorithm ;
    rico:name “SHA-256” ;
    rico:technicalCharacteristics “Secure Hash Algorithm 256-bit (SHA-256)” .

# Data di generazione
:date_hash_generation a rico:Date ;
    rico:normalizedDateValue “2025-01-15T10:30:00”^^xsd:dateTime .

# Esempio di duplicato con stesso hash
:inst_tortuga_backup a rico:Instantiation ;
    prov:atLocation :loc_tortuga_backup ;
    bodi:hasHashCode :fixity_tortuga_001 ;
    bodi:hasSameHashCodeAs :inst_tortuga_docx .
```

Listato 6.3. Modellazione espressa in Turtle del concetto di Fixity.

Figura 6.2. Modellazione grafica del concetto di Fixity.

Modellazione Requisito 3. Rappresentazione dei metadati nativi

La varietà e granularità dei metadati sono intrinsecamente imprevedibili, variando in funzione del formato del file, del software di creazione, del sistema operativo e delle successive operazioni di modifica. Sebbene sia tecnicamente possibile definire a priori un insieme finito e predeterminato di campi descrittivi, tale approccio risulterebbe altamente riduttivo del portato informativo complessivo, escludendo a priori metadati potenzialmente rilevanti per specifici contesti di ricerca. Si rende quindi necessario un approccio di modellazione flessibile e agnostico in grado di accogliere dinamicamente qualsiasi tipologia di metadato nativo senza vincoli predefiniti.

BoDi affronta questa problematica proponendo una modellazione BoDi si articola attraverso la classe `bodi:TechnicalMetadata`, collegata alla `rico:Instantiation` mediante le *object property* `bodi:hasTechnicalMetadata` e la relativa inversa `bodi:isTechnicalMetadataOf`, e tipizzata tramite `bodi:TechnicalMetadataType`. Il modello, infatti, opera una separazione semantica tra il valore del metadato e la sua tipologia:

- Il valore del metadato (ad esempio, 25 Kb per la dimensione di un file, o la stringa “Microsoft Word 2007” per il Software responsabile della sua creazione) è rappresentato tramite un’istanza di `bodi:TechnicalMetadata` attraverso la proprietà `rdf:value`;

- La tipologia del metadato (ad esempio, `FileSize`, `Software`, `cp:revision`, `dc:creator`) è rappresentata dalla classe `bodi:TechnicalMetadataType`.

La modellazione prevede che sia i valori dei metadati che le etichette delle tipologie vengano preservati nella forma esatta in cui sono recuperati dallo strumento utilizzato per l'estrazione, senza operazioni di normalizzazione o standardizzazione. Questa scelta progettuale garantisce la fedeltà alla fonte originale e mantiene la specificità terminologica di ciascun tool di estrazione. Ad esempio, lo stesso metadato relativo al creatore di un documento potrebbe essere etichettato come `Author` da ExifTool, e `dc:creator` da Apache Tika: ciascuna variante può essere preservata come distinta istanza di `bodi:TechnicalMetadataType`, permettendo di documentare sia il contenuto informativo sia la provenienza metodologica e operativa del dato, tramite la contemporanea mappatura della `rico:Activity` di estrazione e del `bodi:Software` responsabile, come si approfondirà nella sezione successiva.

Questo approccio presenta un vantaggio in termini di scalabilità e manutenibilità: l'inserimento di nuove tipologie di metadati richiede esclusivamente la creazione di nuove istanze di `bodi:TechnicalMetadataType`, senza alcuna necessità di modificare la struttura delle classi esistenti. Il modello è quindi intrinsecamente estensibile e in grado di accogliere l'evoluzione dei formati digitali e delle relative specifiche metadatali.

BoDi associa i metadati nativi alla classe `rico:Instantiation` piuttosto che a `rico:RecordResource`, riconoscendo che tali informazioni sono estratte da una specifica istanza fisica del documento e ne riflettono le caratteristiche tecniche contingenti.

Per facilitare l'organizzazione e la gestione di insiemi coerenti di metadati, BoDi introduce inoltre la classe `bodi:TechnicalMetadataTypeSet`, che permette di raggruppare diverse istanze di `bodi:TechnicalMetadataType` secondo criteri funzionali o contestuali. Ad esempio, è possibile definire insiemi quali "metadati EXIF" o "metadati di sistema", facilitando sia la documentazione che l'interrogazione del grafo secondo categorie rilevanti per l'archivista.

L'estrazione dei metadati nativi viene inoltre modellata come un processo documentabile attraverso una `rico:Activity` opportunamente caratterizzata, come illustrato nella sezione dedicata alla modellazione del Requisito 4, in modo tale che valori metadatali provenienti da estrazioni diverse, effettuate in momenti differenti o mediante strumenti alternativi (ad esempio, Apache Tika o ExifTool), coesistano nel grafo come istanze separate di `bodi:TechnicalMetadata` e `bodi:TechnicalMetadataType`, ciascuna collegata alla propria attività di estrazione e alla relativa `rico:Instantiation`. Questo

approccio consente di preservare la tracciabilità delle operazioni di estrazione, documentando quale strumento, con quale configurazione e in quale momento specifico, ha prodotto ciascun metadato, supportando così analisi comparative e verifiche di qualità dei dati estratti.

Riprendendo l'esempio del file "Tortuga.docx", il Listato 6.4 e la Figura 6.3 illustrano la modellazione di alcuni metadati nativi estratti dal file, quali il numero di revisioni (`cp:revision`) e le dimensioni (`FileSize`), scelti a titolo esemplificativo tra le centinaia di metadati che un file può potenzialmente contenere. La rappresentazione della *provenance* dell'attività di estrazione, che completa il quadro della modellazione, sarà trattata nella sezione successiva, dedicata alle soluzioni proposte per rispondere al Requisito 4.

```
# Istanziazione del file Tortuga.docx con metadati
:inst_tortuga_docx a rico:Instantiation ;
    rico:isOrWasInstantiationOf :record_tortuga ;
    bodi:hierarchyDepth 2 ;
    prov:atLocation :loc_tortuga_docx ;
    bodi:hasTechnicalMetadata :metadata_tortuga_revision ;
    bodi:hasTechnicalMetadata :metadata_tortuga_filesize .

# Metadato nativo: numero di revisioni
:metadata_tortuga_revision a bodi:TechnicalMetadata ;
    rdf:value "17"^^xsd:integer ;
    bodi:hasTechnicalMetadataType :type_revision ;
    bodi:isTechnicalMetadataOf :inst_tortuga_docx .

# Tipologia del metadato: revisione
:type_revision a bodi:TechnicalMetadataType ;
    rdfs:label "cp:revision" ;

# Metadato nativo: dimensione del file in byte
:metadata_tortuga_filesize a bodi:TechnicalMetadata ;
    rdf:value "286 KB"^^xsd:integer ;
    bodi:hasTechnicalMetadataType :type_filesize ;
    bodi:isTechnicalMetadataOf :inst_tortuga_docx .

# Tipologia del metadato: dimensione file
:type_filesize a bodi:TechnicalMetadataType ;
    rdfs:label "FileSize" ;

# Localizzazione del file nel file system
:loc_tortuga_docx a prov:Location ;
    rdfs:label "/Desktop/Romanzi/Tortuga.docx" .
```

Listato 6.4. Modellazione espressa in Turtle dei metadati nativi cp:revision e FileSize (senza dichiarazione di provenienza).

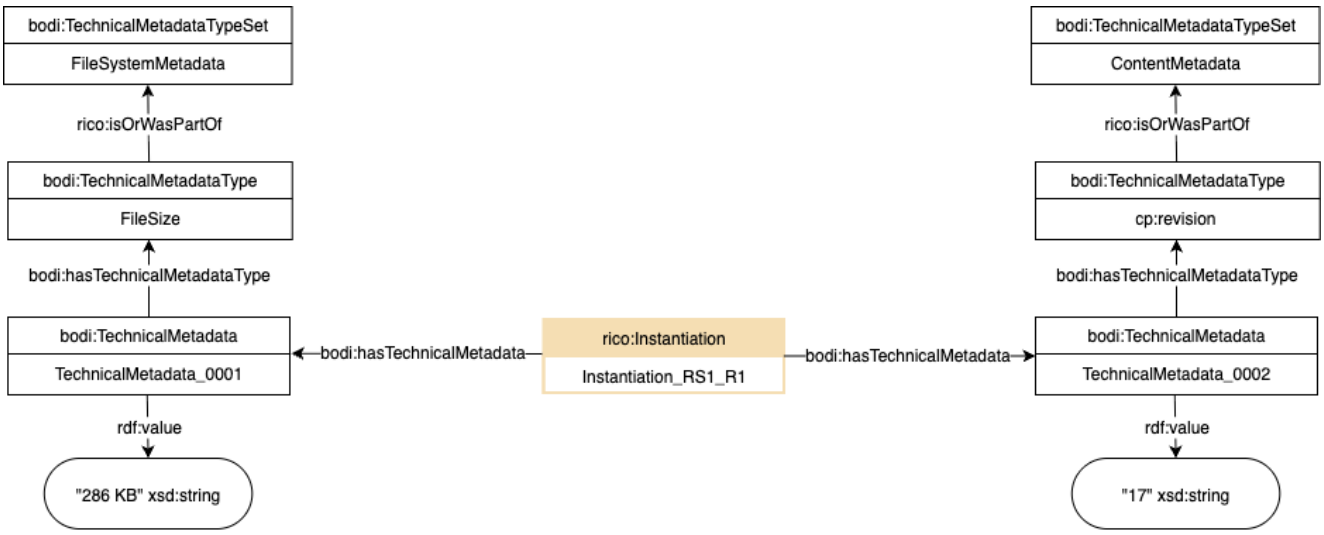


Figura 6.3. Modellazione grafica dei metadati nativi cp:revision e FileSize (senza dichiarazione di provenienza).

Modellazione Requisito 4. Rappresentazione della *provenance*

La provenienza nel contesto digitale richiede un ripensamento che superi la visione monodimensionale tradizionale. Nei flussi documentali digitali intervengono molteplici attori (umani e non) attraverso diversi strati tecnologici, ciascuno dei quali condiziona la ricostruzione della storia dei materiali. Come evidenziato da Bak (2024), la *provenance* digitale è complessa, multipla e ampia, e coinvolge non solo i soggetti produttore ma anche tutti gli agenti e le attività coinvolte nel ciclo di vita dei documenti.

BoDi articola la rappresentazione della provenienza digitale attraverso tre dimensioni principali: (1) la storia conservativa del supporto fisico, (2) l'individuazione degli agenti coinvolti nel ciclo di vita del documento, (3) la documentazione dei processi effettuati sui materiali. La quarta dimensione di *provenance* precedentemente individuata, ossia la documentazione dell'integrità, è già stata affrontata nella modellazione del Requisito 2.

1. **Storia conservativa del supporto fisico.** La documentazione della storia conservativa del supporto fisico è realizzata in BoDi attraverso l'integrazione delle classi PREMIS `premis:StorageMedium` e `premis:StorageLocation`, collegate tra loro mediante la *object property* `premis:medium`. BoDi introduce, inoltre, le proprietà simmetriche `bodi:hasStorageLocation` e `bodi:isStorageLocationOf`, che collegano `rico:Instantiation` a `premis:StorageLocation`, poiché, a causa dei vincoli di dominio e codominio, non è possibile riutilizzare direttamente la proprietà `premis:storedAt`.

   La classe `premis:StorageMedium` rappresenta la natura fisica del supporto di memorizzazione, consentendo di specificare se si tratta, ad esempio, di hard disk, SSD, nastri magnetici o altri tipi di medium. La classe `premis:StorageLocation` documenta invece l'ubicazione fisica o logica

in cui i materiali sono conservati. Questa modellazione si integra con la rappresentazione del livello fisico introdotta nel Requisito 1 tramite la classe `bodi:FileClusterPosition`, completando la descrizione della materialità del digitale.

Per documentare la *chain of custody*, BoDi prevede la possibilità di stabilire relazioni di derivazione tra istanze di `rico:Instantiation`. A tal fine viene utilizzata la classe `rico:DerivationRelation`, che consente di collegare due `rico:Instantiation` in cui una (il *target*) deriva dall'altra (la *source*). Associando a ciascuna `rico:Instantiation` di livello *root-directory* i relativi riferimenti a `premis:StorageLocation` e `premis:StorageMedium`, è possibile documentare i diversi ambienti di conservazione attraversati dai materiali e le modalità di derivazione da un ambiente all'altro, includendo, ove disponibili, anche date, agenti e processi coinvolti.

A titolo di esempio, la *chain of custody* dell'hard disk esterno di Valerio Evangelisti, illustrata nel Listato 6.5 e nella Figura 6.4, mostra il passaggio dal supporto originale custodito dall'associazione, alla copia di sicurezza effettuata e consegnata ad ADLab, fino alla copia derivata da quest'ultima utilizzata per l'elaborazione automatica dei materiali.

```
# Istanziazione originale
:Orig_Inst_RS1_RS2 a rico:Instantiation ;
    bodi:hasStorageLocation :StorageLocation0001 .

# Istanziazione derivata (prima migrazione)
:Der_Inst_RS1_RS2 a rico:Instantiation ;
    bodi:hasStorageLocation :StorageLocation0002 .

# Istanziazione finale (seconda migrazione)
:Inst_RS1_RS2 a rico:Instantiation ;
    bodi:hasStorageLocation :StorageLocation0003 .

# Prima relazione di derivazione
:Derivation0001 a rico:DerivationRelation ;
    rico:relationHasSource :Orig_Inst_RS1_R1 ;
    rico:relationHasTarget :Der_Inst_RS1_R1 .

# Seconda relazione di derivazione
:Derivation0002 a rico:DerivationRelation ;
    rico:relationHasSource :Der_Inst_RS1_R1 ;
    rico:relationHasTarget :Inst_RS1_R1 .

# Storage Location 1 (presso Associazione Valerio Evangelisti)
:StorageLocation0001 a premis:StorageLocation ;
    premis:medium :StorageMedium0001 ;
    rdf:value "Associazione Valerio Evangelisti - Il Sol dell'Avvenire" .

:StorageMedium0001 a premis:StorageMedium ;
    rdfs:label "Western Digital WDBACW0020HBK-01, My Book 2TB USB 3.0 Series
External Hard Drive" .
```

```
# Storage Location 2 (presso ADLab - Armadio 1)
:StorageLocation0002 a premis:StorageLocation ;
    premis:medium :StorageMedium0002 ;
    rdf:value "ADLab - Laboratorio Analogico Digitale, Dipartimento di
Filologia Classica e Italianistica dell'Università di Bologna. Via Zamboni 32,
Bologna, Italia. Armadio 1" .

:StorageMedium0002 a premis:StorageMedium ;
    rdfs:label "hard disk Samsung Portable SSD T7 1 TB USB tipo-C 3.2 Gen 2"
.

# Storage Location 3 (presso ADLab - Sala Server)
:StorageLocation0003 a premis:StorageLocation ;
    premis:medium :StorageMedium0003 ;
    rdf:value "ADLab - Laboratorio Analogico Digitale, Dipartimento di
Filologia Classica e Italianistica dell'Università di Bologna. Via Zamboni 32,
Bologna, Italia. Sala Server" .

:StorageMedium0003 a premis:StorageMedium ;
    rdfs:label "Network Attached Storage Ext4" .
```

Listato 6.5. Modellazione espressa in Turtle della chain of custody attraverso relazioni di derivazione, storage locations e storage media.

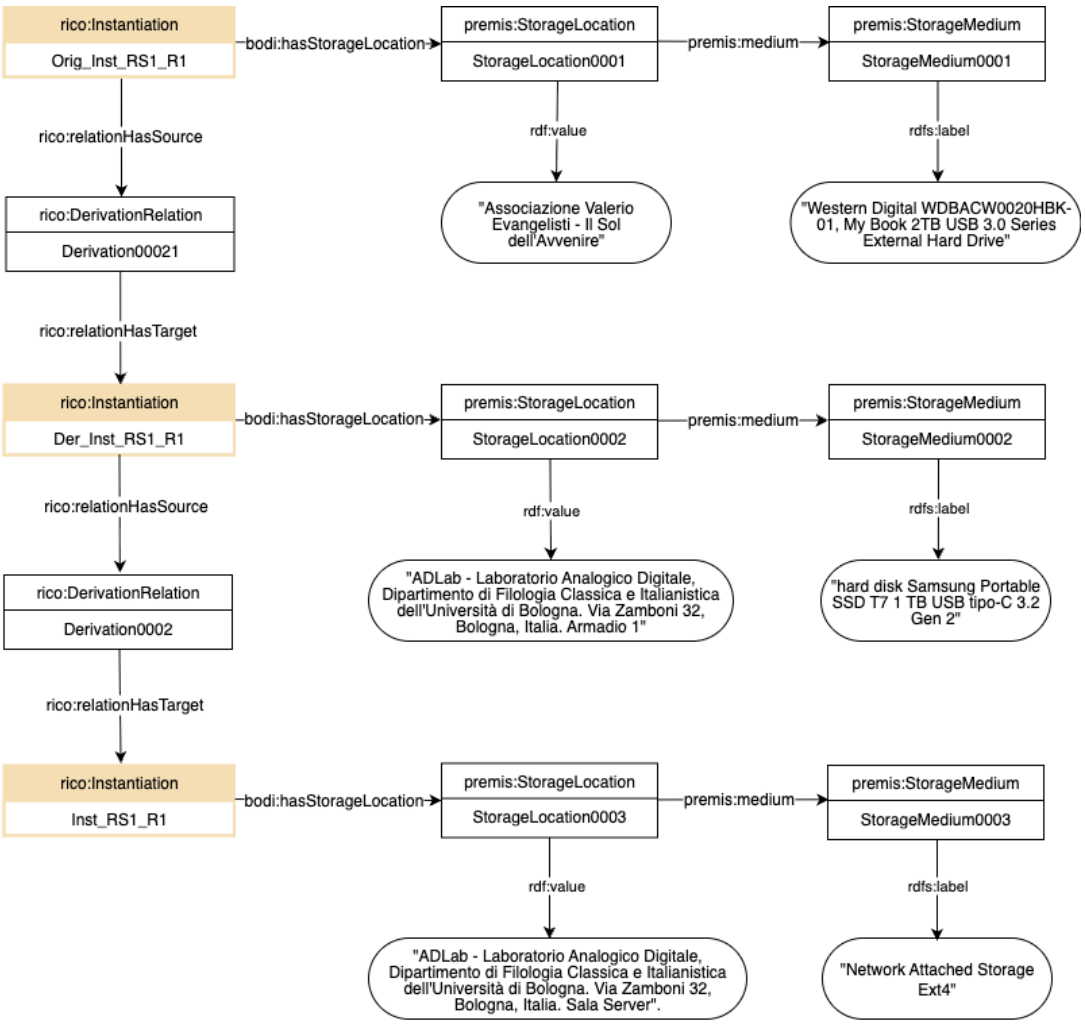


Figura 6.4. Modellazione grafica della chain of custody attraverso relazioni di derivazione, storage locations e storage media.

2. **Individuazione degli agenti coinvolti nel ciclo di vita del documento.** Nel contesto digitale, il principio di provenienza si estende oltre la sola autorialità umana, includendo le forme di agentività esercitate da sistemi automatizzati e da processi ibridi uomo-macchina. Anche la pratica archivistica, in questo contesto, si articola in un continuum di responsabilità che comprende interventi umani diretti, attività supervisionate e operazioni completamente algoritmiche. Riconoscere questa pluralità di agenti è fondamentale per rappresentare in modo completo la storia dei materiali digitali. Muovendo da questa prospettiva, BoDi estende il principio di provenienza includendo l'agentività algoritmica accanto a quella umana, articolandola in tre livelli di responsabilità.

   Per l'agentività umana diretta, RiC-O offre già un sistema articolato, a partire dalla classe `rico:Person`. La superproprietà `rico:hasOrganicOrFunctionalProvenance` collega `rico:RecordResource` o `rico:Instantiation` a un `rico:Agent` che la crea, accumula, riceve

o invia, oppure a una `rico:Activity` che la genera. Le sue sotto-proprietà specificano i diversi ruoli tradizionali nella definizione della *provenance* archivistica: `rico:hasAuthor`, `rico:hasCreator`, `rico:hasAccumulator`, `rico:hasSender`, `rico:hasReceiver`, `rico:hasAddressee`, `rico:hasCollector`. La sottoproprietà `rico:documents`, invece, collega una `rico:RecordResource` all'attività che documenta, piuttosto che all'agente che lo ha prodotto.

In questo quadro, BoDi introduce `bodi:generatedBy` e la sua inversa `bodi:hasGenerated` (sottoproprietà di `rico:isRelatedTo`) per rispondere a esigenze specifiche della *provenance* digitali. Mentre le subproperties di `rico:hasOrganicOrFunctionalProvenance` hanno dominio limitato a `rico:RecordResource` e `rico:Instantiation`, `bodi:generatedBy` ha dominio `owl:Thing`. Questo permette di documentare la generazione di entità non necessariamente archivistiche ma che richiedono tracciabilità, come gli output di un'estrazione di metadati, la generazione di codici hash, descrizioni e annotazioni curatoriali.

In sintesi: mentre le proprietà RiC-O operano nel paradigma della descrizione archivistica tradizionale, `bodi:generatedBy` opera nel paradigma della *provenance* computazionale, documentando chi ha materialmente prodotto specifici output in ecosistemi generativi complessi dove flussi ibridi umano-algoritmici producono continuamente artefatti informativi che richiedono documentazione.

Per l'agentività algoritmica, è importante notare che `rico:Mechanism` è definito in RiC-O come sottoclasse `di rico:Agent`. Di conseguenza, le proprietà con range `rico:Agent` possono tecnicamente includere anche `rico:Mechanism`. Ma l'ecosistema digitale presenta ulteriori sfumature e richiede un'articolazione più complessa del modello, per cui BoDi introduce tre sottoclassi di `rico:Mechanism`: `bodi:Software`, `bodi:Algorithm` e `bodi:SoftwareComponent`.

La classe `bodi:Software`, sottoclasse sia di `premis:SoftwareAgent` che di `rico:Mechanism`, rappresenta programmi o applicazioni caratterizzati da funzionalità specifiche. La classe `bodi:Algorithm`, rappresenta procedure computazionali, formule o insiemi di regole progettati per svolgere compiti specifici o risolvere particolari problemi come, ad esempio, gli algoritmi crittografici utilizzati per il calcolo degli hash, come SHA-256. La classe `bodi:SoftwareComponent` rappresenta parti modulari e riutilizzabili di sistemi software che forniscono funzionalità specifiche e possono essere combinate con altri componenti per creare applicazioni più complesse come, ad esempio i parser specializzati, che analizzano la struttura dei

dati seguendo regole definite per estrarne informazioni. I `bodi:SoftwareComponent` possono essere ricondotti ai `bodi:Software` di riferimento tramite le *object property* `bodi:hasSoftwareComponent` e `bodi:isSoftwareComponentOf`.

BoDi arricchisce ulteriormente questo *framework* attraverso l'introduzione della classe `bodi:SoftwareStack`, sottoclasse di `rico:Agent` ma disgiunta da `rico:Mechanism`, che modella insiemi di rico:Mechanism che eseguono collettivamente `una rico:Activity`. Questa classe rappresenta aggregazioni composite responsabili di un'operazione, distinguendosi dai singoli meccanismi che la costituiscono e permettendo di documentare la natura sistemica e stratificata degli ambienti computazionali contemporanei. Un `bodi:SoftwareStack` è composto da elementi collegati tramite le object pr*operty* `bodi:hasStackElement` e la sua inversa `bodi:isStackElementOf`, entrambe sottoproprietà rispettivamente di `rico:hasOrHadPart` e `rico:isOrWasPartOf`.

Inoltre, la *data property* `bodi:hasDocumentation`, con dominio `rico:Thing` e range `xsd:anyURI`, fornisce riferimenti URI a specifiche tecniche, manuali o documentazione ufficiale che descrivono il funzionamento e le caratteristiche degli strumenti software. Questa proprietà costituisce un ponte essenziale tra l'istanziazione ontologica e la conoscenza tecnica esterna, permettendo a curatori, ricercatori e sistemi automatici di accedere alle informazioni necessarie per comprendere, valutare e eventualmente riprodurre i processi di generazione documentati.

Per l'agentività ibrida, intesa come forma di collaborazione uomo-macchina, BoDi introduce meccanismi specifici per documentare la governance umana sui processi digitali. Le *object property* `bodi:isOrWasSupervisorOf`, sottoproprietà di `rico:hasOrHadAuthorityOver`, e la sua inversa `bodi:hasOrHadSupervisor`, sottoproprietà di `rico:isOrWasUnderAuthorityOf`, permettono di collegare un `rico:Agent` (tipicamente un `rico:Person`) a una `rico:Activity` elaborata da un `rico:Mechanism` o da un `bodi:SoftwareStack`, della quale l'agente umano detiene responsabilità di supervisione. Queste proprietà documentano il controllo strategico e la governance dei workflow computazionali, rendendo esplicita la catena di responsabilità che attraversa processi automatizzati. La supervisione rappresenta una forma di agentività che si esercita non necessariamente attraverso l'intervento diretto sui contenuti, ma mediante la definizione di parametri operativi, validazione di risultati e gestione delle eccezioni, configurando una modalità di controllo umano che opera a livello di orchestrazione dei processi piuttosto che di produzione diretta degli output.

Per contenuti generati da IA e successivamente revisionati, le *object property* `bodi:hasRevisioned` e `bodi:revisedBy` collegano esplicitamente un `rico:Agent` a una `bodi:TechnicalDescription` prodotta automaticamente e validato dall'intervento umano. BoDi prevede due *data property* che documentano la qualità e l'affidabilità dell'output generato dall'IA `bodi:confidenceScore`, che esprime un valore numerico tra 0 e 10 che indica il livello di affidabilità del contenuto generato, dove valori più alti corrispondono a maggiore confidenza nell'accuratezza dell'output. Questo punteggio, spesso prodotto automaticamente dagli stessi sistemi di AI come metadato di processo, supporta workflow di valutazione qualitativa e validazione, permettendo, ad esempio, di stratificare le risorse generate secondo soglie di affidabilità e di prioritizzare gli interventi di controllo umano su output caratterizzati da bassa confidenza. La *data property* `bodi:hasHumanValidation` indica invece se il testo AI-generated è stato revisionato, verificato o validato da esperti umani, documentando l'intervento di controllo e eventuale integrazione. Questa proprietà può registrare un valore booleano (sì/no), ma può accogliere anche informazioni sul tipo di validazione eseguita, il livello di approfondimento della verifica o eventuali note sui cambiamenti apportati.

L'esigenza di documentare i contesti tecnologici era già stata colta da altre ontologie, in particolare da PREMIS attraverso la classe `premis:Environment`. BoDi specializza il concetto di `premis:Environment` fornendo maggiore granularità nella rappresentazione: mentre PREMIS offre descrizioni testuali prevalentemente statiche dell'ambiente operativo, BoDi modella cosa è stato utilizzato, come e per quali operazioni specifiche. Le rappresentazioni risultanti sono modulari e computazionalmente interrogabili, supportando query SPARQL complesse sulla provenienza tecnologica delle risorse e permettendo analisi dettagliate delle dipendenze tecniche.

In sintesi, l'articolazione di BoDi rispetto all'agentività rappresenta un'estensione significativa del *framework* RiC-O. Mentre RiC-O documenta eccellentemente gli agenti coinvolti nella pratica archivistica più tradizionale, BoDi introduce la granularità necessaria per tracciare la stratificazione di responsabilità che caratterizza i processi digitali contemporanei: dalla generazione algoritmica alla supervisione umana, dalla revisione esperta alla validazione qualitativa, ogni forma di agentività trova rappresentazione esplicita attraverso proprietà semanticamente distinte.

3. **Documentazione dei processi effettuati sui materiali.** Ogni processo determinante effettuato sui materiali digitali viene modellato come `rico:Activity` che documenta la storia esecutiva delle operazioni. Come illustrato nella modellazione del Requisito 2, il calcolo degli hash è

rappresentato come `rico:Activity` che collega l'istanza `premis:Fixity` generata, l'algoritmo impiegato (ad esempio SHA-256), e la data di esecuzione. Analogamente, l'estrazione dei metadati nativi, già introdotta nel Requisito 3, viene documentata come `rico:Activity` che specifica quale strumento software o insieme di software, con quale configurazione e in quale momento specifico, ha prodotto l'estrazione di ciascun metadato. Valori metadatali provenienti da estrazioni diverse, effettuate in momenti differenti o mediante strumenti alternativi (ad esempio, Apache Tika o ExifTool), coesistono nel grafo come istanze separate di `bodi:TechnicalMetadata` e `bodi:TechnicalMetadataType`, ciascuna collegata alla propria attività di estrazione e alla relativa `rico:Instantiation`. Questo approccio consente di preservare la tracciabilità completa delle operazioni di estrazione, documentando quale strumento, con quale configurazione e in quale momento specifico, ha prodotto ciascun valore, supportando così analisi comparative e verifiche di qualità dei dati estratti.

Per processi più semplici o consolidati, la proprietà `bodi:generatedBy` può offrire anche uno *shortcut* diretto che bypassa la modellazione esplicita dell'Activity, semplificando la rappresentazione quando il livello di dettaglio processuale non è ritenuto necessario.

BoDi introduce inoltre la classe `bodi:Exception`, sottoclasse di `rico:Concept`, per rappresentare condizioni di errore, eventi imprevisti o situazioni anomale che si verificano durante l'esecuzione dei processi. La *data property* `bodi:exceptionMessage`, con dominio `bodi:Exception` e range `rdfs:Literal`, contiene il messaggio di errore o la descrizione associata all'eccezione, informazione essenziale per il logging, il debugging e la comprensione dei fallimenti nei workflow automatizzati.

La *data property* `bodi:redactedInformation`, con dominio `rico:RecordResource` e range `rdfs:Literal`, è infine implementata per segnalare gli interventi di anonimizzazione dei dati finalizzata alla pubblicazione. Questa proprietà documenta le modifiche effettuate per ragioni di privacy o sicurezza, rendendo trasparenti le scelte curatoriali che hanno comportato l'omissione o la redazione di informazioni sensibili.

Il Listato 6.6 e la Figura 6.5 illustrano, a fini esemplificativi, la modellazione della *provenance* per il file "Tortuga.docx" che integra l'estrazione dei metadati con Apache Tika e supervisione dell'archivistica.

```
# Istanziazione del file Tortuga.docx con informazioni complete di provenance
:inst_tortuga_docx a rico:Instantiation ;
    rico:instantiates :record_tortuga ;
    bodi:hierarchyDepth 2 ;
    prov:atLocation :loc_tortuga_docx ;
```

```
    bodi:hasTechnicalMetadata :metadata_tortuga_revision ;
    bodi:hasTechnicalMetadata :metadata_tortuga_filesize ;
    bodi:hasHashCode :fixity_tortuga_001 ;
    bodi:hasStorageLocation :storage_location_hdd_001 ;
    bodi:hasTechnicalDescription :techdesc_tortuga_ai .

# Record con informazione sull'anonimizzazione
:record_tortuga a rico:Record ;
    rico:hasOrHadTitle :title_tortuga ;
    rico:hasAuthor :author_evangelisti ;
    bodi:redactedInformation "false" .

# Metadato nativo: numero di revisioni estratto da Apache Tika
:metadata_tortuga_revision a bodi:TechnicalMetadata ;
    rdf:value "17"^^xsd:integer ;
    bodi:hasTechnicalMetadataType :type_revision ;
    bodi:isTechnicalMetadataOf :inst_tortuga_docx ;
    bodi:generatedBy :activity_metadata_extraction_tika_001 .

:type_revision a bodi:TechnicalMetadataType ;
    rdfs:label "cp:revision" .

# Metadato nativo: dimensione del file
:metadata_tortuga_filesize a bodi:TechnicalMetadata ;
    rdf:value "286720"^^xsd:integer ;
    bodi:hasTechnicalMetadataType :type_filesize ;
    bodi:isTechnicalMetadataOf :inst_tortuga_docx ;
    bodi:generatedBy :activity_metadata_extraction_tika_001 .

:type_filesize a bodi:TechnicalMetadataType ;
    rdfs:label "FileSize" .

# Attività di estrazione metadati con Apache Tika
:activity_metadata_extraction_tika_001 a rico:Activity ;
    rico:hasOrHadTitle "Estrazione metadati nativi con Apache Tika" ;
    bodi:hasGenerated :metadata_tortuga_revision ;
    bodi:hasGenerated :metadata_tortuga_filesize ;
    rico:isOrWasPerformedBy :softwarestack_tika_extraction ;
    bodi:hasOrHadSupervisor :archivist_rossi ;
    rico:occurredAtDate :date_extraction_2025_01_15 .

# Software Stack per l'estrazione con Apache Tika
:softwarestack_tika_extraction a bodi:SoftwareStack ;
    rdfs:label "Apache Tika Extraction Environment" ;
    bodi:hasStackElement :software_apache_tika ;
    bodi:hasStackElement :software_python ;
    bodi:hasStackElement :component_poi_parser .

:software_apache_tika a bodi:Software ;
    rico:name "Apache Tika" ;
    rico:hasOrHadIdentifier "2.9.1" ;
    bodi:hasDocumentation <https://tika.apache.org/2.9.1/> .

:software_python a bodi:Software ;
    rico:name "Python" ;
    rico:hasOrHadIdentifier "3.11.5" .

:component_poi_parser a bodi:SoftwareComponent ;
    rico:name "Apache POI Parser" ;
```

```
    rdfs:comment "Parser specializzato per documenti Microsoft Office" ;
    bodi:hasDocumentation <https://poi.apache.org/> .

:archivist_giagnolini a rico:Person ;
    rico:name "Giagnolini, Lucia" ;
    bodi:isOrWasSupervisorOf :activity_metadata_extraction_tika_001 ;
    bodi:hasRevisioned :techdesc_tortuga_ai .

:date_extraction_2025_01_15 a rico:Date ;
    rico:normalizedDateValue "2025-01-15T14:30:00"^^xsd:dateTime .
```

Listato 6.6. Modellazione espressa in Turtle della provenance in relazione all'estrazione dei metadati.

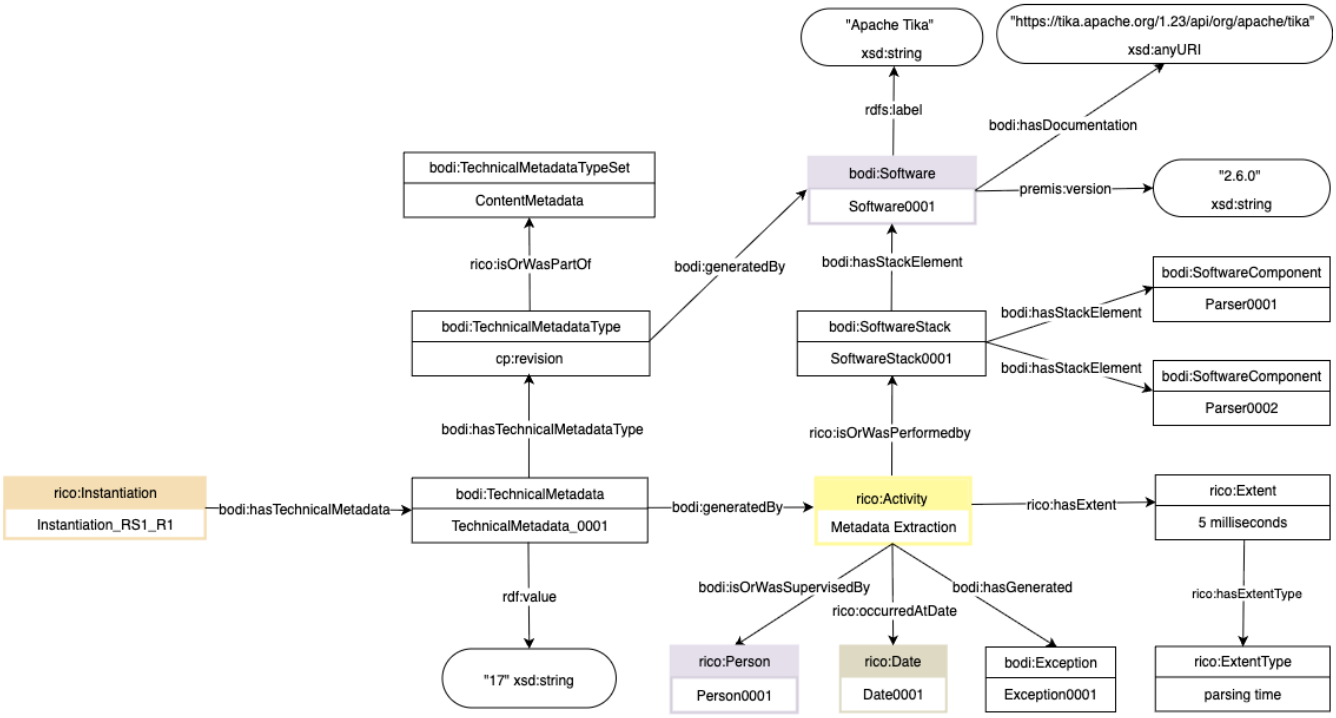


Figura 6.5. Modellazione grafica della provenance in relazione all'estrazione dei metadati.

Modellazione Requisito 5. Rappresentazione dei contesti

Il concetto di contesto archivistico, come evidenziato nel capitolo 4.4, si articola in molteplici dimensioni. In particolare, quella verticale, legata alla tradizionale organizzazione gerarchica dei materiali, e quella orizzontale, che si estende alle molteplici relazioni, circostanze e risorse che arricchiscono il significato dei documenti.

Per la dimensione verticale, BoDi adotta integralmente il *framework* offerto da RiC-O, che rappresenta la struttura gerarchica tra documenti e aggregazioni documentarie attraverso le classi `rico:Record` (documento) e `rico:RecordSet` (aggregazione di documenti), collegate dalle *object property*

`rico:includesOrIncluded` e `rico:isOrWasIncludedIn`, che esprimono la relazione di inclusione verticale tra i livelli della descrizione archivistica.

La scelta progettuale di RiC-O di non prescrivere un numero fisso di livelli descrittivi o una nomenclatura predefinita (fondo, serie, sottoserie, fascicolo, etc.) si rivela particolarmente opportuna nel contesto del digitale d'autore. I materiali digitali, infatti, presentano spesso organizzazioni che non si conformano alle categorizzazioni tradizionali: una struttura di cartelle può riflettere logiche di lavoro personali, aggregazioni tematiche o criteri ibridi che mescolano molteplici dimensioni organizzative e profondità. L'agnosticismo di RiC-O rispetto ai livelli descrittivi canonici consente di rappresentare queste articolazioni senza forzature interpretative, documentando la gerarchia così come si manifesta nei materiali piuttosto che normalizzarla secondo schemi precostituiti.

Questa flessibilità non preclude, comunque, la possibilità di tipizzare le aggregazioni quando necessario. La classe `rico:RecordSetType` permette di qualificare la natura di un `rico:RecordSet`, specificando, ad esempio, se si tratti di un fondo, di una serie, di un dossier tematico o di qualsiasi altra forma di aggregazione rilevante per il contesto specifico. Questa tipizzazione può essere applicata selettivamente, solo dove apporti effettivo valore descrittivo, evitando la rigidità di sistemi che impongono categorizzazioni universali e poco flessibili. Questa architettura si presta a rappresentare simultaneamente molteplici dimensioni organizzative. La classe `rico:RecordSet` rappresenta un'entità aggregatrice flessibile, in grado di accogliere e descrivere insiemi di materiali secondo criteri diversi da quelli canonici. Nel contesto del digitale d'autore, per esempio oltre a rispecchiare la struttura originaria del *file system*, in cui cartelle e sottocartelle corrispondono a `rico:RecordSet` annidati, questa classe consente di costruire aggregazioni alternative basate su criteri tematici, tipologici, cronologici e curatoriali, anche in derivazione dall'interrogazione del *knowledge graph*. In questo modo, `rico:RecordResource` diviene un nodo di convergenza che permette di rappresentare la pluralità dei punti di vista possibili sui materiali digitali, senza vincolarsi a un'unica struttura gerarchica o descrittiva. Questa molteplicità non genera conflitti o ridondanze problematiche: ogni `rico:RecordSet` rappresenta una prospettiva interpretativa legittima sui materiali, e le diverse aggregazioni possono coesistere nel grafo arricchendone la navigabilità.

Se la dimensione verticale documenta la stratificazione gerarchica dei materiali, la dimensione orizzontale apre verso molteplici possibilità, che RiC-O in gran parte già sottende grande alla granularità e alla varietà delle classi e delle proprietà che propone. Nel contesto degli archivi letterari, questa dimensione assume particolare rilevanza attraverso l'integrazione con il dominio bibliografico, realizzata in BoDi mediante l'adozione di LRMoo (IFLA Library Reference Model - Object-Oriented). La classe

centrale di questa integrazione è `lrmoo:F1_Work`, che rappresenta un'opera nel senso di creazione intellettuale o artistica distinta dalle sue manifestazioni materiali. Nel contesto degli archivi d'autore, la classe `lrmoo:F1_Work` consente di rappresentare l'opera letteraria come entità concettuale (ad esempio, il romanzo *Tortuga* inteso come creazione intellettuale), distinguendola dalle sue diverse manifestazioni materiali, come manoscritti, bozze, edizioni e traduzioni[214]. Un'istanza di `lrmoo:F1_Work` può essere collegata a una `rico:RecordResource` attraverso la *object property* `rico:isOrWasSubjectOf` o, in alternativa, mediante la più generale `rico:isRelatedTo`, rendendo così esplicito il legame tra l'opera e i materiali che la documentano.

Questa integrazione abilita percorsi di navigazione trasversale che amplificano considerevolmente i contesti interpretativi. Un ricercatore interessato all'opera "Tortuga" può identificare nel grafo non solo il romanzo pubblicato, ma l'intero ecosistema documentale che lo circonda: manoscritti preparatori, corrispondenze editoriali, note di lavoro, versioni intermedie, traduzioni, recensioni.

LRMoo permette inoltre di modellare ulteriormente questo aspetto tramite la classe `lrmoo:F2_Expression`, che rappresenta la realizzazione intellettuale o artistica di un'opera in una forma specifica (ad esempio, un'edizione, una traduzione, un adattamento, una revisione). Attraverso queste classi è possibile documentare le complesse genealogie delle opere letterarie, le loro derivazioni, trasformazioni e rielaborazioni.

Il Listato 6.7 e la Figura 6.6 illustra l'integrazione tra dominio archivistico e bibliografico, mostrando come il file "Tortuga.docx" si colleghi all'opera letteraria "Tortuga", mostrando, al contempo la relazione gerarchica fra il file e la cartella.

```
# RecordSet: cartella "Romanzi"
:recordset_romanzi a rico:RecordSet ;
    rico:hasOrHadTitle :title_romanzi ;
    rico:includesOrIncluded :record_tortuga ;
    rico:hasOrHadInstantiation :inst_romanzi .

:title_romanzi a rico:Title ;
    rdfs:label "Romanzi" .

# Record: contenuto intellettuale di Tortuga
:record_tortuga a rico:Record ;
    rico:hasOrHadTitle :title_tortuga ;
    rico:isOrWasIncludedIn :recordset_romanzi ;
    rico:hasOrHadInstantiation :inst_tortuga_docx ;
    rico:isRelatedTo :work_tortuga .
```

[214] Una prima sperimentazione basata sul precedente modello IFLA Functional Requirements for Bibliographic Records (FRBR) è stata effettuata nell'ambito del progetto Pavia Archivi Digitali (Carbé 2023, 92).

```
:title_tortuga a rico:Title ;
    rdfs:label "Tortuga" .

# Opera letteraria: Tortuga
:work_tortuga a lrmoo:F1_Work ;
    rdfs:label "Tortuga" ;
    rico:isRelatedTo :record_tortuga .

# Instantiation: directory "Romanzi" nel file system
:inst_romanzi a rico:Instantiation ;
    rdfs:label "Instantiation_Romanzi" ;
    rico:instantiates :recordset_romanzi ;
    rico:hasOrHadPart :inst_tortuga_docx .

# Instantiation: file "Tortuga.docx"
:inst_tortuga_docx a rico:Instantiation ;
    rdfs:label "Instantiation_Tortuga.docx" ;
    rico:instantiates :record_tortuga ;
    rico:isOrWasPartOf :inst_romanzi .
```

Listato 6.7. Modellazione espressa in Turtle dell'integrazione tra contesto archivistico verticale (gerarchia file-cartella) e contesto bibliografico (collegamento all'opera letteraria Tortuga).

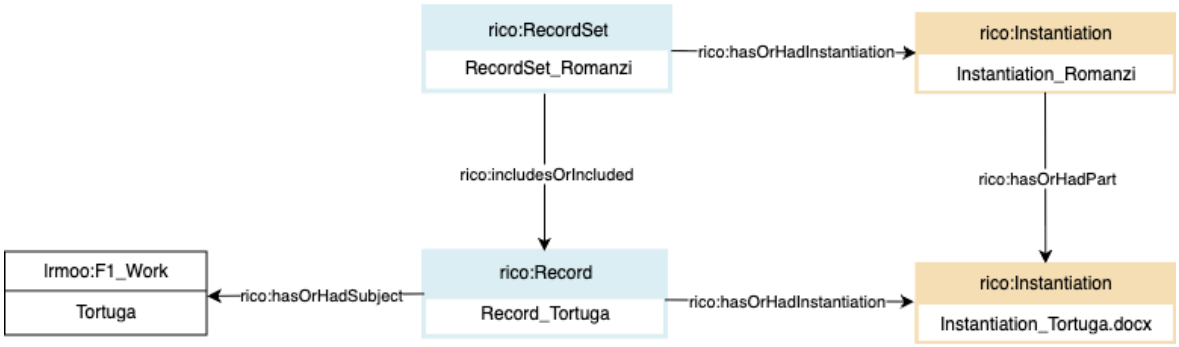


Figura 6.6. Modellazione grafica dell'integrazione tra contesto archivistico verticale (gerarchia file-cartella) e contesto bibliografico (collegamento all'opera letteraria Tortuga).

La combinazione della modellazione dei cinque requisiti dà origine al modello mostrato nella Figura 6.7, che illustra lo schema gerarchico delle classi introdotte da BoDi (in giallo), di quelle estese da RiC-O (in blu), e delle classi provenienti da PREMIS (in verde), PROV-O (in viola) e LRMoo (in rosso). Per ragioni di chiarezza, sono rappresentate soltanto le classi di RiC-O estese da BoDi; l'ontologia RiC-O è tuttavia riutilizzabile nella sua interezza.

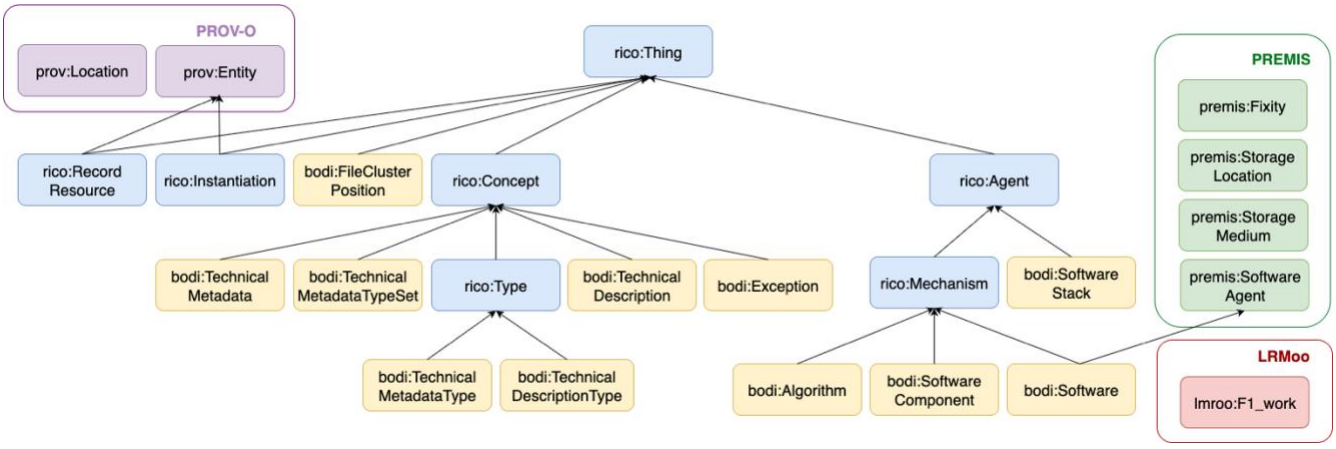


Figura 6.7. Schema delle classi introdotte da BoDi e delle ontologie di riferimento RiC-O, PREMIS, PROV-O e LRMoo. Per chiarezza, sono mostrate solo le classi di RiC-O estese da BoDi.

La navigazione dell'ecosistema informativo così rappresentato può articolarsi secondo molteplici dimensioni simultanee. BoDi abilita percorsi di ricerca che permettono di rappresentare il digitale d'autore secondo logiche:

- gerarchiche: dalla radice del *file system* verso i singoli file e viceversa (attraverso le relazioni gerarchiche fra `rico:Record` e `rico:RecordSet` e le relative `rico:Instantiation`);
- tematiche: dall'opera letteraria a tutti i materiali a essa correlati e viceversa (attraverso le relazioni fra `rico:Record`, `rico:RecordSet` e `lrmoo:F1_work`);
- cronologiche: da una o più date a tutte le entità prodotte o modificate in quel lasso di tempo (attraverso le relazioni fra rico:Date e le varie entità provviste di datazione, come `rico:Record`, `rico:RecordSet`, ma anche `rico:Activity`, `rico:Event`, eccetera);
- tipologiche: da un metadato (formati inclusi) a tutte le risorse che caratterizza e viceversa (attraverso le relazioni tra `bodi:TechnicalMetadata`, `bodi:TecnicalMetadataType` e relative `rico:Instantiation`);
- relazionali: da un agente (`rico:Agent`) a tutte le entità con cui ha interagito e viceversa;
- processuali: da un'attività (`rico:Activity`) a tutte i documenti e le entità coinvolte e viceversa;
- contestuali e trasversali, tramite combinazioni di criteri multipli.

6.3 Validazione del modello

La validità formale dell'ontologia BoDi è stata verificata attraverso diversi livelli di controllo: la validazione sintattica mediante un validatore RDF, la verifica semantica tramite *reasoners* di Protégé,

l'analisi di conformità ontologica con OOPS! (OntOlogy Pitfall Scanner) e infine la sperimentazione su un caso d'uso concreto, dedicato a Valerio Evangelisti, esemplificato nei capitoli 7, 8 e 9.

Il primo livello di controllo, di carattere sintattico, è stato effettuato mediante la validazione RDF. L'ontologia, esportata da Protégé nel formato RDF/XML, è stata sottoposta alla verifica tramite il servizio di validazione RDF[215] fornito dal World Wide Web Consortium (W3C), che utilizza Another RDF Parser (ARP, versione 2-alpha-1). La validazione ha restituito esito positivo, attestando la correttezza sintattica e la coerenza strutturale del file RDF. Si tratta tuttavia di una verifica di base, volta unicamente a garantire la conformità formale del documento ai requisiti dello standard RDF, senza implicazioni sulla coerenza semantica del modello.

Il secondo livello di controllo si è concretizzato tramite l'utilizzo dei *reasoner* messi a disposizione in Protégé, articolato su due verifiche complementari che permettono una validazione completa sia della struttura concettuale che della sua applicabilità operativa. La validazione è stata condotta utilizzando il *reasoner* HermiT (versione 1.4.3.456), basato sul "hypertableau calculus" (Glimm et al. 2014) che fornisce un'implementazione completa del ragionamento in OWL 2 Description Logic (DL). Il primo livello di test, condotto sulla Terminological Box (TBox) dell'ontologia (Baader et al. 2003, 300) composta dalle classi importate da RiC-O, PREMIS, PROV-O e LRMoo e dalle nuove definizioni introdotte in BoDi, ha permesso di verificare la coerenza concettuale del modello attraverso il controllo dell'assenza di classi *unsatisfiable*, ovvero logicamente contraddittorie, della corretta gerarchia e delle relazioni di sottoclasse, nonché del rispetto dei vincoli di disgiunzione tra classi. HermiT ha completato l'analisi della TBox in 81 secondi, pre-computando sistematicamente la gerarchia delle classi e la gerarchia delle *object property* e *data property*, senza rilevare alcuna inconsistenza logica. Questa validazione garantisce che la struttura formale dell'ontologia sia internamente coerente prima ancora di essere popolata con dati concreti, confermando che gli assiomi definiti, le relazioni di sottoclasse, le proprietà inverse e l'integrazione tra le quattro ontologie importate sono correttamente specificati e non introducono contraddizioni nel modello.

La seconda verifica tramite Protégé ha previsto l'arricchimento dell'ontologia con *individuals* appositamente creati per verificare la *reasonability* operativa del modello attraverso scenari che rispecchiano i casi d'uso presentati nei paragrafi precedenti. Gli *individuals* inseriti traducono in istanze concrete i pattern concettuali discussi nella modellazione dei cinque requisiti fondamentali (capitoli 6.1 e 6.2): la stratificazione dei livelli di astrazione del digitale (contenutistico, logico e fisico), la verifica

[215] https://www.w3.org/RDF/Validator/.

dell'integrità attraverso hash code, l'estrazione e tipizzazione dei metadati nativi, la documentazione della *provenance* e la rappresentazione delle relazioni contestuali, permettendo di testare la classificazione automatica delle istanze, il rispetto dei vincoli di dominio e range, e la corretta inferenza delle relazioni[216]. Questa validazione delle asserzioni concrete sugli *individuals*, definita in logiche descrittive come Assertion Box (ABox) (Baader et al. 2003, 300), ha richiesto a HermiT 110 secondi per completare l'elaborazione, un incremento temporale di circa 29 secondi rispetto alla validazione della sola TBox (81 secondi), dovuto alla necessità di verificare la coerenza delle asserzioni concrete: oltre alla ricomputazione della gerarchia delle classi e delle proprietà, il *reasoner* ha elaborato le *class assertions* che definiscono l'appartenenza degli *individuals* alle classi e le *object property assertions* che specificano le relazioni tra *individuals*. Il *reasoner* ha confermato la consistenza dell'ontologia popolata, confermando che tutti gli *individuals* rispettano i vincoli logici definiti dalla TBox e che le inferenze automatiche sulle relazioni inverse e simmetriche vengono correttamente propagate.

Il terzo livello di validazione ha previsto l'utilizzo del validatore OOPS! (OntOlogy Pitfall Scanner), strumento ampiamente utilizzato nella comunità del Web Semantico per l'identificazione di anomalie strutturali e carenze metodologiche nelle ontologie (Poveda-Villalón et al. 2014). OOPS! è indipendente dall'editor ontologico utilizzato e si distingue per la capacità di rilevare un'ampia gamma di problematiche comuni nello sviluppo ontologico (definiti *pitfalls*), fornendo così un controllo sistematico e automatizzato della qualità. Il suo catalogo ufficiale comprende 41 tipi di *pitfalls*, di cui è in grado di verificarne automaticamente 33, rilevando ad esempio errori critici come la definizione di relazioni inverse sbagliate (*P05 - Defining wrong inverse relationships*), l'inclusione di cicli nella gerarchia di classi (*P06 - Including cycles in a class hierarchy*), o la mancata dichiarazione di domini e range nelle proprietà (*P11 - Missing domain or range in properties*). Può altresì evidenziare problemi di minor rilevanza ma comunque significativi, come la presenza di classi con etichette duplicate (*P32 - Several classes with the same label*) o l'uso di convenzioni di naming incoerenti (*P22 - Using different naming conventions in the ontology*).

L'analisi ha rilevato cinque *pitfalls* minori, la cui origine è riconducibile al modo in cui OOPS! analizza ontologie modulari: lo strumento valuta esclusivamente il file XML/RDF locale senza risolvere le direttive di importazione, generando falsi positivi per classi e proprietà che sono dichiarati in forma essenziale nel file BoDi ma completamente definiti nelle ontologie importate (RiC-O 1.1 e PREMIS 3.0).

[216] Il file contenente gli *individuals* di test è disponibile nel *repository* GitHub dedicato all'ontologia BoDi: https://github.com/LuciaGiagnolini12/bodi.

Il Pitfall P08 (*Missing annotations*) di OOPS! segnala l'assenza di `rdfs:label` e `rdfs:comment` per 19 elementi (13 classi e 6 proprietà) provenienti da RiC-O e PREMIS. Si tratta di un falso positivo: le annotazioni sono correttamente definite nelle ontologie di origine, mentre la ri-dichiarazione locale sarebbe ridondante[217]. Le classi `rico:Instantiation`, `rico:Agent`, `rico:Mechanism`, `premis:Fixity` e `premis:StorageLocation` e le proprietà `rico:isRelatedTo`, `rico:hasOrHadPart` e `rico:isOrWasPartOf` sono infatti già documentate nei rispettivi modelli.

Il Pitfall P04 (*Creating unconnected ontology elements*) segnala come "isolati" `premis:SoftwareAgent` e `prov:Entity`, che in BoDi sono dichiarati in forma minimale (`<owl:Class rdf:about = "..."/>`) poiché referenziati come superclassi. In questo senso, svolgono un ruolo chiave di ancoraggio semantico: `bodi:Software` è sottoclasse di `premis:SoftwareAgent` e `rico:Instantiation` e `rico:RecordResource` ereditano da `prov:Entity`, garantendo interoperabilità con PREMIS e PROV-O. La segnalazione deriva dal fatto che OOPS! considera le classi non connesse se prive di proprietà e annotazioni, senza valutare il loro ruolo come collegamento verso vocabolari esterni in una architettura modulare.

Pitfall P11 (*Missing domain or range in properties*), P13 (*Inverse relationships not explicitly declared*) e P35 (*Untyped property*) coinvolgono otto proprietà RiC-O (`rico:sRelatedTo`, `rico:hasOrHadPart`, `rico:isOrWasPartOf`, `rico:hasOrHadType`, `rico:isOrWasTypeOf`, `rico:isInstantiationAssociatedWithInstantiation`, `rico:isOrWasUnderAuthorityOf` `rico:hasOrHadAuthorityOver`) usate in BoDi come super-proprietà. Nel file RDF sono dichiarate solo con le relazioni gerarchiche (`rdfs:subPropertyOf`), senza dominio, codominio o inverse, comunque definiti in RiC-O. Anche in questo caso OOPS! non risolve le importazioni e interpreta la rappresentazione minimale come incompletezza. Le proprietà BoDi derivate (`bodi:generatedBy/hasGenerated`, `bodi:hasHashCode/isHashCodeOf`, `bodi:hasTechnicalDescription/isTechnicalDescriptionOf`) definiscono invece domini, range e inverse in modo completo, garantendo la specializzazione corretta secondo OWL.

[217] L'unico potenziale vantaggio della dichiarazione locale delle classi importate consisterebbe nel rendere esplicito, all'interno del file stesso, l'utilizzo di una determinata definizione. Tuttavia, questa esigenza è già soddisfatta dal meccanismo di `owl:imports`, che specifica la versione dell'ontologia di riferimento. La duplicazione locale, oltre a essere ridondante, introdurrebbe rischi di incoerenza e si porrebbe in contrasto con i principi di riuso e modularità propri della progettazione ontologica.

L'unico pitfall direttamente attribuibile a BoDi (pitfall P41: *no license declared*) ha comportato l'adozione di `dcterms:license` nelle annotazioni dell'ontologia, completando l'allineamento dell'ontologia alle *best practice* del LOD e ai principi FAIR.

Infine, il modello è stato sottoposto a una fase di test iterativa attraverso la sua applicazione al caso di studio dedicato a Valerio Evangelisti, che ha comportato l'analisi e la descrizione della partizione nativa digitale del fondo, come delineato nel capitolo 7. L'implementazione del modello sul caso concreto ha avuto un ruolo decisivo nello sviluppo stesso dell'ontologia, consentendo di valutarne la capacità rappresentativa in situazioni reali e di coglierne progressivamente la visione d'insieme. Questo processo ha favorito un affinamento graduale della modellazione, con l'introduzione di modifiche e integrazioni mirate in risposta alle esigenze emerse dall'osservazione diretta dei dati e dal confronto con i requisiti concettuali.

In sintesi, il processo di validazione ha confermato la correttezza sintattica, la coerenza logica e la consistenza semantica dell'ontologia BoDi. La validazione RDF ha attestato la conformità formale del file agli standard W3C, mentre le verifiche con il *reasoner* HermiT hanno confermato l'assenza di contraddizioni sia a livello terminologico (TBox) sia a livello delle asserzioni (ABox), garantendo la possibilità di inferenza automatica e la coerenza delle relazioni. L'analisi con OOPS! non ha rilevato criticità sostanziali: le segnalazioni emerse riguardano principalmente limiti dello strumento nell'elaborazione di ontologie modulari e non difetti intrinseci del modello. L'unico intervento richiesto ha riguardato l'aggiunta della licenza, in linea con le raccomandazioni FAIR e LOD. La sperimentazione sul caso di studio di Valerio Evangelisti ha consentito, infine, di verificare l'applicabilità del modello in un contesto reale e di affinarne progressivamente la struttura. Nel complesso, le verifiche effettuate confermano la solidità formale e l'adeguatezza concettuale dell'ontologia rispetto agli obiettivi di rappresentazione.

Parte III - Automazione

7. Automazione del processo di rappresentazione

In risposta alla RQ3 - Automazione, questo capitolo indaga in che modo le caratteristiche del digitale possano essere impiegate per automatizzare e facilitare i processi di descrizione archivistica. A tale scopo, illustra il workflow metodologico sviluppato per la rappresentazione di archivi nativi digitali in LOD e ne presenta l'applicazione alla partizione digitale dell'Archivio di Valerio Evangelisti.

Il *workflow* prende in input le cartelle dell'archivio nativo digitale e produce in output una *knowledge base* in formato RDF, conforme al modello BoDi (Born-Digital Ontology). Il workflow si articola nelle seguenti cinque fasi operative:

Fase 1 - Sistematizzazione dell'informazione esistente: questa fase costituisce la base dell'intero processo e mira a produrre una rappresentazione formalizzata della struttura dei contenuti dell'archivio. La fase è articolata in dieci passaggi progressivi e trasforma file e cartelle in una rappresentazione semantica strutturata. Il processo inizia con l'impostazione di misure di protezione che rendono l'archivio immutabile, prosegue attraverso un censimento gerarchico completo e la generazione di impronte digitali crittografiche SHA-256 per ogni di file. La fase procede con l'estrazione sistematica dei metadati attraverso tre strumenti specializzati (libreria Python os, Apache Tika, ExifTool) garantendo una copertura delle caratteristiche tecniche di ogni documento. Tutte le informazioni catturate vengono infine sottoposte ad un processo di trasformazione semantica che permette di rappresentarle come entità RDF conformi a BoDi (e alle ontologie che importa). Controlli di integrità finali assicurano che nessuna operazione abbia compromesso l'autenticità dei materiali originali.

Fase 2 - Validazione e arricchimento rule-based: questa fase opera esclusivamente nel dominio semantico per validare la coerenza logica dei grafi prodotti ed esplicitare conoscenza implicita attraverso processi di inferenza controllata. Il sistema di validazione implementa una serie di *query* SPARQL che verificano la consistenza strutturale, la presenza dei metadati e la coerenza delle relazioni prodotte dalla prima fase. Attestata la validità del grafo, si procede con l'arricchimento semantico della *knowledge base* attraverso sei operazioni di elaborazione e consolidamento dei dati: correlazione di file con identico contenuto mediante analisi dei codici hash; allineamento di tipologie di metadati estratti da strumenti diversi; suddivisione dei metadati in nove categorie funzionali; assegnazione di tipologie documentali sulla base dei *media type*; attribuzione di titolo e date (creazione e modifica) alle varie entità secondo il modello RiC-O.

Fase 3 - Arricchimento semantico mediante conoscenza specialistica e modelli generativi: questa fase estende il perimetro informativo oltre la *knowledge base* esistente attraverso l'integrazione di fonti di conoscenza specialistica e l'applicazione di tecnologie di IA. L'approccio prevede due diversi processi di arricchimento: un processo semiautomatico di mappatura delle opere letterarie di Evangelisti secondo il modello LRMoo (IFLA Library Reference Model), che collega sistematicamente i documenti d'archivio alle opere di riferimento attraverso analisi bibliografica specialistica; un sistema automatizzato basato su LLMs per produrre descrizioni sinottiche in linguaggio naturale derivate dall'analisi dei metadati estratti.

Fase 4 - Documentazione dei contesti di origine: questa fase riguarda la ricostruzione dei contesti di origine dei materiali e la documentazione della catena di custodia che collega i supporti originali alle copie di lavoro utilizzate per l'analisi. Prevede un censimento dei diversi ambienti digitali, la ricostruzione delle loro caratteristiche e delle relazioni temporali e processuali che li coinvolgono. Dal censimento si ricavano dati minimi descrittivi, rappresentati secondo il modello BoDi. In questo modo, il grafo prodotto nelle fasi precedenti si arricchisce di informazioni sui contesti di provenienza e sulla storia conservativa dei documenti digitali.

Fase 5 -Anonimizzazione per la pubblicazione: l'ultima fase implementa strategie di anonimizzazione selettiva che preservano la struttura e il portato informativo dell'archivio garantendo al contempo la protezione delle informazioni sensibili. Il sistema distingue tra diverse tipologie di risorse archivistiche applicando logiche differenziate e modificabili a seconda delle esigenze del caso.

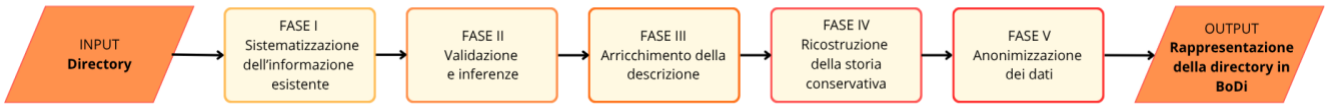


Figura 7.1. Workflow in cinque fasi per la trasformazione di archivi nativi digitali in una knowledge base RDF conforme al modello BoDi.

Il *workflow* (Figura 7.1) è stato progettato per perseguire obiettivi metodologici specifici:

- **Preservazione dell'integrità**: ogni operazione deve mantenere inalterata l'integrità fisica e logica dei documenti originali, implementando meccanismi di verifica crittografica e sistemi di backup preventivi che assicurano la tracciabilità completa delle trasformazioni.
- **Reversibilità e trasparenza**: tutte le trasformazioni devono essere documentate e reversibili, e collegano inequivocabilmente i documenti fisici alle loro rappresentazioni semantiche attraverso processi verificabili e riproducibili.

- **Conformità agli standard**: le rappresentazioni generate devono aderire a modelli di rappresentazione documentati per garantire portabilità, interoperabilità e riusabilità.

L'applicazione sperimentale del workflow è stata condotta sull'Archivio Valerio Evangelisti, utilizzando come caso di studio le prime tre *directories* del fondo: la copia dell'hard disk del computer principale, la copia dell'hard disk esterno e l'estrazione dei contenuti dei floppy disk[218].
Le sezioni seguenti presentano nel dettaglio l'implementazione tecnica di ciascuna fase, documentando le scelte metodologiche, gli strumenti utilizzati e i controlli di qualità implementati, con l'obiettivo di fornire un modello replicabile per la rappresentazione di archivi nativi digitali d'autore[219].

7.1 Sistematizzazione dell'informazione esistente

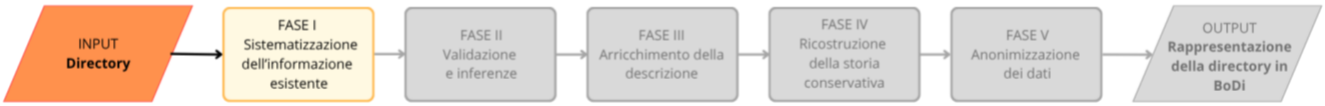


Figura 7.2. Prima fase del workflow: sistematizzazione dell'informazione esistente.

L'asse portante del progetto è rappresentato dalla sistematizzazione dell'informazione esistente, fase iniziale e al contempo più cruciale dell'intero processo, che si configura come una fotografia computazionale dello stato dell'archivio. A partire da una o più *directory*[220], questa fase ha l'obiettivo di produrre una baseline di dati stabile, immutabile e verificabile, che rappresenti l'archivio nel rispetto della struttura e dei metadati tecnici associati a file e cartelle. Tale baseline è semantizzata in un grafo RDF, rappresentante lo stato fattuale della *knowledge base*, che fungerà da riferimento per tutte le trasformazioni successive.

[218] Il computer portatile, pervenuto solo in tempi recenti, non è stato incluso nell'implementazione corrente a causa delle tempistiche ristrette del progetto. L'Associazione Evangelisti si riserva di verificarne preliminarmente i contenuti nel rispetto della loro integrità. La sua integrazione è prevista come sviluppo futuro e consentirà l'avvio di una nuova fase di testing del modello e del workflow.
[219] Il codice che implementa il workflow è consultabile nel *repository* dedicato: https://github.com/LuciaGiagnolini12/FileSystem_to_BoDi.
[220] In questa sede, con il termine *directory* si intendono i livelli di radice (*root*) delle cartelle analizzate, cioè le cartelle principali da cui si diramano tutti i file e le sottocartelle di ciascun supporto considerato. In particolare, si fa riferimento alla cartella contenente i dati dell'hard drive del computer fisso, a quella corrispondente all'hard drive esterno e a quella che raccoglie i riversamenti dei floppy disk.

Questa prima fase è articolata in dieci passaggi (figura 7.3) elaborati secondo una logica di raffinamento progressivo, dalla gestione della materialità dei supporti digitali fino alla loro rappresentazione semantica.

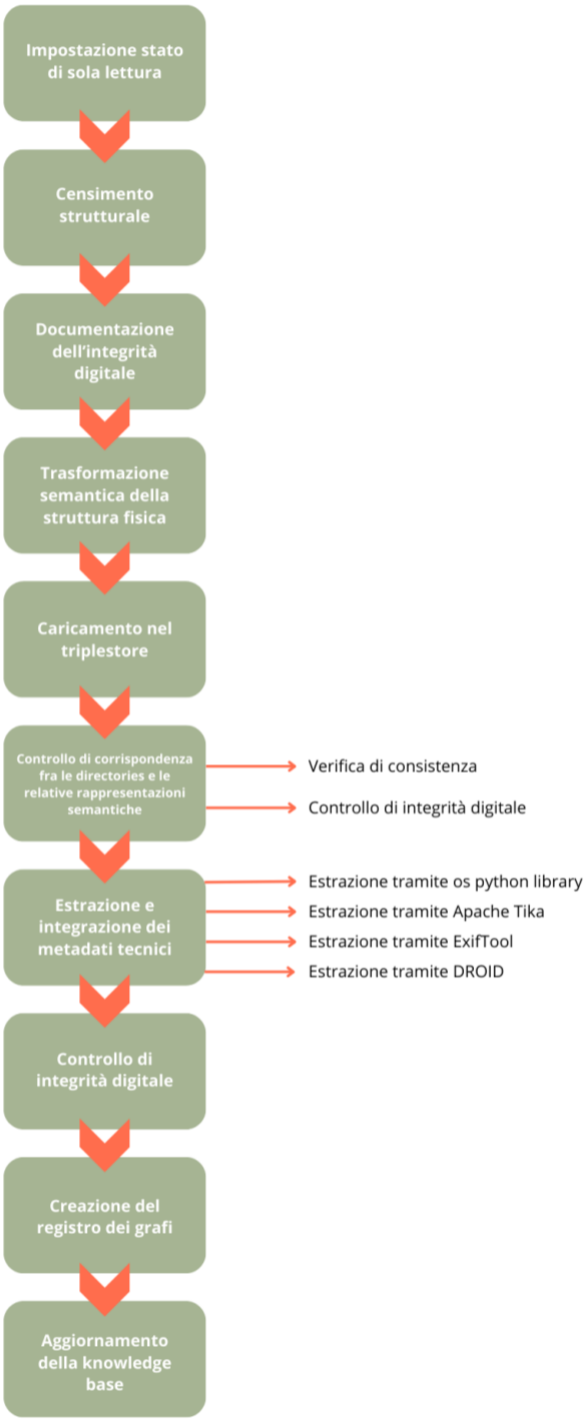


Figura 7.3. Rappresentazione grafica degli step previsti dalla prima fase del workflow.

Tutti gli step vengono eseguiti nello stesso ambiente operativo[221] (in termini sistema e configurazione) al fine di eliminare variabili legate all'eterogeneità delle piattaforme e dei sistemi che potrebbero portare a idiosincrasie fra i vari step. Di seguito viene presentata la descrizione dettagliata di ciascun passaggio, con particolare attenzione agli aspetti tecnici e metodologici implementati.

1. **Impostazione stato di sola lettura**. Per prima cosa, si procede all'applicazione di misure di protezione volte a rendere immutabile la fonte: l'intero corpus viene impostato in modalità *read-only*, onde evitare che le operazioni successive possano apportare modifiche. Il sistema genera automaticamente uno script di *recovery* specifico per ogni *directory* analizzata in un dato momento, garantendo la reversibilità completa del processo di protezione.

2. **Documentazione dell'integrità digitale**. Per ciascun documento viene generata un'impronta digitale univoca utilizzando l'algoritmo crittografico SHA-256. Questo identificatore garantisce l'integrità dei file poiché qualsiasi modifica del contenuto, anche minima, produce una stringa di hash completamente diversa. L'hash funziona quindi come chiave univoca del contenuto, permettendo di verificare che i file non siano stati alterati nel tempo e di identificare facilmente versioni diverse dello stesso documento. Il calcolo viene effettuato separatamente per ogni *directory* e documentato in file JSON dedicati (`FloppyDisk_HASH.json`, `HD_HASH.json`, `HDEsterno_HASH.json`). Ciascun registro JSON contiene: codici hash SHA-256 per ogni file, dimensioni dei file in bytes, *timestamp* di ultima modifica, percorsi completi dei file, statistiche del processo di calcolo (file elaborati, errori, performance) e comando utilizzato per il calcolo.

   I registri garantiscono la possibilità di verificare in futuro che nessun file sia stato modificato o corrotto rispetto allo stato di partenza, assicurando che eventuali compromissioni dell'integrità possano essere immediatamente rilevate.

3. **Trasformazione semantica della struttura fisica.** La struttura del *file system* viene convertita in una rappresentazione semantica conforme a BoDi (e alle ontologie che importa). In particolare,

[221] Nello specifico, l'ambiente sperimentale del caso di studio appartiene all'infrastruttura tecnica del Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna. L'architettura hardware è basata su un server Dell PowerEdge R7525, dotato di due CPU AMD EPYC 7313 16-Core Processor (versione 25.1.1, 1500 MHz, capacità massima 3729 MHz, architettura a 64 bit), 256 GB di RAM, un'unità SSD da 1 TB e un array RAID da 12 TB su hard disk rotativi. Il sistema operativo è Ubuntu 24.04.3 LTS ("noble") con shell Bash 5.2.21-2ubuntu4 (/bin/bash). L'implementazione utilizza Python 3.12.3, gestito tramite un *virtual environment* contenente librerie principali per l'elaborazione e l'interrogazione semantica: certifi 2025.6.15, charset-normalizer 3.4.2, duration 1.1.1, idna 3.10, pyparsing 3.2.3, rdflib 7.1.4, requests 2.32.4, setuptools 80.9.0, SPARQLWrapper 2.0.0, tika 3.1.0 e urllib3 2.5.0.

vengono generate[222]:

- entità di tipo `rico:Record` e `rico:RecordSet` per rappresentare file e cartelle dal punto di vista concettuale-contenutistico;
- entità di tipo `rico:Instantiation` per rappresentare le dimensioni logiche di `rico:Record` e `rico:RecordSet`, ossia la concreta istanziazione in *media* di file e cartelle;
- relazioni gerarchiche tra le rappresentazioni concettuali-contenutistiche delle entità archivistiche (`rico:RecordSet rico:includesOrIncluded rico:Record` e inversa) e, specularmente, relazioni gerarchiche tra le rappresentazioni dei *media* ad esse correlate (`rico:Instantiation rico:hasOrHadPart rico:Instantiation` e inversa), così come le relazioni tra le rappresentazioni concettuali-contenutistiche e le rappresentazioni dei media fisici (`rico:Record/RecordSet rico:hasOrHadInstantiation rico:Instantiation` e inversa);
- entità di tipo `premis:Fixity` per rappresentare il codice hash associato ad ogni documento (che viene ricalcolato in questa fase, sia per finalità di rappresentazione che per finalità di controllo dell'integrità del fondo (v. step. 6));
- entità di tipo `bodi:Algorithm` per rappresentare il l'algoritmo utilizzato per il processo di calcolo degli SHA256 e proprietà correlate per definire specifiche tecniche (`rico:hasTechnicalCharacteristics`) e relazioni derivanti dal processo;
- entità di tipo `rico:Identifier` e `rico:IdentifierType`, per la rappresentazione di un identificativo univoco associato ad ogni `rico:Record` e `rico:RecordSet` e le rispettive relazioni con questi ultimi;
- entità di tipo `prov:Location` per rappresentare i percorsi dei file all'interno delle *directory*;
- relazioni fra `rico:Instantiation` e le rispettive `premis:Fixity` e `prov:Location`;
- una serie di specifiche identificative o descrittive delle entità individuate tramite *data property* (`rdfs:label`, `rdf:value` o più specifiche, come `bodi:hierarchyDepth` per individuare la profondità della posizione del file all'interno della gerarchia della *directory*, dove 0 corrisponde al livello root).

La rappresentazione viene formalizzata in file in formato N-Quads (uno per ogni *directory*:

[222] Il processo si limita alla modellazione del livello logico del sistema operativo, senza estendersi alla rappresentazione di aspetti fisico-strutturali quali la configurazione delle partizioni disco. Sebbene il modello BoDi preveda l'utilizzo della classe `bodi:FileClusterPosition` per descrivere tali dettagli, nel caso d'uso dell'Archivio Evangelisti l'analisi è stata eseguita su copie non forensi dei supporti originali conservati su server. Di conseguenza, non sono disponibili dati attendibili per una rappresentazione accurata di questo livello di granularità.

`structure_floppy.nq`, `structure_hd.nq`, `structure_hdesterno.nq`), ossia file che contengono quadruple (la tripla RDF associata al grafo di appartenenza, ad es. `http://ficlit.unibo.it/ArchivioEvangelisti/RS1_RS4_R17_inst; https://www.ica.org/standards/RiC/ontology#isOrWasInstantiationOf; http://ficlit.unibo.it/ArchivioEvangelisti/RS1_RS4_R17; http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS4`). Infatti, per ogni *directory* viene generato un *named graph* dedicato ad accogliere le rispettive informazioni strutturali derivanti da questa fase:

- `http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1` contiene le triple relative alla rappresentazione della struttura della *directory* contenente la copia dell'hard disk del computer principale di Evangelisti;
- `http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2` contiene le triple relative alla struttura della copia dell'hard disk esterno;
- `http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3` contiene le triple relative alla struttura della *directory* contenente l'estrazione dei contenuti dei floppy disk.

Questa strategia di suddivisione in *named graph* rappresenta un approccio architetturale adottato sistematicamente in tutta la pipeline di processamento: ogni fase produce uno o più grafi specifici in base alla finalità e alla tipologia delle informazioni contenute. Tale metodologia facilita la gestione modulare dei dati nel *triplestore*[223], permettendo aggiornamenti mirati e ottimizzando l'esecuzione di query contestualizzate su porzioni specifiche della *knowledge base*.

4. **Caricamento nel *triplestore*.** I file RDF vengono caricati in un *triplestore* per consentire interrogazioni semantiche avanzate. Il sistema implementa un'architettura modulare che separa la generazione dei dati (generazione dei file N-Quads) dal caricamento specifico (*triplestore-dependent*), garantendo portabilità e interoperabilità. La configurazione di default utilizza Blazegraph[224] come *triplestore* di destinazione, implementando un processo di caricamento che sfrutta le REST API native per il trasferimento diretto mediante HTTP POST con Content-Type application/n-quads.

   La selezione di Blazegraph risponde a due ordini di motivazioni: dal punto di vista metodologico, il carattere *open source*, la documentazione esaustiva e l'utilizzo consolidato nella ricerca ne fanno una

[223] Le triple di dati RDF possono essere archiviate in uno specifico database (*triplestore*), attraverso cui le interrogazioni (*query*) possono essere fatte attraverso il linguaggio SPARQL (Open Knowledge s.d.).
[224] https://blazegraph.com/.

buona soluzione per garantire la riproducibilità; dal punto di vista tecnico-operativo, per la compatibilità con ResearchSpace[225], l'ambiente selezionato per la pubblicazione e consultazione del dataset (v. capitolo 9). In ogni caso, l'approccio architetturale garantisce flessibilità di *deployment*: i file N-Quads generati, essendo conformi agli standard W3C, possono essere caricati in qualsiasi *triplestore* compatibile (Apache Jena Fuseki[226], QLever[227], Amazon Neptune[228], ecc.) senza modifiche strutturali, assicurando piena interoperabilità e riusabilità del dataset in contesti eterogenei.

5. **Controllo di corrispondenza fra le directory e le relative rappresentazioni semantiche.** Una fase cruciale del processo è costituita dal controllo sistematico di coerenza tra la realtà fattuale documentata negli step 2 e 3 e la rappresentazione semantica generata nello step 4. Il controllo si articola in due dimensioni:

   - **Verifica di consistenza**: mediante *query* SPARQL avviate via API sulla *knowledge base* caricata su Blazegraph, si confrontano i conteggi dei file per ogni cartella desumibili dai grafi con quelli precedentemente registrati nel JSON prodotto dallo step 2. Le due fotografie di consistenza devono essere perfettamente sovrapponibili: l'obiettivo è assicurare che nessun documento sia stato perso, ignorato o duplicato durante la generazione dei grafi.
   - **Controllo di integrità**: seguendo la stessa logica, si verifica che gli SHA-256 (hash code) calcolati nello step 3 corrispondano esattamente a quelli riportati nelle entità RDF (frutto di un secondo ricalcolo). Anche questa verifica viene svolta tramite query SPARQL e costituisce la garanzia formale che l'integrità binaria dei documenti sia stata preservata durante le operazioni di migrazione semantica.

   Sebbene tutte le operazioni siano eseguite in un ambiente operativo omogeneo, è opportuno sottolineare che, sia per i conteggi sia per il calcolo dei valori hash, sono stati deliberatamente adottati strumenti differenti, allo scopo di verificare la neutralità dell'output rispetto alla tecnologia impiegata. In particolare, la modalità di conteggio nello step 2 si basa su un comando Bash primario (`find`), mentre nello step 6 il conteggio è derivato dai dati RDF interrogati con SPARQL, contando i `rico:Record` legati da `rico:isOrWasIncludedIn` ai `rico:RecordSet` con query e computazione ricorsiva implementata in Python. Analogamente, per il calcolo dei codici hash, nello step 3 si

[225] https://researchspace.org/.
[226] https://jena.apache.org/index.html.
[227] https://github.com/ad-freiburg/qlever.
[228] https://aws.amazon.com/it/neptune/.

utilizzano i comandi nativi del sistema operativo (`shasum`[229] su macOS, `sha256sum`[230] su Linux), chiamati tramite *subprocess*, mentre nello step 4 l'hash viene ricalcolato con la libreria hashlib.sha256() di Python[231]. Questa doppia validazione, indipendentemente dagli strumenti utilizzati, è stata implementata come criterio di verifica della qualità del dato e per assicurare la riproducibilità dei risultati in contesti eterogenei, lasciando all'utente la possibilità di selezionare gli strumenti più adatti al proprio ambiente.

6. **Estrazione e integrazione dei metadati tecnici**. Superata con esito positivo la fase di validazione, il sistema procede con l'estrazione sistematica dei metadati da ciascun documento presente nell'archivio. Questa operazione si articola su tre livelli distinti, ciascuno affidato a strumenti specializzati in grado di restituire prospettive analoghe e/o complementari sulla stessa entità digitale:
   - **Estrazione tramite Libreria Python `os`**[232], che permette l'interazione con il sistema operativo, utilizzata per l'estrazione diretta dei metadati di *file system*. Tra le informazioni acquisite figurano dimensioni dei file, *timestamp* (creazione, modifica, accesso), permessi di accesso e identificativi univoci (*inode*), dati essenziali per documentare lo stato tecnico e il contesto operativo del documento all'interno dell'ambiente digitale originario. A differenza dei successivi due strumenti, questa libreria, agendo a livello di *file system*, riesce a recuperare e mappare anche i metadati relativi alle cartelle.
   - **Estrazione tramite Apache Tika**, software *open source* di *parsing* multiformato, impiegato per l'estrazione di metadati di contenuto da un ampio spettro di tipologie documentarie (oltre cento formati, tra cui PDF, documenti Office, immagini raster, archivi compressi, file multimediali)[233]. I metadati ottenuti includono elementi informativi interni come creatore, titolo, data di creazione e ultima modifica, abstract, keywords, e altri attributi rilevanti a seconda della tipologia del file. Apache Tika è pensato come un estrattore generalista: oltre a leggere i metadati nativi del file, cerca armonizzarli mappandoli su vocabolari standard, come Dublin Core e XMP (Apache Software Foundation 2012).
   - **Estrazione tramite ExifTool**[234], strumento *open source* specializzato nell'analisi di file

[229] https://ss64.com/mac/shasum.html.
[230] https://sha256sum.com/.
[231] https://docs.python.org/3.12/library/hashlib.html Il caso di studio è stato implementato utilizzando la versione Python 3.12.3.
[232] https://docs.python.org/3/library/os.html.
[233] https://tika.apache.org/. Per il caso di studio è stata utilizzata la versione server 3.2.1.
[234] https://exiftool.org/. Per il caso di studio è stata utilizzata la versione 12.76.

multimediali, in particolare immagini e video. Oltre ad alcune informazioni sovrapponibili a quelle restituite da `os` e Apache Tika, ExifTool fornisce metadati tecnici estremamente dettagliati e specifici per ciascun tipo di media, tra cui risoluzione, profondità di colore, metadati EXIF, codec utilizzati, *bitrate*, durata, frequenze audio, parametri di compressione e numerosi altri dati.

La strategia adottata risponde all'esigenza di condurre un'analisi multidimensionale del medesimo file, incrementando notevolmente gli elementi descrittivi disponibili per l'indagine archivistica e filologica. Tale approccio si fonda sulla considerazione, approfondita nei capitoli 4.2 e 4.3, che l'utilizzo di strumenti differenti può produrre descrizioni diverse a partire dagli stessi contenuti e quindi generare metadati diversi per lo stesso oggetto. Normalizzando i dati in conformità con BoDi, questo step si configura, dunque, come una pluralizzazione semantica, offrendo al ricercatore una base informativa articolata e comparabile sullo stesso documento[235].

Per ogni *directory* vengono generati tre *named graphs* dedicati ad accogliere le informazioni provenienti dai tre strumenti di estrazione:

- `http://ficlit.unibo.it/ArchivioEvangelisti/FS_TechMeta_hd`,
  `http://ficlit.unibo.it/ArchivioEvangelisti/AT_TechMeta_hd`,
  `http://ficlit.unibo.it/ArchivioEvangelisti/ET_TechMeta_hd`
  contengono rispettivamente le triple relative all'estrazione prodotta dalla libreria Python os, da Apache Tika e ExifTool della *directory* contenente la copia dell'hard disk del computer principale di Evangelisti;
- `http://ficlit.unibo.it/ArchivioEvangelisti/FS_TechMeta_HDEsterno`,
  `http://ficlit.unibo.it/ArchivioEvangelisti/AT_TechMeta_HDEsterno`,
  `http://ficlit.unibo.it/ArchivioEvangelisti/ET_TechMeta_HDEsterno`
  contengono rispettivamente le triple relative all'estrazione prodotta per la copia dell'hard disk

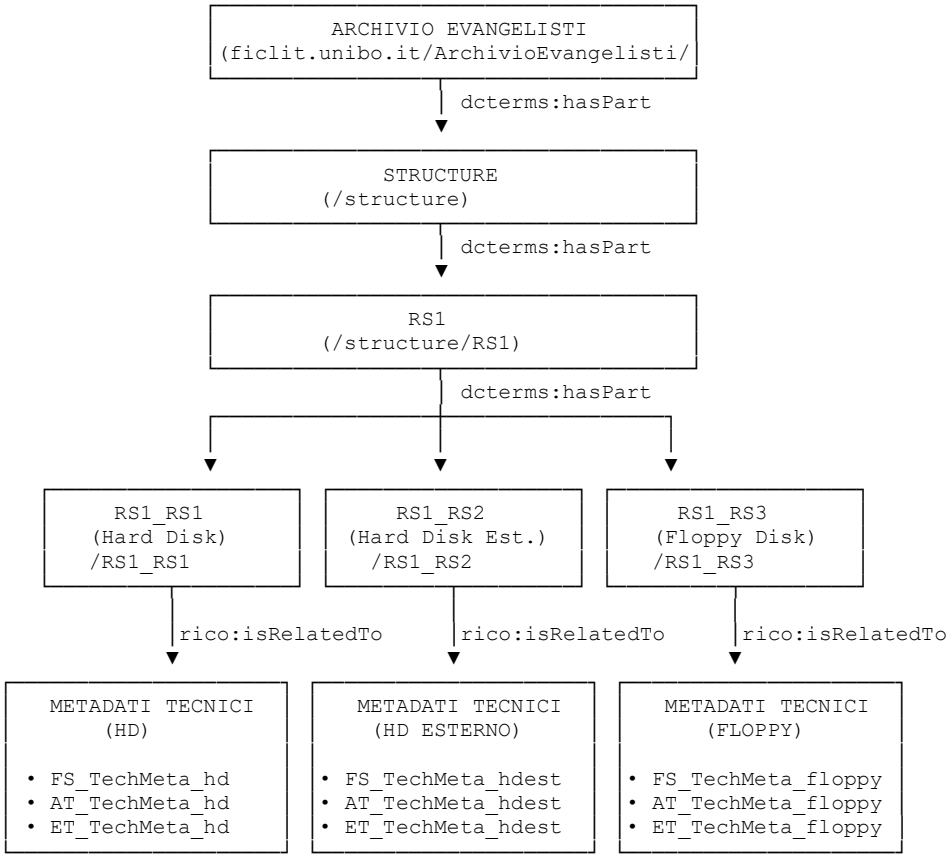
[235] Il sistema implementa un meccanismo a due livelli per garantire l'unicità degli identificatori URI dei metadati nonostante esecuzioni multiple del processo di estrazione. Il primo livello utilizza strutture dati in memoria (dizionari Python) per controlli rapidi, associando coppie di valori che identificano univocamente ogni tipo di metadato all'URI corrispondente. Il secondo livello persiste contatori incrementali e mappature nel file *evangelisti_uri_counters.json*. Ogni tipologia di metadato è identificata da due componenti: il nome del campo estratto (come "File Size", "Content-Type", "PUID") e il software che lo ha generato (ExifTool, Apache Tika, Python os). All'avvio, il sistema carica il JSON popolando le cache. Durante l'elaborazione, viene verificata l'esistenza della tipologia di metadato: se presente, viene riutilizzato, altrimenti viene generato un nuovo URI con contatore incrementale. Ad esempio, "File Size" da ExifTool crea ("File_Size", "exiftool"), e la desinenza URI exiftool_tmtype_0015, mentre lo stesso campo da Python os genera ("File_Size", "os") e la desinenza URI os_tmtype_0008, mantenendo distinte le provenienze. Successivamente, ogni volta che, ad esempio, verrà estratto "File Size" da ExifTool dall'analisi di un nuovo file, il sistema riutilizzerà exiftool_tmtype_0015 come base per l'URI. Le cache vengono serializzate periodicamente nel JSON convertendo le tuple in stringhe, garantendo identità persistenti degli URI tra diverse esecuzioni ed evitando la generazione di duplicati e ridondanze.

esterno;

- `http://ficlit.unibo.it/ArchivioEvangelisti/FS_TechMeta_floppy`, `http://ficlit.unibo.it/ArchivioEvangelisti/AT_TechMeta_floppy`, `http://ficlit.unibo.it/ArchivioEvangelisti/ET_TechMeta_floppy` contengono rispettivamente le triple relative all'estrazione prodotta per la *directory* contenente l'estrazione dei contenuti dei floppy disk.

7. **Secondo controllo di integrità digitale**. Terminato il processo di estrazione metadati, per validare il fatto che l'estrazione non abbia compromesso in alcun modo l'integrità dei materiali digitali, viene implementata una seconda procedura di verifica dell'integrità. Prima del ricalcolo degli hash, il sistema crea automaticamente un backup del file JSON contenente i codici hash originali calcolati nello step 2. Successivamente, viene eseguito un nuovo calcolo completo degli hash di tutti i file nelle *directory* elaborate. Il sistema confronta quindi gli hash pre-estrazione (conservati nel backup) con quelli post-estrazione (appena ricalcolati) attraverso un'analisi file per file. Se tutti gli hash risultano identici, la verifica ha successo e il backup viene rimosso, certificando che nessuna modifica è stata apportata ai materiali durante l'estrazione dei metadati. In caso di discrepanze, il sistema mantiene entrambi i file per analisi diagnostica, consentendo di identificare precisamente quali file sono stati alterati e facilitando la risoluzione di eventuali problematiche nell'integrità dei dati.

8. **Creazione del registro dei grafi**. Fino a questo step sono stati creati tanti grafi strutturali quante le *directory*; e per ogni *directory* sono stati generati tre grafi contenenti l'esito dell'estrazione dei metadati con i tre strumenti (libreria os, Apache Tika, Exiftool). Con la generazione di decine di *named graph* distribuiti tra informazioni strutturali e metadati estratti con diversi strumenti, si presenta la necessità di implementare un meccanismo di governance che garantisca tracciabilità e la navigabilità dell'ecosistema informativo. La complessità derivante dalla molteplicità di grafi rende infatti problematico determinare, ad esempio, quale grafo contenga specifiche tipologie di informazioni, o stabilire connessioni tra la rappresentazione strutturale di una *directory* e i relativi metadati estratti. Per stabilire una relazione intelligibile fra i grafi prodotti, in questo step viene generato un ulteriore grafo (*NGRegistry*) che identifica e formalizza i rapporti tra le diverse componenti del sistema, fungendo da registro centralizzato che documenta le dipendenze, facilita l'identificazione dei grafi rilevanti per specifiche interrogazioni e permette operazioni di manutenzione mirate. Questo approccio, dunque, mantiene la modularità dei singoli grafi preservando la visione d'insieme, identificando i rapporti in questo modo:



9. **Aggiornamento della knowledge base**. Infine, i grafi generati dall'estrazione dei metadati e il registro dei grafi vengono caricati in Blazegraph (con le stesse modalità descritte al punto 5).

Ogni step del processo descritto è implementato in uno o più script Python dedicati, responsabili di specifici compiti operativi (impostazione *read-only*, censimento, calcolo dei codici hash, estrazione dei metadati, eccetera). Tali moduli processano le *directory* una per una, garantendo che tutte le operazioni previste da uno step vengano completate su ogni *directory* prima che il sistema prosegua con la fase successiva. Il flusso di lavoro è orchestrato attraverso una pipeline centralizzata (`pipeline.py`), che funge da orchestratore del processo. Lo script di pipeline gestisce l'esecuzione dei nove moduli Python principali:

- `config_loader.py`, per il caricamento centralizzato della configurazione (`directory_config.json`) che permette la gestione del processamento delle *directory* archivistiche, dando la possibilità di individuare tutti i *path* delle cartelle da analizzare e impostando la configurazione degli output ad essi correlati (step 2);
- `file_count.py`, per il censimento strutturale (step 1 e 2);
- `hash_calc.py`, per la generazione degli hash SHA-256 (step 3);
- `structure_generation.py`, per la trasformazione semantica della struttura (step 4);

- `blazegraph_loader.py`, per caricare le triple generate tramite le REST API di Blazegraph (step 5 e 10);
- `count_check.py`, per la verifica della consistenza strutturale (step 6);
- `integrity_check.py`, per la verifica dell'integrità binaria dei file (step 6 e 8);
- `metadata_extraction.py`, per l'estrazione sistematica dei metadati con la libreria os, Apache Tika e ExifTool (step 7 e 9);
- `journal_restore.py`, per il *restore* del journal blazegraph, ad ogni nuovo avvio della pipeline.

L'ordine di esecuzione e le dipendenze fra i moduli sono rappresentati mediante lo pseudocodice visibile nel Listato 7.1, che documenta il comportamento della pipeline.

```
FUNZIONE pipeline_evangelisti(directory_configs)
    # GESTORE PRINCIPALE
    FILE: evangelisti_pipeline_test.py

    # CONFIGURAZIONE CENTRALIZZATA
FILE: config_loader.py
SCOPO: Caricamento configurazioni unified da directory_config.json

    PER OGNI directory IN [floppy, hd, hdesterno]:

        // STEP 1: Impostazione stato di sola lettura
        FILE: file_count.py
        INPUT:  directory.path (directory fisica target)
        EXECUTE: python file_count.py -w -f directory.path
        OUTPUT:
            - Script recovery dei permessi r/w
            - Directory in modalità read-only

        // STEP 2: Censimento strutturale
        FILE: file_count.py
        INPUT:  directory.path (directory read-only)
        EXECUTE: python file_count.py -r -e directory.path -o
                 directory.count_output
        OUTPUT:
            - File JSON conteggi (es. FloppyDisk_CNT.json)
            - Struttura: conteggio_totale, sottocartelle[path, file_count]

        // STEP 3: Documentazione integrità digitale
        FILE: hash_calc.py
        INPUT:  directory.path (directory read-only)
        EXECUTE: python hash_calc.py directory.path
        OUTPUT:
            - File JSON hash SHA-256 (es. FloppyDisk_HASH.json)con struttura:
              file_hashes[path, sha256, size, modified]

        // STEP 4: Trasformazione semantica struttura fisica
        FILE: evangelisti_structure_generation.py
        INPUT:
            - directory.path (directory fisica)
```

```
    - directory.root_id (identificatore struttura archivistica)
EXECUTE: python evangelisti_structure_generation.py --type
         directory.structure_type
OUTPUT:
    - File N-Quads struttura RDF (es. structure_floppy.nq)
      Named graph:
       http://ficlit.unibo.it/ArchivioEvangelisti/structure/ROOT_ID
      Entità create: rico:Record, rico:RecordSet, rico:Instantiation,
       premis:Fixity, prov:Location

// STEP 5: Caricamento nel sistema RDF
FILE: blazegraph_loader.py (chiamato da evangelisti_pipeline_test.py)
CLASS: BlazegraphJournalGeneratorREST
INPUT:  File N-Quads struttura (structure_*.nq)
EXECUTE: BlazegraphRESTLoader.load_multiple_files()
OUTPUT:
    - Dati struttura caricati in Blazegraph triplestore
    - Endpoint SPARQL attivo:
      http://localhost:9999/blazegraph/namespace/kb/sparql

// STEP 6: Controllo di qualità sistematico
// STEP 6a: Verifica consistenza conteggi
FILE: count_check.py
INPUT:
    - File JSON conteggi (step 2)
    - Blazegraph triplestore (step 5)
EXECUTE: python count_check.py directory.check_type
OUTPUT: Report validazione consistenza conteggi

// STEP 6b: Verifica integrità hash
FILE: integrity_check.py
INPUT:
    - File JSON hash (step 3)
    - Blazegraph triplestore (step 5)
EXECUTE: python integrity_check.py directory.check_type
OUTPUT: Report validazione integrità hash

// STEP 7: Estrazione e integrazione metadati tecnici
FILE: evangelisti_metadata_extraction.py
INPUT:
    - directory.path (directory fisica)
    - File N-Quads struttura (per URI instantiation)
EXECUTE: python evangelisti_metadata_extraction.py
         directory.metadata_directory
TOOLS: Python os + Apache Tika + ExifTool
OUTPUT:
    - FileSystem_TechMeta_SUFFIX.nq (metadati estratti mediante .os
      library)
    - ApacheTika_TechMeta_SUFFIX.nq (metadati estratti mediante Apache
      Tika)
    - ExifTool_TechMeta_SUFFIX.nq (metadati estratti mediante ExifTool)
    - evangelisti_uri_counters.json (contatori URI persistenti)

// STEP 8: Verifica consistenza hash post-estrazione metadati
FILE: evangelisti_pipeline_test.py (metodo
      step_6_5_verify_hash_consistency)
INPUT:
    - File JSON hash originali (step 3: FloppyDisk_HASH.json)
    - directory.path (directory fisica post-estrazione)
```

```
       EXECUTE:
           1. Backup automatico hash originali (FloppyDisk_HASH_backup.json)
           2. Ricalcolo hash correnti: python hash_calc.py directory.path
           3. Confronto dettagliato file-per-file: _compare_hash_files()
       OUTPUT:
              - Certificazione integrità: hash identici = file integri durante
                estrazione
              - In caso successo: rimozione backup temporaneo
              - In caso discrepanze:
                  * FloppyDisk_HASH.json (hash pre-estrazione, ripristinato)
                  * FloppyDisk_HASH_post_metadata.json (hash post-estrazione, per
                    analisi)
                  * Report diagnostico file modificati

          // STEP 9: Aggiornamento knowledge graph
          // STEP 9a: Generazione Named Graph Registry
          FILE: evangelisti_pipeline_test.py
          CLASS: NGRegistryGenerator
          INPUT:  Blazegraph triplestore con struttura + metadati
          EXECUTE: NGRegistryGenerator.generate_and_load_registry()
          OUTPUT:
              - NGRegistry.nq (indice dei grafi caricati)
              - Knowledge base completa con struttura + metadati + indice

          // STEP 9b: Caricamento metadati
          FILE: blazegraph_loader.py (chiamato da evangelisti_pipeline_test.py)
          CLASS: BlazegraphJournalGeneratorREST
          INPUT:  File N-Quads metadati (step 7)
          EXECUTE: BlazegraphRESTLoader.load_multiple_files()
          OUTPUT: Metadati integrati in Blazegraph triplestore


        // OUTPUT FINALE PIPELINE
        RETURN:
            - Blazegraph triplestore completo
            - File JSON documentazione (conteggi + hash)
            - File N-Quads di backup (struttura + metadati + indice)
            - Report validazione qualità dati
            - Report integrità digitale post-processing


# UTILITY BACKUP/RESTORE
FILE: journal_restore.py
CLASS: BlazegraphJournalRestorer
SCOPO: Ripristino journal da backup
```

Listato 7.1. Pseudocodice della pipeline per la prima fase del workflow.

La *pipeline* garantisce piena tracciabilità attraverso la documentazione di ogni step del processo mediante tracce log e il salvataggio sistematico su disco dei risultati intermedi in formato JSON e N-Quads, creando una memoria del processo in grado di supportare audit futuri. L'architettura implementa controlli

di dipendenza che impediscono l'avanzamento agli step successivi senza la verifica preliminare delle condizioni necessarie[236].

Dal punto di vista operativo, il sistema è progettato per gestire archivi di dimensione variabile attraverso elaborazione parallela e gestione ottimizzata della memoria, minimizzando l'impatto sulle risorse del sistema ospitante, consentendo di processare anche dataset di grandi dimensioni attraverso salvataggi periodici e gestione incrementale dei dati.

Il risultato finale è una *knowledge base* che integra la struttura archivistica dell'archivio con i metadati tecnici di cartelle e documenti, relazioni e informazioni di *provenance* delle operazioni come esemplificato nella Figura 7.4. La baseline semantica così costituita rappresenta il fondamento per le fasi successive, che tratteranno dell'arricchimento di questa struttura semantica di base. È importante sottolineare che questa rappresenta l'unica fase dell'intero workflow che prevede la l'accesso diretto alle *directory* fornite: tutte le operazioni successive opereranno esclusivamente nel dominio semantico.

[236] I controlli di dipendenza implementati nella pipeline sono strutturati attraverso una verifica preliminare che blocca l'esecuzione in caso di prerequisiti non soddisfatti. Prima dell'avvio, il sistema valida l'esistenza fisica delle *directory* target, interrompendo l'esecuzione se anche una sola risultasse inaccessibile o non conforme a quella prevista. Parallelamente, viene verificata la presenza di tutti gli script essenziali per l'esecuzione della pipeline, quali `file_count.py`, `hash_calc.py`, `evangelisti_structure_generation.py`, `count_check.py` e `integrity_check.py`, con fallimento immediato se mancanti. Il sistema effettua inoltre controlli di consistenza sui suffissi associati alle *directory* (come "floppy", "hd", "hdesterno") validando che siano in formato *lowercase* e coerenti tra configurazioni di output e metadati, impedendo incongruenze che comprometterebbero la nomenclatura dei file generati. Durante l'esecuzione, ogni step implementa controlli intermedi che verificano la presenza dei file generati precedentemente, come i file adibiti ai controlli di consistenza e integrità (ad es. `FloppyDisk_CNT.json` e `HD_HASH.json`), i file contenenti la rappresentazione della struttura delle *directory* (come `structure_floppy.nq`) o i file con la rappresentazione dei metadati (ad es. `FileSystem_TechMeta_hdesterno.nq`), bloccando l'avanzamento se i risultati attesi non sono disponibili. Questa architettura di controllo crea una catena di dipendenze dove ogni fase dipende dal successo della precedente, con logging che documenta ogni controllo effettuato e il motivo di eventuali interruzioni, permettendo la diagnosi dei problemi.

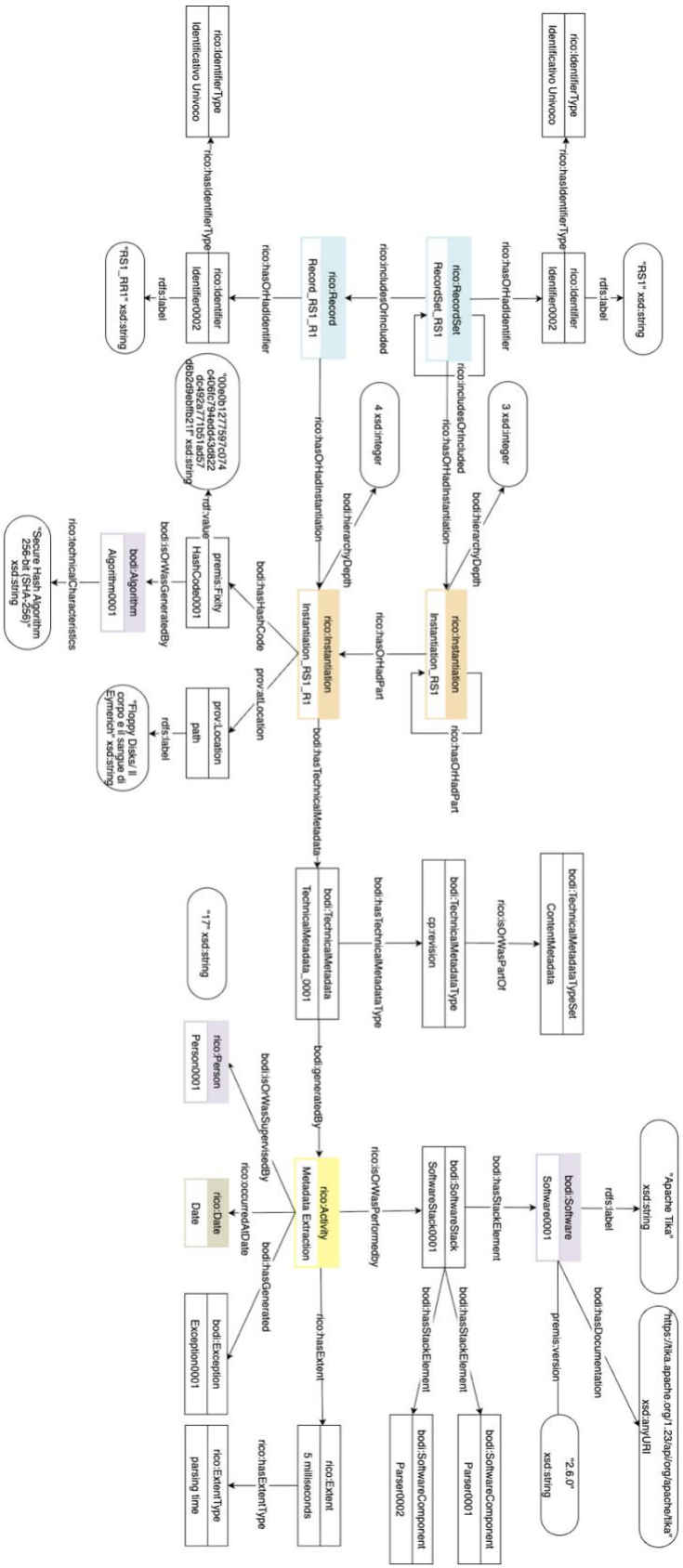


Figura 7.4. Esempio di struttura della knowledge base al termine della prima fase del workflow, che integra la rappresentazione semantica della struttura archivistica con metadati tecnici, relazioni e informazioni di provenance, ossia la baseline per le fasi successive.

7.2 Validazione e inferenze

Figura 7.5. Seconda fase del workflow: validazione e inferenze.

Questa fase opera sulla *knowledge base* generata dalla Fase 1 estraendo e rendendo esplicita la conoscenza già contenuta nei dati attraverso processi di inferenza logica controllata. La fase si distingue dalla precedente per il suo carattere puramente analitico e inferenziale: mentre la prima fase operava una trasformazione dal mondo fisico-digitale a quello semantico, la seconda fase lavora interamente nel dominio semantico per identificare pattern, eventuali inconsistenze e relazioni latenti che emergono dalla lettura sistematica del grafo.

Le attività di questa fase si suddividono in due principali processi: la validazione della logica dei grafi prodotti dalla Fase 1 e l'analisi dei grafi finalizzata all'inserimento di nuova conoscenza derivata dall'inferenza di conoscenza a partire dai dati esistenti[237]. Le attività di questa seconda fase si articolano nel seguente modo:

1. **Query di validazione della struttura, dei metadati e delle relative statistiche**.

Elaborazione ed esecuzione di *competency questions* finalizzate a verificare la validità dell'output della Fase 1 e porre le basi per l'arricchimento successivo (step 2). Le *query*, lanciate attraverso l'esecuzione di un file Python (`validation_queries.py`) che permette la connessione a Blazegraph, individuano statistiche generali e controlli di consistenza del dataset, integrità strutturale e coerenza gerarchica, analisi del sistema di metadati tecnici, consistenza e coerenza dei codici hash. Nello specifico, vengono eseguite cinquantadue *query* distinguibili fra *query* a carattere statistico/informativo e *query* di validazione:

[237] L'attività di analisi è stata condotta direttamente all'interno dell'istanza Blazegraph integrata a ResearchSpace, ambiente prescelto per la visualizzazione e la pubblicazione dei dati. I file in formato N-Quads (.nq) prodotti dalla fase 1 della pipeline sono stati caricati nel *triplestore* Blazegraph integrato in ResearchSpace mediante uno script di *bulk loading* sviluppato ad hoc da Remo Grillo disponibile nella documentazione del progetto al link: https://github.com/LuciaGiagnolini12/ValerioEvangelisti_Project. La pipeline della Fase 1 è configurabile per scrivere direttamente sull'endpoint Blazegraph di ResearchSpace; tuttavia, per esigenze gestionali di controllo si è preferito operare in una macchina di test separata per la gestione degli output. La generazione degli N-Quads, da questo punto di vista, offre grande flessibilità di sviluppo e testing in maniera agnostica rispetto all'ambiente.

Query a carattere statistico/informativo:

- *Conteggio totale delle triple RDF*. Misura la dimensione complessiva del dataset attraverso il conteggio di tutte le triple RDF, fornendo un indicatore della dimensione dell'archivio.
- *Distribuzione entità archivistiche*. Conteggia separatamente `rico:Record`, `rico:RecordSet` e `rico:Instantiation`, verificando che il bilanciamento logico-fisico sia corretto (ogni entità logica (`rico:Record` e `rico:RecordSet`) deve avere una corrispondente istanziazione fisica (`rico:Instantiation`).
- *Conteggio identificativi.* Verifica la presenza e consistenza numerica di `rico:Identifier` (identificativi univoci per ogni risorsa) e `rico:IdentifierType` (tipologie di identificativi utilizzati), essenziali per il tracciamento e la citazione delle risorse archivistiche.
- *Conteggio entità di contesto*. Enumera le `rico:Person` (persone associate alla gestione delle operazioni), `rico:Activity` (attività di processamento e gestione), fondamentali per documentare la *provenance* e la storia archivistica.
- *Conteggio di metadati tecnici e relativa tipologia*. Quantifica i `bodi:TechnicalMetadata` (metadati estratti dai file) e i `bodi:TechnicalMetadataType` (tipologie di metadati).
- *Conteggio infrastruttura software*. Verifica presenza di `bodi:Software`, `bodi:SoftwareStack`, `bodi:SoftwareComponent` e `bodi:Algorithm` utilizzati nel processamento, essenziali per la documentazione delle trasformazioni applicate.
- *Conteggio codici hash*. Enumera `premis:Fixity` (codici hash)
- *Conteggio percorsi.* Enumera `prov:Location` (percorsi di file e cartelle)
- *Conteggio estensioni*. Enumera `rico:Extent` (informazioni dimensionali di entità diverse).
- *Distribuzione dei metadati per strumento di estrazione*. Quantifica i metadati estratti da Apache Tika, ExifTool e libreria os.
- *Distribuzione delle tipologie di metadati per strumento di estrazione*. Quantifica le tipologie di metadati estratti da Apache Tika, ExifTool e libreria os e li salva su un file .csv per successive analisi e operazioni (step 2).
- *Statistiche globali sull'attività di estrazione*. Fornisce conteggi di `rico:Activity` (operazioni di estrazione), `bodi:TechnicalMetadata` (metadati estratti) e `bodi:TechnicalMetadataType`

(tipologie di metadati riconosciute), offrendo una panoramica quantitativa dei risultati del processamento.

- *Distribuzione media type.* Estrae tutti i MIME type presenti nell'archivio con statistiche dettagliate (conteggi, percentuali, ranking), fornendo un quadro sulla composizione tipologica dell'archivio e identificando formati inaspettati o problematici. I risultati sono salvati in un file .csv finalizzato a successive analisi e operazioni (step 2).
- *Individuazione di file con lo stesso codice hash*. Identifica `rico:Instantiation` con `premis:Fixity` equivalenti, rivelando file identici in maniera trasversale rispetto alle directory.

Query di validazione:

- *Analisi e validazione delle root*. Identificazione dei punti di partenza della gerarchia archivistica nella *knowledge base* e verifica che corrispondano effettivamente alle *root* dell'archivio, per escludere errori di strutturazione.
- *Rilevamento di auto-inclusioni gerarchiche* - esclude che esistano entità che includano se stesse nella relazione gerarchica (`rico:isOrWasIncludedIn`), situazione logicamente impossibile che indicherebbe errori nella generazione delle relazioni o problemi nell'algoritmo di costruzione della gerarchia.
- *Rilevamento di inclusioni gerarchiche bidirezionali.* Esclude che esistano situazioni dove due entità si includono reciprocamente (A `rico:includesOrIncluded` B e B `rico:includesOrIncluded` A), che indicherebbero errori nella determinazione della direzione gerarchica corretta.
- *Verifica della presenza di label per le varie entità*. Controlla che ogni `rico:Record`, `rico:RecordSet`, `rico:Instantiation`, `bodi:TechnicalMetadata`, `bodi:TechnicalMetadataType`, `rico:Activity`, `rico:Person`, `bodi:Software`, `bodi:SoftwareStack`, `bodi:SoftwareComponent` e `bodi:Algorithm` possiedano un'etichetta descrittiva (`rdfs:label`), indispensabile per la comprensibilità e le interfacce di ricerca e visualizzazione.
- *Validazione della profondità gerarchica*. Verifica la coerenza delle profondità calcolate per le entità archivistiche (`bodi:hierarchyDepth`) controllando che la differenza tra *parent* e *child* sia esattamente 1, identificando incongruenze che indicherebbero errori nell'algoritmo di calcolo delle profondità o relazioni gerarchiche malformate.

- *Rilevamento di entità archivistiche prive di instanziazione fisica*. Verifica che non esistano `rico:Record` e `rico:RecordSet` privi di `rico:Instantiation` corrispondenti, individuando eventuali errori nel collegamento tra rappresentazione contenutistica e logica dell'archivio.
- *Rilevamento di instanziazioni fisiche prive di collegamento con entità archivistiche*. Verifica che non esistano `rico:Instantiation` non collegate alla struttura archivistica logica, individuando eventuali errori nel collegamento tra rappresentazione logica e fisica dell'archivio.
- *Validazione della struttura logica dei files*. Verifica che i `rico:Record` non abbiano figli nella gerarchia, poiché solo i `rico:RecordSet` dovrebbero contenere altri elementi.
- *Controllo dell'unicità dei path*. Verifica che ogni percorso fisico (`prov:Location`) sia associato a una sola `rico:Instantiation`, evitando conflitti o duplicazioni.
- *Verifica della copertura dei metadati*. Controlla che ogni `rico:Instantiation` abbia metadati estratti da almeno uno strumento, segnalando eventuali fallimenti nella pipeline di estrazione.
- *Validazione delle attività di estrazione metadati.* Verifica che ogni `rico:Activity` corrispondente ad un'attività di estrazione sia associata a una `rico:Person` supervisore (responsabile dell'operazione) e a una `rico:Date` (momento di esecuzione dell'operazione).
- *Rilevamento di metadati privi di collegamenti*. Verifica l'assenza di metadati tecnici non relazionati né alla `rico:Activity` che li ha generati né alla `rico:Instantiation` a cui si riferiscono.
- *Controllo della relazione bidirezionale fra i metadati e la relativa tipologia*. Verifica che ogni `bodi:TechnicalMetadata` sia correttamente associato al proprio `bodi:TechnicalMetadataType` e viceversa, assicurando l'integrità referenziale del sistema di tipizzazione metadati.
- *Identificazione di tipologie di metadati inutilizzati.* Verifica che non ci siano `bodi:TechnicalMetadataType` definiti nel sistema ma mai effettivamente utilizzati da istanze di `bodi:TechnicalMetadata`.
- *Verifica del collegamento fra codici hash e algoritmo di generazione*. Controlla che ogni hash (`premis:Fixity`) sia correttamente collegato `bodi:Algorithm` che lo ha generato.

- *Validazione dell'applicazione dello standard crittografico SHA-256*. Verifica che il sistema utilizzi esclusivamente l'algoritmo SHA-256 per il calcolo degli hash, assicurando uniformità crittografica.
- *Controllo formato hash*. Verifica che tutti gli hash rispettino il formato SHA-256 standard (64 caratteri esadecimali), identificando eventuali hash malformati che indicherebbero errori nel calcolo o nella memorizzazione.
- *Verifica dell'unicità di hash per risorsa*. Controlla che ogni `rico:Instantiation` abbia associato esattamente un hash (né zero, ché implicherebbe mancanza di controllo di integrità, né multipli) assicurando un mapping 1:1 tra risorse e codici.

Gli outputs dello step 1 di questa seconda fase sono:

- Un report JSON completo con statistiche dettagliate per ogni categoria di validazione, con tempi di esecuzione e breakdown di eventuali errori che permette di monitorare, valutare e ottimizzare progressivamente la qualità del sistema informativo.
- Un file CSV contenente tutte le tipologie di metadati estratti.
- Un file CSV contenente tutti i Media types presenti nel fondo.

2. Applicazione di *rule-based inferencing* per l'arricchimento della knowledge base.

Questo step è finalizzato ad arricchire la *knowledge base* esplicitando conoscenza latente con una logica *data-driven*, attraverso un'analisi sistematica degli output del primo step. Le operazioni di *rule-based inferencing* consistono nel rendere esplicite relazioni già contenute nei dati, mettendo in relazione le informazioni presenti nei grafi prodotti per la rappresentazione delle tre *directory* secondo due modalità complementari: un dialogo verticale, che connette il grafo di struttura con i grafi contenenti i metadati della stessa *directory*, e un dialogo trasversale, che per la prima volta nel processo stabilisce relazioni tra grafi riferiti a *directory* diverse. Le operazioni effettuate consistono in:

- **Correlazione di file aventi lo stesso codice hash**. Lo step 1 (*query* di validazione della struttura) ha permesso di individuare file con lo stesso codice hash; dunque, file identici in termini di contenuto sia all'interno di una singola *directory* che trasversalmente fra le tre analizzate. Si è così proseguito a generare una relazione fra di essi (`bodi:hasSameHashCodeAs`) in modo tale da esplicitare questa caratteristica. Tale approccio permette di ricostruire le possibili traiettorie evolutive di un file attraverso le sue diverse istanziazioni, identificando copie di backup, processi

di riutilizzo e pattern di duplicazione che altrimenti rimarrebbero impliciti nella struttura gerarchica dei *file system*.

- **Allineamento fra metadati tecnici estratti da strumenti diversi**. Lo step 1 (*query* di validazione della struttura) ha consentito l'estrazione di tutte le tipologie di metadati tecnici (`bodi:TechnicalMetadataType`) presenti nel dataset, per un totale di 111.688 tipologie di metadati tecnici. Ciascuna tipologia è stata automaticamente associata allo strumento di estrazione che l'ha generata e alla sua frequenza di occorrenza nel dataset, permettendo un'analisi quantitativa della distribuzione dei metadati. Dall'analisi del file CSV contenente queste informazioni, sono stati selezionati i metadati più rilevanti dal punto di vista statistico, ossia quelli con almeno dieci occorrenze, dunque nella descrizione di almeno dieci file nel fondo. Ciò corrisponde ai primi 4.285 metadati tecnici in ordine di frequenza. Su questo sottoinsieme è stato condotto un processo di allineamento semantico per identificare *metadata types* che, pur presentando denominazioni diverse e provenendo da strumenti di estrazione differenti, veicolano semanticamente lo stesso tipo di informazione derivando dall'interpretazione della stessa proprietà. L'allineamento consente l'identificazione di contrasti e discrepanze tra le estrazioni operate da tool diversi. Allo stato attuale, nessun singolo estrattore automatico può garantire un'affidabilità assoluta senza un'analisi diretta della struttura esadecimale del file, e le divergenze tra gli output di strumenti diversi possono costituire un'informazione rilevante quanto le concordanze (Gorini e Giagnolini 2025). Inoltre, l'identificazione di equivalenze tra tipologie di metadati consente di unificare semanticamente informazioni distribuite tra diversi strumenti di estrazione, superando l'eterogeneità terminologica. Tale unificazione abilita interrogazioni che riconoscono automaticamente equivalenze concettuali, migliora la qualità delle inferenze successive e facilita la comparazione diretta fra valori estratti con tool diversi, trasformando la frammentazione strumentale in una risorsa di validazione incrociata e controllo qualità dei metadati estratti. L'allineamento è stato realizzato attraverso l'analisi comparativa della documentazione degli strumenti di estrazione e la verifica empirica sui valori associati alle varie tipologie di metadato. Il risultato è documentato nella Tabella 7.1, che presenta le equivalenze semantiche individuate tra metadata types provenienti da fonti diverse. L'allineamento è stato operativamente effettuato aggiornando la *knowledge base* mediante l'inserimento di relazioni `owl:sameAs` tra `bodi:TechnicalMetadataType` equivalenti, eseguito sul grafo dedicato all'arricchimento semantico previsto in questo step.

Label normalizzata	ExifTool	Apache Tika	Os Library
Denominazione	File Name	File Name	
Data di creazione del contenuto	CreateDate	dcterms:created	
Data di modifica del contenuto	MediaModifyDate	dcterms:modified	
Data di creazione del media	MediaCreateDate	Media Created Date	st_birthtime
Data di modifica del media fs	FileModifyDate	File Modified Date	st_mtime
Ultima data di accesso del media fs	FileAccessDate		st_atime
Creatore del contenuto	Creator	dc:creator	
Ultimo agente che ha modificato il contenuto	LastModifiedBy	meta:last-author	
Consistenza (size)	FileSize	File Size ; Content-Length	st_size
N. di revisioni	RevisionNumber	cp:revision	
Tempo di lavoro	TotalEditTime	extended-properties:TotalTime	
N. di parole	Words	meta:word-count	
N. di caratteri	Characters	meta:character-count	
N. di paragrafi	Paragraphs		
Software	Application	extended-properties:Application	
MIME Type	MIMEType	Content-Type	
N. di pagine	Pages	meta:page-count	
Altezza immagine	ImageHeight	tiff:ImageLength	
Larghezza immagine	ImageWidth	Image Width / tiff:ImageWidth	
Bit per campione	BitsPerSample	tiff:BitsPerSample	
Componenti colore	ColorComponents	Number of Components	

Risoluzione X	XResolution	X Resolution	
Risoluzione Y	YResolution	Y Resolution	
Durata	Duration	xmpDM:duration	
Commento	Comment	xmpDM:logComment	
Frequenza campionamento audio	AudioSampleRate	audioSampleRate	
Versione PDF	PDFVersion	pdf:PDFVersion	
Titolo	Title	dc:title	
Produttore	Producer	pdf:producer	
Strumento di creazione	CreatorTool	xmp:CreatorTool	
ID documento	DocumentID	xmpMM:DocumentID	
Data cambio inode	FileInodeChangeDate		st_ctime
Permessi file	FilePermissions		st_mode
Compressione	Compression	Compression Type	
Spazio colore	ColorSpace	Exif SubIFD:Color Space	
Orientamento	Orientation	Exif IFD0:Orientation, tiff:Orientation	
Unità risoluzione	ResolutionUnit	Resolution Units; tiff:ResolutionUnit	
Marca fotocamera	Make	Exif IFD0:Make; tiff:Make	
Modello fotocamera	Model	Exif IFD0:Model; tiff:Model	
Flash	Flash	Exif SubIFD:Flash; exif:Flash	
Numero F	FNumber; Aperture	Exif SubIFD:F-Number; exif:FNumber	
Lunghezza focale	FocalLength	Exif SubIFD:Focal Length; exif:FocalLength	

Tempo esposizione	ExposureTime	Exif SubIFD:Exposure Time; exif:ExposureTime	
ISO	ISO	Iso	
Bilanciamento bianco	WhiteBalance	Exif SubIFD:White Balance Mode	
Modalità misurazione	MeteringMode	Exif SubIFD:Metering Mode	
Modalità esposizione	ExposureMode	Exif SubIFD:Exposure Mode	
Compensazione esposizione	ExposureCompensation	Exif SubIFD:Exposure Bias Value	
Versione EXIF	ExifVersion	Exif SubIFD:Exif Version	
Lingua	LanguageCode	language	
Codifica contenuto	MIMEEncoding	Content-Encoding	
Lunghezza thumbnail	ThumbnailLength	Exif Thumbnail:Thumbnail Length	
Offset thumbnail	ThumbnailOffset	Exif Thumbnail:Thumbnail Offset	
Soggetto	Subject	dc:subject	
Commenti	Comment	w:Comments	
Nitidezza	Sharpness	Sharpness	
Altitudine globale	GlobalAltitude	Global Altitude	
Sub-secondi data originale	SubSecDateTimeOriginal	Exif SubIFD:Sub-Sec Time Original	
Sub-secondi tempo digitalizzato	SubSecTimeDigitized	Exif SubIFD:Sub-Sec Time Digitized	
Sub-secondi tempo	SubSecTime	Exif SubIFD:Sub-Sec Time	
Altezza area AF	AFAreaHeight	AF Area Height	
Larghezza area AF	AFAreaWidth	AF Area Width	

Valore luminosità	BrightnessValue	Exif SubIFD:Brightness Value	
Editore	Publisher	Dc:publisher	
Versione record applicazione	ApplicationRecordVersion	Application Record Version	
Altitudine GPS	GPSAltitude	GPS:GPS Altitude	
Riferimento altitudine GPS	GPSAltitudeRef	GPS:GPS Altitude Ref	
Timestamp GPS data	GPSDateStamp	GPS:GPS Date Stamp	
Timestamp GPS ora	GPSTimeStamp	GPS:GPS Time-Stamp	
Metodo elaborazione GPS	GPSProcessingMethod	GPS:GPS Processing Method	
Scala stampa	PrintScale	Print Scale	
Set caratteri codificato	CodedCharacterSet	Coded Character Set	
Distanza soggetto	SubjectDistance	Exif SubIFD:Subject Distance	
Frame rate video	VideoFrameRate	Video Frame Rate	
Rapporto aspetto pixel	PixelAspectRatio	Pixel Aspect Ratio	
Conteggio campioni audio	AudioSampleCount	Audio Sample Count	
Byte medi per secondo	AvgBytesPerSec	Avg Bytes Per Sec	
Numero canali	NumChannels	Num Channels	
Dimensione campione	SampleSize	Sample Size	
Conteggio stream	StreamCount	Stream Count	
Codec video	VideoCodec	Video Codec	
Conteggio frame video	VideoFrameCount	Video Frame Count	
Conteggio frame	FrameCount	Frame Count	
Velocità dati massima	MaxDataRate	Max Data Rate	
Interpretazione fotometrica	PhotometricInterpretation	Exif IFD0:Photometric Interpretation	

Campioni per pixel	SamplesPerPixel	Exif IFD0:Samples Per Pixel; tiff:SamplesPerPixel	
Compositore	Composer	xmpDM:composer	
Famiglia font	FontFamily	FontFamilyName	
Nome font	FontName-en-US	FontName	
Sottofamiglia font	FontSubfamily-en-US	FontSubFamilyName	
Sicurezza documento	DocSecurity	extended-properties:DocSecurity	
Pattern CFA	CFAPattern	Exif SubIFD:CFA Pattern	
Configurazione planare	PlanarConfiguration	Exif IFD0:Planar Configuration	
Avviso copyright	CopyrightNotice	Copyright Notice	
Offset strip	StripOffsets	Exif IFD0:Strip Offsets	
Versione Maker Note	MakerNoteVersion	Maker Note Version	
Tipo dispositivo	DeviceType	Device Type	
Rilevamento volti	Face Detect	Face Detect	
Ordine byte dati raw	RawDataByteOrder	Raw Data Byte Order	
Pattern CFA dati raw	RawDataCFAPattern	Raw Data CFA Pattern	
Formato audio	AudioFormat	Audio Format	
Codice lingua del media	MediaLanguageCode	Media Language Code	
Modalità grafica	GraphicsMode	Graphics Mode	
Predittore	Predictor	Exif IFD0:Predictor	
Cicli editing	Editing-cycles	editing-cycles	

Tabella 7.1. Allineamento tra metadata types provenienti da strumenti di estrazione diversi.

- **Suddivisione dei metadati tecnici in raggruppamenti semantici.** Dati i metadati tecnici più frequenti individuati nello step precedente, si è proseguito con un raggruppamento degli stessi in nove categorie semantiche distinte (`bodi:TechnicalMetadataTypeSet`): *file system metadata* (attributi strutturali del *file system* quali dimensioni, date di modifica e permessi), *document content*

metadata (metadati intrinseci al contenuto documentale come autore, titolo e informazioni editoriali), *image-specific metadata* (parametri tecnici delle immagini inclusi risoluzione, profondità colore e metadati EXIF), *audio-specific metadata* (caratteristiche tecniche dei file audio quali frequenza di campionamento, *bitrate* e informazioni musicali), *video-specific metadata* (proprietà specifiche dei contenuti video come *frame rate*, codec e tracce multiple), *email metadata* (intestazioni e attributi dei messaggi di posta elettronica), *executable metadata* (informazioni sui file eseguibili inclusi architettura, versioni e caratteristiche binarie), *archive metadata* (parametri dei file compressi e archivi), e *security metadata* (attributi relativi a crittografia, permessi e protezioni). Questa operazione di classificazione è finalizzata da un lato ad ampliare la semantica del singolo `rico:TechnicalMetadataType`, mettendolo in correlazione anche con quei metadati che, pur non corrispondendo ad un allineamento diretto, individuano caratteristiche affini; dall'altro, come verrà mostrato nel capitolo 6.5, questa suddivisione è finalizzata anche ad un miglioramento della fruizione in lettura dell'elenco dei metadati, che apparendo raggruppati in categorie tematicamente coerenti risultano più facilmente navigabili e comprensibili rispetto ad un'elencazione sequenziale indifferenziata. Tale approccio consente inoltre di implementare strategie di interrogazione e analisi più mirate, permettendo agli utenti di focalizzarsi su specifiche tipologie di metadati in relazione alle proprie esigenze di ricerca.

- *Assegnazione di una tipologia ai file sulla base del media type*. Lo step di validazione della struttura, dei metadati e delle relative statistiche ha permesso di estrarre tutti i *media type* presenti nell'archivio, fornendo una panoramica completa e quantificata della diversità tipologica del patrimonio digitale conservato (Tabella 7.2). I *media type*, anche noti come MIME types (Multipurpose Internet Mail Extensions), sono definiti e standardizzati nell'RFC 6838 dell'IETF, con l'Internet Assigned Numbers Authority (IANA) responsabile per tutti i tipi MIME ufficiali[238].

Media Type	Occorrenze
`image/jpeg`	17.778
`text/rtf`	9.319
`text/plain`	6.204
`image/png`	4.933
`image/svg+xml`	3.462
`application/msword`	2.833

[238] https://www.iana.org/assignments/media-types/media-types.xhtml

Media Type	Occorrenze
application/xml	2.731
text/html	1.893
application/json	1.445
application/octet-stream	1.010
application/pdf	773
application/x-gzip	751
image/vnd.fpx	679
image/gif	507
image/bmp	458
application/unknown	349
application/zip	331
audio/mpeg	269
application/rdf+xml	250
video/mpeg	133
font/woff	128
image/webp	108
application/x-font-ttf	71
application/vnd.openxmlformats-officedocument.wordprocessingml.document	60
video/x-msvideo	58
font/woff2	57
image/x-icon	49
application/epub+zip	46
image/vnd.djvu	43
video/mp4	27
video/x-ms-wmv	27
application/x-shockwave-flash	26
application/x-7z-compressed	26
image/tiff	22
audio/ogg	22
application/vnd.ms-officetheme	21
image/x-cursor	19
application/vnd.oasis.opendocument.text	14
image/x-jps	11

Media Type	Occorrenze
application/vnd.ms-excel	10
application/x-rar-compressed	8
audio/x-wav	5
application/vnd.adobe.air-application-installer-package+zip	5
application/x-bittorrent	5
application/vnd.ms-powerpoint	4
application/vnd.openxmlformats-officedocument.wordprocessingml.template	4
audio/x-pn-realaudio	4
audio/x-matroska	3
application/x-mobipocket-ebook	3
video/webm	3
image/pcx	2
application/vnd.ms-word.template.macroEnabledTemplate	2
application/x-iso9660-image	2
application/vnd.iccprofile	2
video/quicktime	2
audio/mp4	2
video/x-m4v	2
application/postscript	2
application/ResEdit	1
application/bzip2	1

Tabella 7.2. Media types presenti nell'Archivio Valerio Evangelisti e relative frequenze.

L'analisi dei dati riportati nella Tabella 7.2 ha permesso di organizzare i *media type* per categorie funzionali, al fine di individuare una classificazione tipologica dei file che ne facilitasse la lettura[239]. A ogni *media type* è stata associata un'etichetta specifica (ad esempio `audio/aac` → `Audio (AAC)`, `video/mp4` → `Video (MP4)`, `application/pdf` → `Documento (PDF)`). Sulla base di questo dizionario di mappatura è stato eseguito un aggiornamento della *knowledge base* per associare ad

[239] Le statistiche sui *media type* utilizzati in questa fase sono state impiegate in funzione dell'arricchimento della *knowledge base*. Un'analisi più approfondita delle statistiche di *media type* presenti in archivio è delineata al capitolo 8.4. In questa sede, però, vale la pena sottolineare che, laddove il metadato tecnico riferito al *media type* non sia stato estratto, al file in questione non è stata assegnata alcuna categoria.

ogni file (`rico:Instantiation`) la tipologia corrispondente, tramite l'assegnazione della proprietà `rico:type` contenente la stringa identificativa (`Video (MP4)`, `Immagine (JPG)`, etc.).
Questo approccio consente di ottimizzare la ricercabilità delle tipologie documentali attraverso query più dirette e di migliorare la loro leggibilità complessiva.

Tuttavia, è importante sottolineare anche le limitazioni intrinseche nella gestione puramente data-driven di questa operazione di classificazione. Ad esempio, il formato PDF costituisce un caso paradigmatico di ambiguità: il MIME type `application/pdf` può identificare documenti testuali strutturati, così come scansioni sostanzialmente equivalenti a file grafici, o altri contenuti non testuali. Tale indeterminatezza non può essere risolta mediante la sola analisi dei metadati di formato, indipendentemente dal livello di sofisticazione della fase di configurazione iniziale, ma richiede necessariamente un accesso diretto al contenuto per una classificazione accurata. Nonostante queste limitazioni, l'approccio rappresenta un compromesso strategico tra precisione classificatoria ed efficienza operativa, risultando vantaggioso per archivi digitali di grandi dimensioni, poiché l'analisi manuale sistematica di ogni oggetto digitale per ciascuna operazione risulterebbe temporalmente ed economicamente insostenibile. È importante sottolineare, comunque, che la curatela manuale non viene eliminata, ma strategicamente rimandata a una fase successiva, quando il lavoro computazionale massivo è già stato completato, permettendo così di beneficiare dell'elaborazione automatica preliminare. Qualora l'intervento umano rilevi successivamente discrepanze tra la classificazione automatica e l'effettivo contenuto, il file può essere facilmente riclassificato attraverso *query* di *update*. In ambienti come ResearchSpace (si veda il Capitolo 9), tale operazione si traduce potenzialmente in modifiche dirette tramite interfaccia di *data entry*.

- *Attribuzione di titolo a Record e Record Set*. Il processo di assegnazione dei titoli riflette le proprietà `rdfs:label` esistenti sui `rico:Record` e `rico:RecordSet` in entità `rico:Title` dedicate, seguendo il modello concettuale di Records in Contexts che tratta i titoli come entità separate. Il sistema identifica tutte le entità archivistiche che possiedono già una `rdfs:label` dalla fase di generazione della struttura e crea per ciascuna una nuova entità `rico:Title` con URI strutturato, replicando il valore testuale originale e stabilendo relazioni bidirezionali attraverso le proprietà `rico:hasOrHadTitle` e `rico:isTitleOf`. Questo approccio permette una rappresentazione più ricca e flessibile dei titoli archivistici, consentendo in futuro l'associazione di metadati specifici alle entità *Title*.
- *Attribuzione di date a Record, Record Sets e Instantiations*. Il processo di assegnazione delle date riflette i metadati temporali estratti in entità `rico:Date` dedicate, distinguendo semanticamente tra

date del contenuto e date del supporto fisico/digitale. In questa prospettiva, il sistema identifica e normalizza i metadati temporali presenti utilizzando per `rico:Record` e `rico:RecordSet` le date di contenuto (`dcterms:created`, `dcterms:modified` e i relativi allineamenti) che si riferiscono alla creazione e modifica del contenuto intellettuale, mentre per le `rico:Instantiation` vengono impiegate le date del *file system* (`st_mtime` e relativi allineamenti) che documentano la modifica del supporto digitale stesso. Ogni data genera un'entità `rico:Date` con URI strutturato, incorporando sia il valore standardizzato *machine-readable* in formato ISO 8601[240] (tramite la *data property* `rico:normalizedDateValue`) che la rappresentazione in linguaggio naturale (tramite la *data property* `rico:expressedDate`), e stabilendo relazioni bidirezionali con `rico:Instantiation`, `rico:Record` e `rico:RecordSet` attraverso le *object property* `rico:hasCreationDate` e `rico:hasModificationDate` e le corrispondenti inverse `rico:isCreationDateOf` e `rico:isModificationDateOf`. Questa distinzione permette di preservare la differenza concettuale tra la temporalità del contenuto archivistico e quella del suo supporto come approfondito nel capitolo 8.5.

L'esecuzione delle sei operazioni di arricchimento semantico appena descritte è gestita attraverso un unico script Python (`relations_update_graph.py`) che implementa un sistema *dual-track* con grafo dedicato e backup N-Quads. Il sistema genera automaticamente tutte le nuove triple all'interno del grafo `http://ficlit.unibo.it/ArchivioEvangelisti/updated_relations` che viene caricato direttamente nell'istanza Blazegraph di ResearchSpace, mentre parallelamente salva una copia completa in formato N-Quads con *timestamp* per garantire tracciabilità e backup dell'operazione.

Questa strategia garantisce trasparenza metodologica, riproducibilità e reversibilità delle operazioni attraverso la documentazione completa delle trasformazioni e l'isolamento delle triple inferite dal grafo di default. La gestione dell'aggiornamento in un grafo dedicato rende infatti possibile l'eliminazione selettiva delle modifiche e la rigenerazione controllata dell'intero processo di arricchimento.

Complessivamente, la Fase 2 – che comprende il processo di validazione e analisi (step 1) e di inferenza e arricchimento della *knowledge base* (step 2) – è completamente documentata e tracciabile. Ogni query SPARQL utilizzata e i relativi outputs costituiscono la descrizione della logica eseguita, permettendo di verificare esattamente come ogni conclusione sia stata raggiunta a partire dai dati disponibili.

[240] Applicato, in questo caso, fino alla componente del giorno. Cfr. https://www.iso.org/standard/70907.html.

7.3 Arricchimento della descrizione

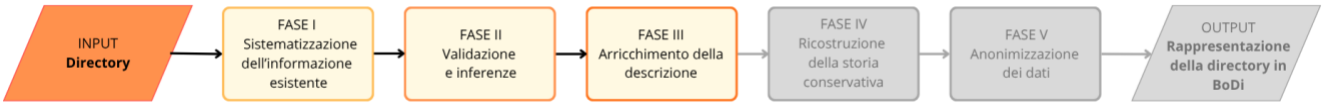


Figura 7.6. Terza fase del workflow: arricchimento della descrizione mediante conoscenza specialistica e modelli generativi.

Questa fase rappresenta un'operazione di arricchimento semantico che estende il perimetro informativo oltre la *knowledge base* esistente attraverso l'applicazione di fonti di conoscenza esterne. A differenza della Fase 2, dove la conoscenza esterna era nota a priori, qui le fonti possono essere applicate solo dopo l'estrazione dei dati: il processo non inferisce conoscenza latente già presente, ma produce conoscenza ex novo attraverso l'accesso combinato ai metadati, ai contenuti dei documenti digitali e a conoscenza esterna. L'arricchimento viene implementato attraverso due approcci diversi. Il primo procedimento, di natura semiautomatica, integra l'intervento umano specializzato con processi computazionali. Si tratta di una categorizzazione che prevede l'assegnazione di un riferimento specifico a un'opera di Evangelisti ai singoli file e cartelle (laddove possibile). Il processo viene condotto attraverso l'analisi dei contenuti documentali, dei relativi metadati e della bibliografia prodotta da Evangelisti. Tale classificazione manuale alimenta successivamente un sistema di generazione automatica di triple RDF.

Il secondo approccio, completamente automatizzato, si basa su un sistema ibrido di *Retrieval-Augmented Generation* (RAG) e *Knowledge graph question answering* (KGQA) che utilizza il modello linguistico *open source* Llama 3.2[241] per la generazione automatica di descrizioni tecniche derivate dai metadati estratti.

L'accuratezza del primo approccio dipende dalla combinazione di tre elementi fondamentali: la conoscenza specialistica della bibliografia di Evangelisti, l'analisi sistematica dei metadati e l'esame diretto dei contenuti, processo che può essere avviato esclusivamente dopo il completamento delle fasi preliminari di preservazione dell'integrità digitale e strutturazione semantica. Analogamente, l'efficacia dell'approccio automatizzato dipende dalle capacità linguistiche intrinseche del modello LLM e dalla qualità dei metadati forniti come input. In entrambi i casi, diversamente dalla fase precedente, la conoscenza generata deriva da processi non inferenziali e da conoscenze esterne non applicabili a priori. Tali procedure di arricchimento semantico si articolano operativamente attraverso specifiche attività che vengono illustrate nei paragrafi seguenti.

[241] https://ollama.com/library/llama3.2.

1. Mappatura e modellazione delle opere di Evangelisti: dall'analisi bibliografica al dato archivistico

All'interno della prospettiva archivistica contemporanea orientata alla descrizione relazionale e contestuale del patrimonio documentario, risulta oggi evidente la necessità di sperimentare modalità di rappresentazione concettuale che superino i paradigmi tradizionali (Tomasi 2022). Ad esempio, l'integrazione della descrizione archivistica con modelli come LRMoo (IFLA Library Reference Model object-oriented) consente un'apertura in questa direzione, permettendo di descrivere le entità bibliografiche come parte integrante dell'ecosistema archivistico, sia come risorse bibliografiche vere e proprie, che in termini di entità astratte in relazione ai contenuti.

Nel quadro di una tale sperimentazione, la produzione letteraria di Valerio Evangelisti e i suoi riflessi in archivio si offrono come terreno d'indagine particolarmente adatto. Prolifico autore di romanzi storici, fantascientifici, politici e narrativamente ibridi, Evangelisti ha generato nel corso della sua carriera un corpus letterario articolato in cicli, trilogie, raccolte e opere autonome, ciascuna delle quali presenta una stratificazione importante sul piano narrativo, paratestuale, contestuale e archivistico.

Sulla base delle classi definite da LRMoo, il focus è stato impostato sulla classe *Work*, definita come «distinct intellectual ideas conveyed in artistic and intellectual creations, such a poems, stories or musical compositions. A Work is the outcome of an intellectual process of one or more persons. […]» (IFLA LRMOO Working Group et al. 2024, 24). Intendiamo cioè rappresentare ogni opera narrativa come idea autoriale espressa in potenza e declinata nel tempo attraverso molteplici *Expressions*, ma comunque connotata da una propria autonomia identitaria, riconoscibile anche nei casi di revisione, ampliamento o trasposizione.

Rispetto alla produzione di Evangelisti, la mappatura si è concentrata su romanzi, racconti e cicli narrativi formalizzati come *Work* secondo la logica dell'ontologia LRMoo, con l'intento specifico di individuare e correlare, per ciascuna entità `f1_work`, i materiali d'archivio pertinenti, siano essi relativi o prodromici alla stesura (note, manoscritti, stesure integrali o parziali), contestuali alla pubblicazione (copertine, interviste, rassegne stampa) o successivi (riedizioni, traduzioni, ricezione critica).

In primo luogo, è stata effettuata una ricognizione bibliografica come fase preparatoria alla modellazione delle opere narrative di Evangelisti secondo la classe `f1_work`. Questa attività ha portato, in prima battuta, alla mappatura dei rapporti tra cicli, trilogie e i rispettivi romanzi che li compongono, come

illustrato nella Tabella 7.3[242].

Work - Ciclo/Trilogia	Work - Romanzo
Ciclo di Eymerich	Nicolas Eymerich, inquisitore
Ciclo di Eymerich	Le catene di Eymerich
Ciclo di Eymerich	Il mistero dell'inquisitore Eymerich
Ciclo di Eymerich	Il corpo e il sangue di Eymerich
Ciclo di Eymerich	Cherudek
Ciclo di Eymerich	Picatrix. La scala per l'inferno
Ciclo di Eymerich	Metallo urlante
Ciclo di Eymerich	Il castello di Eymerich
Ciclo di Eymerich	Mater Terribilis
Ciclo di Eymerich	La Sala dei Giganti
Ciclo di Eymerich	La luce di Orione
Ciclo di Eymerich	Rex tremendae maiestatis
Ciclo di Eymerich	Eymerich risorge
Ciclo di Eymerich	Il fantasma di Eymerich
Ciclo di Pantera	Metallo urlante
Ciclo di Pantera	Black Flag
Ciclo di Pantera	Antracite
Trilogia di Magus	Il presagio
Trilogia di Magus	L'inganno
Trilogia di Magus	L'abisso
Trilogia americana	Antracite
Trilogia americana	Noi saremo tutto
Trilogia americana	One Big Union

[242] Per il momento sono stati esclusi altri lavori di Evangelisti (ad esempio introduzioni alle opere, racconti in volumi collettanei, ecc.), in quanto richiedono un'analisi filologica più approfondita, che esula dallo scopo dimostrativo dell'applicazione di LRMoo. La mappatura va dunque intesa come uno strumento di lavoro in evoluzione, aperto a futuri ampliamenti e integrazioni, in collaborazione con altri esperti di dominio.

Ciclo messicano	Il collare di fuoco
Ciclo messicano	Il collare spezzato
Ciclo dei pirati	Tortuga
Ciclo dei pirati	Veracruz
Ciclo dei pirati	Cartagena
Ciclo Il Sole dell'Avvenire	Il sole dell'avvenire. Vivere lavorando o morire combattendo
Ciclo Il Sole dell'Avvenire	Il sole dell'avvenire. Chi ha del ferro ha del pane
Ciclo Il Sole dell'Avvenire	Il sole dell'avvenire. Nella notte ci guidano le stelle

Tabella 7.3. Relazioni tra cicli narrativi e romanzi nell'opera di Valerio Evangelisti.

La ricognizione ha portato alla creazione di una mappatura in cui i file sono stati associati al romanzo di riferimento; nei casi in cui un testo non risultasse chiaramente associabile ad un romanzo, ma comunque riconducibile a un ciclo specifico (principalmente *Eymerich*), l'associazione è stata per il momento attribuita solo al ciclo nel suo complesso.

A partire da queste due mappature (ciclo/trilogia-romanzo e file-romanzo) è stato sviluppato un codice Python (`works_evangelisti.py`) per tradurre le relazioni in triple RDF le cui funzioni sono sintetizzabili nello pseudocodice presentato nel Listato 7.2.

```
INIZIALIZZAZIONE dataset RDF con named graph
http://ficlit.unibo.it/ArchivioEvangelisti/works
DEFINIZIONE namespace:
  - lrmoo: per le classi F1_Work e proprietà R67
  - rico: per Record, RecordSet e proprietà isRelatedTo
  - BASE_URI_WORKS e BASE_URI_RECORDS per gli identificativi

CARICAMENTO relazioni predefinite cicli/trilogie-romanzi da file tabulare
CARICAMENTO mappature file-romanzo da file tabulare

# Fase 1: Creazione entità F1_Work e relazioni gerarchiche
PER OGNI (ciclo/trilogia, romanzo) IN relazioni_predefinite:

  // Normalizzazione URI (lowercase, rimozione caratteri speciali)
  ciclo_uri = normalizza_titolo_per_uri(ciclo)
  romanzo_uri = normalizza_titolo_per_uri(romanzo)

  // Dichiarazione entità come F1_Work
  SE ciclo_uri NON già_dichiarato:
    AGGIUNGI_AL_NAMED_GRAPH: ciclo_uri rdf:type lrmoo:F1_Work
    AGGIUNGI_AL_NAMED_GRAPH: ciclo_uri rdfs:label "Titolo originale ciclo"
    MARCA ciclo_uri come dichiarato

  SE romanzo_uri NON già_dichiarato:
```

```
      AGGIUNGI_AL_NAMED_GRAPH: romanzo_uri rdf:type lrmoo:F1_Work
      AGGIUNGI_AL_NAMED_GRAPH: romanzo_uri rdfs:label "Titolo originale romanzo"
      MARCA romanzo_uri come dichiarato

    // Relazioni gerarchiche bidirezionali
    AGGIUNGI_AL_NAMED_GRAPH: ciclo_uri lrmoo:R67_has_part romanzo_uri
    AGGIUNGI_AL_NAMED_GRAPH: romanzo_uri lrmoo:R67i_forms_part_of ciclo_uri

  #Fase 2: Collegamento con materiali d'archivio
  PER OGNI (titolo_opera, identificatore_record) IN mappature_da_xlsx:

    // Creazione URI per opera e record
    opera_uri = normalizza_titolo_per_uri(titolo_opera)
    record_uri = crea_record_uri(identificatore_record)

    // Assicura che l'opera sia dichiarata come F1_Work
    SE opera_uri NON già_dichiarato:
      AGGIUNGI_AL_NAMED_GRAPH: opera_uri rdf:type lrmoo:F1_Work
      AGGIUNGI_AL_NAMED_GRAPH: opera_uri rdfs:label "Titolo originale opera"

    // Collegamenti bidirezionali opera-file
    AGGIUNGI_AL_NAMED_GRAPH: opera_uri rico:isRelatedTo record_uri
    AGGIUNGI_AL_NAMED_GRAPH: record_uri rico:isRelatedTo opera_uri

  #Fase 3: Propagazione gerarchica: se un file è associato a un romanzo, viene
  automaticamente associato anche al ciclo/trilogia a cui il romanzo appartiene
  PER OGNI (titolo_opera, identificatore_record) IN mappature_da_xlsx:
    // Trova cicli/trilogie che contengono l'opera
    cicli = trova_cicli_contenitori(titolo_opera, relazioni_predefinite)

    PER OGNI ciclo IN cicli:
      ciclo_uri = normalizza_titolo_per_uri(ciclo)

      // Propaga collegamenti ai livelli superiori della gerarchia
      AGGIUNGI_AL_NAMED_GRAPH: ciclo_uri rico:isRelatedTo record_uri
      AGGIUNGI_AL_NAMED_GRAPH: record_uri rico:isRelatedTo ciclo_uri

  #Fase 4: Controllo duplicati e persistenza
  // Prevenzione duplicati tramite controllo esistenza in Blazegraph
  FUNZIONE aggiungi_tripla_se_non_esiste(soggetto, predicato, oggetto):
    SE NON esiste_in_blazegraph(soggetto, predicato, oggetto):
      AGGIUNGI_AL_DATASET: (soggetto, predicato, oggetto)
      RITORNA vero
    ALTRIMENTI:
      RITORNA falso

  // Upload finale
  CARICA dataset su Blazegraph tramite SPARQL UPDATE INSERT DATA
  VERIFICA caricamento con query COUNT nel named graph
  SALVA copia locale in formato N-Quads
```

Listato 7.2. Pseudocodice per la generazione del grafo RDF delle opere narrative di Valerio Evangelisti e il loro collegamento con i materiali dell'archivio.

L'esecuzione del file ha consentito, in primo luogo, di individuare le relazioni tra cicli/trilogie (`f1_work`) e romanzi (`f1_work`), tramite la proprietà `R67_has_part` e la relativa inversa `R67i_forms_part_of`; in

secondo luogo, di collegare il `rico:Record` o il `rico:RecordSet` identificato nella mappatura al relativo `f1_work` attraverso la proprietà `rico:isRelatedTo`. Grazie alla logica inferenziale, è stato così possibile stabilire un collegamento diretto anche tra i `rico:Record` e i rispettivi cicli.

Le nuove triple vengono aggiunte nel grafo dedicato `http://ficlit.unibo.it/ArchivioEvangelisti/works` e salvate in un file NQuads, coerentemente con i processi effettuati nelle precedenti fasi.

Questo approccio ha permesso di ricostruire le dinamiche di produzione e di organizzazione interna del corpus evangelistiano, offrendo una prima mappatura delle corrispondenze tra documenti d'archivio e opere letterarie. Da un lato, esso fornisce indizi utili per la ricostruzione della genesi testuale; dall'altro, consente di tracciare le successive traiettorie di ricezione e rielaborazione dell'opera, aprendo la strada a un uso archivistico insieme filologico, storico, critico e computazionale.

In prospettiva, oltre a un ampliamento della mappatura dei *Work*, un approfondimento filologico e bibliografico più mirato potrà condurre alla modellazione di entità *Expression*, utili a rappresentare in modo distinto le diverse edizioni di uno stesso romanzo, le varianti testuali, le traduzioni e gli eventuali interventi autoriali successivi alla prima pubblicazione. Ciò permetterà di rendere conto in maniera più granulare della storia editoriale e materiale delle opere.

2. Generazione automatica di descrizioni tecniche

Nel contesto della descrizione archivistica di patrimoni digitali di grandi dimensioni, emerge con chiarezza la necessità di strategie capaci di bilanciare qualità descrittiva e sostenibilità operativa. I metodi tradizionali, pur garantendo elevati standard di accuratezza, risultano difficilmente scalabili di fronte ai volumi della produzione digitale contemporanea. La descrizione manuale, infatti, richiede risorse consistenti e comporta inevitabili variazioni qualitative, con il rischio di compromettere la coerenza complessiva della documentazione prodotta.

In questo scenario, l'impiego di processi automatizzati per velocizzare la descrizione rappresenta una prospettiva sempre più rilevante, poiché consente una scansione sistematica e rapida delle collezioni digitali. Tale approccio non mira a sostituire l'*expertise* archivistica, ma a potenziarla, permettendo ai professionisti di concentrare le proprie competenze sui casi complessi e sulle decisioni strategiche, mentre i sistemi computazionali gestiscono le operazioni ripetitive e meccaniche.

Questa fase del *workflow* prevede un sistema per la generazione automatica di descrizioni tecniche sintetiche, finalizzato a facilitare la lettura e la comprensione dei metadati tecnici (comunque accessibili in forma completa e strutturata) offrendo un'anteprima che orienta la consultazione e consente di valutare

rapidamente se approfondire lo studio della risorsa mediante la scheda dettagliata.

In questo contesto, per "descrizione tecnica" si intende la descrizione delle caratteristiche del contenitore "file", ossia delle sue proprietà tecniche e strutturali, e non del contenuto informativo. La descrizione contenutistica resta affidata a una curatela specialistica, eventualmente, supportata da sistemi di IA per l'analisi testuale, fatti salvi i metadati di contenuto già presenti nel dataset (come autore, date di creazione e modifica) precedentemente analizzati.

Il sistema sviluppato implementa i paradigmi RAG e KGQA per automatizzare il processo di descrizione a partire dai metadati nativi estratti dai materiali. Introdotto originariamente da Lewis et al. (2020) nel contesto dei sistemi di elaborazione del linguaggio naturale, l'approccio RAG costituisce un'evoluzione significativa rispetto ai modelli linguistici tradizionali, combinando la capacità generativa dei LLMs con meccanismi di recupero informativo che consentono di ancorare le generazioni testuali a fonti documentali specifiche e verificabili. L'idea fondamentale è quella di migliorare i risultati di uno strumento di IA generativa tramite la ricerca in tempo reale di dati che si vanno ad aggiungere all'input, arricchendo il contesto disponibile al modello per la generazione della risposta.

Come documentato in *survey* recenti sui sistemi RAG (Gao et al. 2024), l'indicizzazione dei dati può avvenire attraverso diversi metodi, tra cui database vettoriali (l'approccio più diffuso nella pratica corrente) (Lewis et al. 2020), sistemi basati su parole chiave (TF-IDF, BM25), o grafi di conoscenza (Pan et al. 2024; Edge et al. 2024). L'implementazione qui proposta adotta un *knowledge graph* come sistema di indicizzazione per il *retrieval*, in luogo dei database vettoriali più comunemente impiegati, per tre ragioni fondamentali: la capacità di rappresentare relazioni esplicite tra entità, la garanzia di tracciabilità del processo di *retrieval*, e il controllo sui criteri di selezione dei dati da recuperare.

Le implementazioni RAG più comuni, infatti, utilizzano *vector similarity search* su *embeddings* per recuperare documenti testuali rilevanti rispetto a una query in linguaggio naturale. Nel sistema qui descritto, il *retrieval* opera invece attraverso query SPARQL su un *knowledge graph* strutturato secondo ontologie formali, recuperando metadati strutturati di risorse già identificate piuttosto che informazioni testuali da un corpus. Invece di utilizzare ricerca probabilistica su rappresentazioni del testo, il sistema impiega interrogazioni deterministiche che sfruttano le relazioni ontologiche esplicite presenti nel *knowledge graph* RDF, garantendo precisione e tracciabilità del processo di recupero informativo[243]. In

[243] In questo senso, il sistema si distingue anche da GraphRAG (Edge et al. 2024), che adotta un approccio finalizzato a costruire grafi automaticamente a partire dal testo (usando *vector similarity* per sintesi globali) poiché il grafo è già definito e viene fornito come contesto, non viene generato dal modello.

questo senso, il sistema implementato si colloca fra RAG e approcci di KGQA, che cercano di fornire risposte concrete a domande in linguaggio naturale sfruttando i grafi di conoscenza. I sistemi KGQA hanno l'obiettivo primario di tradurre una domanda in linguaggio naturale in una query formale (ad esempio SPARQL) per estrarre fatti precisi da un grafo (Guo et al. 2024; Toroghi et al. 2025). La risposta è spesso una tupla di dati o una breve frase costruita direttamente da tali tuple, con una componente generativa limitata. L'implementazione qui proposta, pur adottando il retrieval deterministico tipico dei KGQA, ne supera i limiti attraverso una componente generativa molto più ampia. Il sistema non si limita a estrarre e presentare i metadati tecnici (ad esempio, formato: JPEG, dimensione: 5MB), ma li interpreta e sintetizza in un discorso narrativo coerente e umanamente leggibile. In altre parole, il grafo funge da base di conoscenza strutturata e verificabile per un modello generativo, il quale ha il compito di produrre una *descrizione*, non solo una *risposta*.

Il cuore del sistema è il modello AI Llama 3.2 3B[244], eseguito localmente tramite il software Ollama. Come evidenziato nel capitolo 5.3.2, optare per questa modalità invece di soluzioni API cloud (come quelle offerte da OpenAI[245] o Anthropic[246]), risponde a esigenze di privacy, sicurezza e controllo sui dati. L'esecuzione locale garantisce maggior controllo sulle informazioni, riducendo al minimo il rischio di dispersione o trasferimento verso infrastrutture esterne dei dati. Inoltre, Llama è una famiglia di modelli *open source* con pesi aperti, preferibile in contesti accademici per garantire trasparenza e riproducibilità della ricerca (Touvron et al. 2023; White et al. 2024).

Il sistema è strutturato in quattro fasi principali:

1. **Retrieval strutturato.** Per ogni `rico:Instantiation`, una query SPARQL recupera tutti i metadati tecnici (`bodi:TechnicalMetadata`) associati. Il recupero è deterministico: i metadati ottenuti corrispondono esattamente a quelli associati alla risorsa interrogata. I metadati estratti vengono selezionati e forniti al LLM in ordine di priorità, ossia organizzati per garantire che il modello interpreti prima i dati più rilevanti (identificativi minimi e strutturali) costruendo un contesto semantico specifico per la generazione.
2. **Generazione controllata**. Il LLM interpreta il contesto e produce una descrizione sintetica sulla base del seguente *prompt*: «Based on these technical metadata, write a concise description of this digital file for archival purposes. File: [percorso_file] [metadati]. Describe in 2-3 sentences: file

[244] https://huggingface.co/meta-llama/Llama-3.2-3B.
[245] https://platform.openai.com/docs/api-reference/.
[246] https://www.claude.com/platform/api.

type, format, key technical properties, and creation/modification details. Only include information that is available in the metadata»[247]. Il risultato, per ciascuna `rico:Instantiation`, viene formalizzato come istanza di `bodi:TechnicalDescription` e collegata alla `rico:Instantiation` tramite la *object property* `bodi:hasTechnicalDescription` (e relativa inversa, `bodi:isTechnicalDescriptionOf`).

3. **Documentazione della *provenance***. Il sistema genera per ogni descrizione un'entità `rico:Activity` che rappresenta il processo di generazione, collegandola temporalmente a un'entità `rico:Date` condivisa dall'intera sessione di lavoro, espressa sia in formato normalizzato ISO 8601 sia in forma leggibile. La `rico:Activity` è inoltre connessa a un'entità `bodi:Software` che identifica il modello LLM utilizzato [248] . Ogni Software include automaticamente un collegamento alla documentazione ufficiale tramite `bodi:hasDocumentation`.

   La proprietà `bodi:hasHumanValidation`, inizialmente impostata su "false", permette di tracciare eventuali revisioni manuali successive, distinguendo le descrizioni automatiche da quelle validate. L'entità `rico:Person` identifica infine il supervisore umano responsabile del controllo qualità sul processo automatizzato, collegandosi alla `rico:Activity` attraverso la proprietà `bodi:hasOrHadSupervisor`.

   In questo modo, ogni descrizione può essere ricondotta al processo specifico, al software, al supervisore e al momento preciso di generazione.

4. **Integrazione con la knowledge base.** Le descrizioni generate e le relative informazioni di *provenance* (attività, software, supervisore umano, data) vengono integrate nel grafo RDF attraverso una strategia di inserimento incrementale: ogni cento descrizioni prodotte (valore configurabile), il sistema effettua automaticamente un inserimento *batch* nel *triplestore*, riducendo il rischio di perdita di dati in caso di interruzione e distribuendo il carico

[247] Il modello opera con tre parametri di generazione: temperature=0.3 riduce la variabilità nella selezione dei *token* (le unità lessicali elementari: parole, frammenti di parole, punteggiatura) privilegiando scelte ad alta probabilità per garantire omogeneità descrittiva; num_predict=150 limita la lunghezza massima dell'output a 150 *token* (circa 2-3 frasi); top_p=0.9 applica *nucleus sampling*, tecnica che ad ogni passo generativo considera solo i *token* il cui peso probabilistico cumulativo raggiunge il 90%, escludendo le scelte lessicali a probabilità residuale che, pur grammaticalmente valide, risultano semanticamente poco pertinenti o stilisticamente incongruenti. Il processamento procede in *batch* da 10 `rico:Instantiation` con intervalli di 2 secondi per prevenire saturazione del servizio API.

[248] Il sistema mantiene una cache persistente di entità `bodi:Software` per evitare duplicazioni, riutilizzando la stessa istanza per tutte le descrizioni generate dallo stesso modello anche attraverso esecuzioni multiple.

computazionale nel tempo. Dopo ogni inserimento, viene aggiornato un file di checkpoint che registra le `rico:Instantiation` già processate, consentendo la ripresa automatica e prevenendo duplicazioni in esecuzioni successive.

L'inserimento nel *triplestore* avviene nel *named graph* dedicato `http://ficlit.unibo.it/ArchivioEvangelisti/ai_descriptions`, preservando modularità in continuità con gli step precedenti del workflow. Anche in questa fase, il sistema genera contestualmente un file N-Quads che replica l'intero contenuto del grafo, fornendo una copia persistente utilizzabile per backup, analisi o migrazione.

Per esemplificare il funzionamento del sistema di generazione automatica delle descrizioni, consideriamo il caso del file *05 - Notte di attesa.rtf*., un documento di lavoro relativo alla stesura di Tortuga (Mondadori, 2008), uno dei romanzi del Ciclo dei Pirati. Quando il sistema interroga il *triplestore* tramite query SPARQL, recupera i metadati tecnici estratti associati al file visibili nel Listato 7.3.

L'interpretazione dei metadati presentati nel Listato 7.3 non è immediata per un utente non specialista. La presenza di valori apparentemente duplicati attraverso diversi *namespace* (`dc:creator` e `Author`, entrambi con valore "Valerio") insieme a un `Last-Author` ("Valerio Evangelisti") non permette di comprendere la differenza fra i due profili. Valori numerici come `st_size`: 100735 o `st_mtime` 1.217299405E9 richiedono conversioni per essere valutati; *timestamp* in formato ISO come `CreateDate`: 2008:02:15 12:00:00 e `ModifyDate`: 2008:07:29 04:43:00 devono essere contestualizzati come arco di lavoro.

```
Author: Valerio
Characters: 9057
CharactersWithSpaces: 10775
Content-Length: 100735
Content-Type: application/rtf
CreateDate: 2008:02:15 12:00:00
Dc:creator: Valerio
FileType: RTF
InternalVersionNumber: 32895
LastModifiedBy:Valerio Evangelisti
ModifyDate: 2008:07:29 04:43:00
Pages: 5
RevisionNumber:33
st_mtime 1.217299405E9
St_size: 100735
Title: 3
TotalEditTime: 6.0 hours
Words: 1739
```

Listato 7.3. Esempio di metadati tecnici estratti da un file RTF.

L'architettura implementata affronta questa complessità organizzando i metadati secondo criteri di priorità e trasmettendoli al modello Llama 3.2 3B insieme al *prompt* con la richiesta di generazione della descrizione. Il modello, elaborando il contesto strutturato, produce una descrizione di questo tipo:

> Rich Text Format (RTF) document, 5 pages with 1,739 words. Created by Valerio in February 2008, revised through 33 versions with 6 hours total editing time, last modified by Valerio Evangelisti in July 2008. Titled "3", likely a chapter or section of a larger work. File size 98.4 KB.

Questa sintesi opera diverse trasformazioni interpretative e assolve a una duplice funzione. In primo luogo, offre un'anteprima sintetica che velocizza le operazioni di consultazione: l'utente può valutare rapidamente se il documento corrisponde alle proprie esigenze di ricerca senza dover necessariamente accedere ai metadati completi. In secondo luogo, consente un'interpretazione più agevole dei metadati, traducendo etichette tecniche non immediatamente comprensibili in informazioni contestualizzate e mettendo in relazione dati frammentati. In questo modo, l'informazione diventa più accessibile anche a utenti privi di competenze specialistiche, senza compromettere la disponibilità dei metadati completi in forma strutturata.

Il processo operativo descritto è stato implementato attraverso uno script Python (`ai_technical_descriptions.py`) che gestisce l'intero processo, dalla query SPARQL al

triplestore fino alla generazione e integrazione delle descrizioni. Il Listato 7.4 presenta lo pseudocodice che sintetizza la logica implementativa, evidenziando le quattro fasi sequenziali del processo e i meccanismi di persistenza, *cache* e *checkpointing*[249].

```
INIZIALIZZAZIONE sistema:
  - Endpoint Blazegraph, Client Ollama (llama3.2)
  - File contatori URI e checkpoint per stato persistente
  - Named graph: http://ficlit.unibo.it/ArchivioEvangelisti/ai_descriptions
  - Parametri LLM: temperature=0.3, num_predict=150, top_p=0.9

CARICAMENTO stato persistente:
  - Contatori (ai_text_counter, activity_counter, software_counter)
  - Cache software: {nome_modello → URI_software}
  - Checkpoint: set(URI instantiation processate)

# Fase 1: Retrieval strutturato
PER OGNI pagina (page_size=100):
  query = "SELECT ?instantiation WHERE { ?instantiation rdf:type
rico:Instantiation }"

  PER OGNI instantiation:
    SE instantiation IN checkpoint: SALTA

    metadati = query_sparql("SELECT metadati tecnici per instantiation")
    AGGIUNGI a lista_da_processare

# Fase 2: Augmentation e Generazione
  PER batch IN dividi_in_batch(lista_da_processare, batch_size=10):
  PER instantiation IN batch:

    // Organizza metadati per priorità
    contesto = [
      identificativi (MIME, size, hash),
      strutturali (width, height, duration),
      descrittivi (creator, dates)
    ] // Limita a 3000 caratteri

    // Genera descrizione
    prompt = "Based on these metadata... {contesto}. Describe in 2-3 sentences..."
    risposta = ollama.generate(prompt, temperature=0.3, num_predict=150,
    top_p=0.9)

    // Normalizza output
    descrizione = normalizza(risposta) // spazi, virgolette, punteggiatura
```

[249] Nel *workflow* implementato, le operazioni di generazione automatica delle descrizioni tecniche qui descritte sono state effettuate a valle del processo di anonimizzazione (step 5), anziché in sequenza immediata all'estrazione dei metadati. Questa scelta risponde a esigenze di ottimizzazione delle risorse computazionali: data la mole di documenti da processare e l'*effort* richiesto dall'utilizzo locale del modello Llama, eseguire la generazione su un dataset già ridotto consente di concentrare le risorse sui soli file che saranno effettivamente resi disponibili, riducendo i tempi di elaborazione e garantendo coerenza tra le descrizioni generate e i dati accessibili. Per dataset di minori dimensioni, è suggeribile effettuare questa operazione precedentemente all'anonimizzazione, anche per permettere all'archivista di avere chiavi di accesso agevolate su tutto il dataset prima delle operazioni di anonimizzazione.

```
    // Crea URI
    ai_text_uri = BASE_URI + formatta_contatore(++ai_text_counter, 6)
    AGGIUNGI a descrizioni_generate
    AGGIUNGI a checkpoint_set

# Fase 3: Creazione provenance
  PER descrizione IN descrizioni_generate:

  // Software (con cache)
  SE model NOT IN cache:
    software_uri = BASE_URI + formatta_contatore(++software_counter, 4)
    cache[model] = software_uri

  // Activity
  activity_uri = BASE_URI + formatta_contatore(++activity_counter, 4)

  // Entità condivise (sessione)
  date_uri = genera_date_uri_sessione()

  // Triple RDF
  triple = [
    (ai_text_uri, rdf:type, bodi:TechnicalDescription),
    (ai_text_uri, rdf:value, descrizione),
    (ai_text_uri, bodi:hasHumanValidation, "false"),
    (ai_text_uri, bodi:generatedBy, activity_uri),
    (instantiation_uri, bodi:hasTechnicalDescription, ai_text_uri),
    // Activity, Software, Date...]
  AGGIUNGI triple a buffer

# Fase 4: Integrazione incrementale
  OGNI 100 descrizioni (~600 triple):
  inserisci_in_blazegraph(buffer, chunk_size=1000)
  salva_checkpoint(checkpoint_set)
  salva_contatori()
  buffer = []

// Inserimento finale e export
   inserisci_triple_rimanenti()
   esporta_nquads("ai_descriptions_" + timestamp() + ".nq")

# Funzioni supporto
  FUNZIONE inserisci_in_blazegraph(triple_list, chunk_size):
  PER chunk IN dividi(triple_list, chunk_size):
    query = "INSERT DATA { GRAPH <...> { {chunk} } }"
    esegui_sparql_update(query)
    ATTENDI 0.5s

  FUNZIONE salva_contatori():
    backup(file_esistente)
    scrivi_json({contatori, cache, timestamp})

  FUNZIONE salva_checkpoint(processate):
    backup(file_esistente)
    scrivi_json({processate, timestamp})
```

Listato 7.4. Pseudocodice per la generazione automatica di descrizioni tecniche tramite RAG con modello LLM locale, gestione cache persistente e inserimento incrementale nel triplestore.

7.4 Ricostruzione della storia conservativa

Figura 7.7. Quarta fase del workflow: ricostruzione della storia conservativa.

La documentazione della storia conservativa dei materiali digitali costituisce un requisito fondamentale per la loro comprensione. Nel caso dell'Archivio Valerio Evangelisti si segnala l'assenza di un inventario pregresso dei supporti fisici, così come l'assenza di un *preservation plan*. Questa condizione ha richiesto lo sviluppo di un approccio che integrasse il censimento iniziale dei supporti con la loro modellazione in BoDi, finalizzato a documentare i passaggi conservativi subiti dai materiali all'interno della *knowledge base*.

Sulla base delle risorse prese in considerazione dalla *pipeline*, sono stati identificati i seguenti supporti:

- un hard disk contenente la copia dei contenuti un PC Windows di Valerio Evangelisti conservato dagli eredi;
- un hard disk esterno di Valerio Evangelisti conservato dall'Associazione;
- 93 floppy disk 3.5" conservati presso l'Associazione.

Per ciascun supporto è stata documentata una serie di informazioni essenziali: tipologia del medium, provenienza, ubicazione e condizioni di conservazione. Questi dati, sebbene minimi, hanno costituito la base informativa per la creazione delle entità rappresentanti i supporti fisici e le rispettive localizzazioni, che sono state poi collegate alle istanziazioni digitali generate dal workflow di processamento.

A partire da questo censimento, la documentazione della storia conservativa del supporto fisico è realizzata in BoDi tramite l'integrazione delle classi PREMIS `premis:StorageMedium` e `premis:StorageLocation`, collegate tra loro mediante la *object property* `premis:medium`. BoDi introduce, inoltre, le proprietà simmetriche `bodi:hasStorageLocation` e `bodi:isStorageLocationOf`, che collegano `rico:Instantiation` a `premis:StorageLocation`.

La classe `premis:StorageMedium` rappresenta la natura fisica del supporto di memorizzazione, consentendo di specificare se si tratta, ad esempio, di hard disk, SSD, nastri magnetici o altri tipi di medium. Ogni istanza di `premis:StorageMedium` è caratterizzata da una label descrittiva che identifica il tipo specifico di supporto (ad esempio, "Western Digital WDBACW0020HBK-01, My Book 2TB USB 3.0 Series External Hard Drive", "Samsung Portable SSD T7 1 TB USB Type-C 3.2 Gen 2 hard disk").

La classe `premis:StorageLocation` documenta invece l'ubicazione fisica o logica in cui i materiali sono conservati. Nel caso dell'Archivio Evangelisti, le ubicazioni identificate comprendono:

- la sede dell'Associazione Valerio Evangelisti - Il Sol dell'Avvenire (Castel D'Aiano, Bologna);
- l'abitazione degli eredi di Evangelisti (Castel D'Aiano, Bologna);
- il Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna, ADLab, Armadio 1 (per i supporti intermedi durante il processamento) e Server Room (per le istanziazioni finali conservate su server) (Bologna).

Per documentare la *chain of custody*, BoDi prevede la possibilità di stabilire relazioni di derivazione tra istanze di `rico:Instantiation`, che rappresentano gli insiemi di file contenuti nei diversi supporti fisici e che vengono gestiti, copiati o trasformati nel corso del processo conservativo. A tale scopo viene utilizzata la classe `rico:DerivationRelation`, che collega due istanze di `rico:Instantiation` (una *source* e una *target*) indicando che la seconda deriva dalla prima, ad esempio in seguito a un'operazione di copia.

Per ogni supporto censito, BoDi prevede quindi la creazione delle corrispondenti `rico:Instantiation`, collegate tra loro tramite le relazioni di derivazione, che descrivono le operazioni di copia, migrazione o riversamento effettuate tra i diversi supporti. In questo modo è possibile rappresentare, per ciascun medium, la sequenza delle copie successive, distinguendo tra istanziazioni originali, intermedie e finali. Le relazioni di derivazione possono essere contestualizzate nel tempo mediante entità `rico:Date`, collegate alla relazione attraverso la proprietà `rico:occurredAtDate`, mentre la partecipazione degli attori coinvolti nelle operazioni di copia o migrazione può essere documentata associando alla relazione le corrispondenti istanze di `rico:Agent`.

Nello specifico caso dell'Archivio Evangelisti, i primi passaggi ricostruiti riguardano i materiali estratti dal PC Windows che Evangelisti usava come stazione di lavoro principale. La catena conservativa comprende:

1. **Istanziazione originale** (`Orig_Inst_RS1_RS1`): conservata nell'abitazione degli eredi, identificata come "Unknown Windows PC" data l'impossibilità di recuperare le specifiche tecniche complete del sistema (v. capitolo 3.3).
2. **Istanziazione derivata intermedia** (`Der_Inst_RS1_RS1`): copia su SSD Samsung recuperata dall'Associazione Valerio Evangelisti a partire da PC e realizzata nell'Agosto 2024.

3. **Istanziazione finale** (`RS1_RS1_inst`): copia derivata dai materiali della SSD Samsung effettuata su server del FICLIT per il processamento dei dati (ultima copia realizzata nell'Agosto 2025).

Anche per l'hard disk esterno Western Digital si documentano tre fasi distinte:

1. **Istanziazione originale** (`Orig_Inst_RS1_RS2`): supporto originale custodito presso l'Associazione, identificato come Western Digital WDBACW0020HBK-01, My Book 2TB USB 3.0 Series External Hard Drive.
2. **Istanziazione derivata intermedia** (`Der_Inst_RS1_RS2`): copia su SSD Samsung recuperata dall'Associazione Valerio Evangelisti a partire da PC e realizzata nell'Agosto 2024.
3. **Istanziazione finale** (`RS1_RS2_inst`): copia derivata dai materiali della SSD Samsung effettuata su server del FICLIT per il processamento dei dati (ultima copia realizzata nell'Agosto 2025).

Come per gli altri supporti, anche per i floppy disk dell'Archivio Evangelisti il processo di documentazione prevede tre fasi di istanziazione; a differenza degli altri casi, però, ogni fase è replicata per ciascun singolo floppy disk, generando un'istanza originale distinta per ciascun supporto:

1. **Istanziazioni originali** (`Orig_Inst_RS1_RS3_RSXX`): una per ciascun floppy disk preso in considerazione, corrispondente ai file presenti sul supporto fisico originale conservato presso l'Associazione Valerio Evangelisti - Il Sol dell'Avvenire.
2. **Istanziazioni derivate intermedie** (`Der_Inst_RS1_RS3_RSXX`): una per ogni istanziazione originale, realizzata mediante copia su SSD Samsung utilizzata come supporto di lavoro intermedio durante le operazioni di acquisizione e riversamento; ciascuna istanziazione derivata corrisponde a una cartella individuale contenente i file estratti da un singolo floppy, e tutte le cartelle sono riunite all'interno di un'unica macro-cartella sull'SSD, creata per la gestione unitaria del corpus.
3. **Istanziazione finale** (`RS1_RS3_RSXX_inst`): copia della macro-cartella sull'SSD e di tutti i suoi contenuti, riversata su server del FICLIT per il processamento e la conservazione a lungo termine.

Per formalizzare questa mappatura in triple RDF è stato elaborato un codice Python le cui funzionalità sono riassunte dal Listato 7.5.

```
INIZIALIZZAZIONE ChainOfCustodyGenerator CON:
  - named_graph "chain_of_custody"
  - namespaces PREMIS, RiC-O, BODI
  - endpoint Blazegraph
  - contatori ID univoci e cache date
```

```
# Creazione entità storage
FUNZIONE add_storage_location(descrizione_ubicazione, tipo_supporto):
  storage_location_uri = genera_id(“StorageLocation”)
  storage_medium_uri = genera_id("StorageMedium")

  CREAZIONE_TRIPLE:
    storage_location_uri a premis:StorageLocation
    storage_location_uri rdf:value descrizione_ubicazione
    storage_location_uri premis:medium storage_medium_uri
    storage_medium_uri a premis:StorageMedium
    storage_medium_uri rdfs:label tipo_supporto

# Documentazione derivazioni
FUNZIONE add_derivation(istanziazione_source, istanziazione_target,
info_temporale):
  derivation_uri = genera_id(“Derivation”)
  CREA_TRIPLE:
    derivation_uri a rico:DerivationRelation
    derivation_uri rico:relationHasSource istanziazione_source
    derivation_uri rico:relationHasTarget istanziazione_target
    derivation_uri rico:occurredAtDate add_date(info_temporale)

# Generazione chain completa
FUNZIONE create_chain_of_custody(sequenza_fasi_conservative):
  lista_istanziazioni = []

  PER OGNI fase IN sequenza_fasi_conservative:
    (location, medium) = add_storage_location(fase.storage)
    CREAZIONE_TRIPLE_BIDIREZIONALI:
      fase.instantiation bodi:hasStorageLocation location
      location bodi:isStorageLocationOf fase.instantiation
    lista_istanziazioni.aggiungi(fase.instantiation)

# Generazione automatica delle instantiation originali per i floppy
FUNZIONE generate_floppy_chains():
  recordsets = QUERY_SPARQL {
    ?recordset a rico:RecordSet ;
               rico:isOrWasIncludedIn <RS1_RS3> ;
               rdfs:label ?label .
  }
  PER OGNI recordset IN recordsets:
    chain = [
      {uri: “Orig_Inst_” + id, storage: {Associazione, “Floppy 3.5\” – “ +
label}},
      {uri: “Der_Inst_” + id, storage: {ADLab Cabinet, Samsung SSD}},
      {uri: id + “_inst”, storage: {ADLab Server}}
    ]
    create_chain_of_custody(chain)

  # Persistenza dual-track
  SALVA_LOCALE(“chain_of_custody_” + timestamp + “.nq”)
  CARICA_BLAZEGRAPH(named_graph, formato='trig', encoding=’utf-8’)
  VERIFICA_COUNT_TRIPLE(named_graph)
```

Listato 7.5. Pseudocodice per la documentazione della chain of custody e delle relazioni di derivazione tra istanziazioni dei supporti dell'archivio.

7.5 Anonimizzazione dei dati

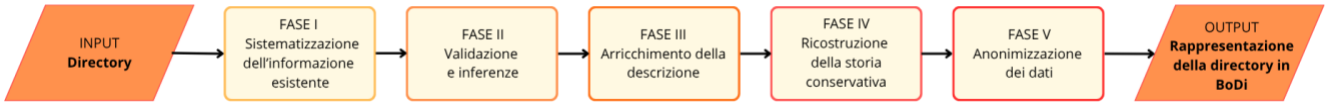


Figura 7.8. Quinta fase del workflow: anonimizzazione dei dati.

L'ultima fase del workflow è dedicata alla preparazione dei dati per la pubblicazione. Il sistema si basa su una logica di anonimizzazione selettiva che distingue tra diverse tipologie di risorse archivistiche, applicando regole specifiche per ciascuna categoria.

La gestione dell'accesso agli archivi letterari contemporanei solleva complesse questioni etiche che investono il confine tra interesse pubblico della ricerca e tutela della riservatezza delle persone coinvolte nella produzione documentale (Smith et al. 2021; Note 2025). La presenza di corrispondenza privata, documentazione personale e riferimenti a soggetti viventi non costituisce una novità per gli archivi di persona, ma la preservazione del digitale introduce elementi di complessità inediti: da un lato, la quantità di materiali prodotti è incomparabilmente maggiore rispetto al passato; dall'altro, l'urgenza di intervenire tempestivamente per contrastare l'obsolescenza tecnologica impone di confrontarsi con documentazione spesso recentissima, per la quale non si sono ancora consolidati i tempi di sedimentazione e selezione tipici delle pratiche archivistiche tradizionali. A questi fattori si aggiunge una peculiarità specifica degli oggetti digitali: i dati personali non si limitano ai contenuti espliciti (mittenti, destinatari, contenuti testuali, ecc.) ma si sedimentano anche nei metadati dei file. Informazioni di sistema, tracce di modifica, percorsi di archiviazione possono rivelare ulteriori informazioni sensibili sulla sfera privata e professionale dell'autore e dei suoi interlocutori. Il bilanciamento tra l'esigenza di preservare l'integrità informativa del fondo e la necessità di proteggere dati sensibili risulta quindi particolarmente delicato. Questa tesi non ha l'obiettivo di fornire soluzioni definitive a tali problematiche, che richiederebbero competenze interdisciplinari in ambito giuridico, archivistico ed etico, oltre a tempi di analisi ben più estesi, ma avanza una proposta metodologica che tenta di operare in presenza di tali vincoli, cercando di minimizzare l'impatto delle restrizioni d'accesso sulla futura fruibilità della risorsa

L'approccio adottato per l'anonimizzazione si basa sulla preservazione strutturale del grafo semantico: le entità non vengono rimosse dal sistema, ma le loro etichette descrittive e le informazioni personali[250]

[250] Dove per informazioni personali si intende qualsiasi dato che identifichi o renda identificabile una persona fisica secondo il Codice in materia di protezione dei dati personali (Decreto Legislativo 30 giugno 2003, n. 196, come modificato dal Decreto Legislativo 10 agosto 2018, n. 101 di adeguamento al GDPR) e il Regolamento (UE) 2016/679 del Parlamento europeo e del

collegate vengono sistematicamente sostituite con la dicitura standardizzata "Redacted information"[251]. Questa metodologia consente di mantenere intatta la struttura relazionale dell'archivio, preservando i collegamenti gerarchici e le associazioni trasversali tra entità, pur impedendo l'accesso diretto ai contenuti sensibili. La strategia garantisce che i ricercatori possano comprendere l'organizzazione generale del fondo e condurre analisi strutturali senza compromettere la riservatezza dei materiali non autorizzati.

Il sistema di anonimizzazione opera su quattro tipologie di entità: `rico:Record`, `rico:RecordSet`, `rico:Instantiation` e `bodi:TechnicalMetadata`.

Per i `rico:Record` viene adottata una strategia di protezione secondo la quale tutti i record sono anonimizzati di default, ad eccezione di quelli che soddisfano specifici criteri di visibilità. In accordo con gli eredi e l'Associazione Valerio Evangelisti, detentrice dei materiali, è stato definito un perimetro di accessibilità limitato ai dati relativi ai romanzi di Valerio Evangelisti e ai materiali di lettura e autodocumentazione. Un record rimane visibile quando è collegato tramite la proprietà `rico:isRelatedTo` a entità di tipo `lrmoo:F1_Work`, indicando una correlazione diretta tra il documento d'archivio e una specifica opera narrativa di Evangelisti mappata secondo il modello LRMoo, oppure quando è presente nella lista esplicita delle entità autorizzate alla visualizzazione (*white list*), indipendentemente dalla presenza di collegamenti diretti alle opere letterarie, o ancora quando è contenuto tramite la proprietà `rico:isOrWasIncludedIn` in un `rico:RecordSet` presente nella lista autorizzata, beneficiando della protezione estesa per propagazione gerarchica.

L'anonimizzazione dei `rico:RecordSet` segue una logica opposta, privilegiando la visibilità per favorire la ricerca e la comprensione della struttura archivistica. Un `rico:RecordSet` viene anonimizzato solamente quando è esplicitamente incluso in una lista di esclusione (*black list*), costruita sulla base dell'identificazione di cartelle contenenti materiali strettamente privati o cartelle con denominazioni che rivelano informazioni personali.

Le `rico:Instantiation` seguono automaticamente lo stato di riservatezza dell'entità archivistica a cui sono collegate attraverso la proprietà `rico:hasOrHadInstantiation`. Quando un `rico:RecordSet` o `rico:Record` viene sottoposto ad anonimizzazione, tutte le sue rappresentazioni fisiche subiscono il

Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati (GDPR).

[251] "Data redaction" o "Information redaction" sono i termini tecnici utilizzati in lingua inglese per individuare le operazioni di anonimizzazione finalizzate alla protezione dei dati personali (The National Archives 2016).

medesimo trattamento, garantendo coerenza nella protezione dei dati e prevenendo la divulgazione indiretta di informazioni attraverso i livelli di `rico:Instantiation`.

Il sistema implementa inoltre una protezione differenziata per i metadati: anche quando le rappresentazioni fisiche sono sottoposte ad anonimizzazione, vengono mantenuti selettivamente visibili i metadati di natura tecnico-informatica non sensibile, come le tipologie di formato, le dimensioni dei documenti, i programmi utilizzati per la creazione, e le informazioni temporali, permettendo così di condurre analisi quantitative automatizzate e studi di *distant reading* pur garantendo la protezione delle informazioni riservate. Tutti gli altri metadati vengono anonimizzati per prevenire la divulgazione di informazioni anche solo potenzialmente sensibili. Dall'altro lato, per le rappresentazioni fisiche che rimangono visibili, qualora nei metadati associati vengono identificate informazioni personali riferite a terze parti, il sistema procede comunque all'anonimizzazione selettiva del singolo campo problematico, preservando la visibilità di tutti gli altri elementi non compromessi. Questa doppia strategia garantisce il massimo livello di accessibilità per la ricerca compatibile con i requisiti di riservatezza, operando attraverso un controllo granulare che distingue tra il valore informativo tecnico-scientifico dei metadati e il loro potenziale di compromissione della privacy.

Il sistema relazionale rimane intatto e vengono eventualmente sovrascritte solo le stringhe informative laddove necessario. Come elemento di novità del grafo, viene aggiunta a tutti i `rico:Record` e `rico:RecordSet` la *data property* `bodi:redactedInformation` con valore "yes" se l'entità è stata anonimizzata, con valore "no" se l'entità è stata preservata nella sua forma originale (con l'eventuale eccezione di singoli metadati sensibili). Questa marcatura esplicita consente di tracciare in modo trasparente e interrogabile lo stato di accessibilità di ciascuna risorsa archivistica, facilitando le operazioni di ricerca e filtraggio.

Prima di procedere all'applicazione delle politiche di anonimizzazione, il sistema esegue automaticamente un salvataggio completo della *knowledge base*, creando una copia di sicurezza che consente di ripristinare lo stato precedente qualora si rendesse necessario modificare i parametri di anonimizzazione. Questa procedura di salvaguardia preventiva garantisce la reversibilità completa del processo e la possibilità di adattare le strategie di riservatezza in base all'evoluzione dei requisiti di ricerca o delle normative sulla protezione dei dati.

Il sistema include meccanismi di controllo qualità che verificano la coerenza dell'anonimizzazione applicata attraverso interrogazioni SPARQL per verificare potenziali inconsistenze come, ad esempio

entità `rico:Record` o `rico:RecordSet` anonimizzate con `rico:Instantiation` che mantengono ancora la *label* originale.

Questa architettura consente di definire politiche di riservatezza specifiche per diverse tipologie di contenuto archivistico, permettendo di allargare o restringere il campo delle informazioni visibili a seconda delle esigenze. Inoltre, la distinzione metodologica adottata tra `rico:Record` e `rico:RecordSet` consente di bilanciare efficacemente le esigenze di ricerca con i requisiti di riservatezza, garantendo al contempo la preservazione dell'intelligibilità strutturale dell'archivio.

La logica di anonimizzazione è stata implementata attraverso un algoritmo che opera sul *triplestore* secondo la procedura sintetizzabile nello pseudocodice presentato nel Listato 7.6:

```
INIZIALIZZAZIONE sistema_anonimizzazione con:
  - namespaces RiC-O, LRMoo e BoDi
  - endpoint di interrogazione Blazegraph
  - liste di autorizzazione (white list) ed esclusione (black list) in file.xslx

# Fase 1: Recupero entità dal triplestore
IDENTIFICA tutte_le_entità rico:Record E rico:RecordSet

# Fase 2: Classificazione Record secondo logica di autorizzazione
PER OGNI record IN rico:Record:
  SE record.isRelatedTo(lrmoo:F1_Work):
    MARCA record COME visibile_opera

  SE record IN lista_autorizzazione_esplicita:
    MARCA record COME visibile_whitelist

  SE record.isIncludedIn(recordset_autorizzato):
    MARCA record COME visibile_gerarchia_whitelist

  SE    NON    visibile_opera    AND    NON    visibile_whitelist    AND    NON
   visibile_gerarchia_whiteslist:
    AGGIUNGI record A entità_da_anonimizzare

# Fase 3: Classificazione RecordSet secondo logica di esclusione
PER OGNI recordset IN rico:RecordSet:
  SE recordset IN lista_esclusione_esplicita:
    AGGIUNGI recordset A entità_da_anonimizzare
  SE recordset.contiene(record_visibile):
    RIMUOVI recordset DA entità_da_anonimizzare

# Fase 4: Anonimizzazione completa entità selezionate
PER OGNI entità IN entità_da_anonimizzare:
  AGGIORNA_SPARQL {
    ELIMINA { entità rdfs:label ?etichetta_precedente }
    ELIMINA { entità rico:title ?titolo_precedente }
    INSERISCI { entità rdfs:label "Redacted information" }
    INSERISCI { entità rico:title "Redacted information" }
    INSERISCI { entità bodi:redactedInformation "yes" }
  }
```

```
# Fase 5: Marcatura entità visibili
PER OGNI entità IN entità_visibili:
  AGGIORNA_SPARQL {
    INSERISCI { entità bodi:redactedInformation "no" }
  }

# Fase 6: Controllo selettivo metadati per entità visibili
PER OGNI entità WHERE { entità bodi:redactedInformation "no" }:
  PER OGNI metadato_tecnico IN entità.rappresentazioni.metadatiTecnici:
    SE metadato_tecnico.tipo IN ["Creator", "Author", "LastModifiedBy"]:
      SE NON autore_accettabile(metadato_tecnico.valore):
        ANONIMIZZA metadato_tecnico CON "Redacted information"
```

Listato 7.6. Pseudocodice rappresentante l'algoritmo di anonimizzazione selettiva per le entità archivistiche.

Parte IV - Visualizzazione

8. Scalable reading d'archivio: ipotesi e risultati

Il *distant reading*, metodologia teorizzata da Franco Moretti per l'analisi quantitativa della letteratura attraverso strumenti computazionali (2000, 2013), trova nell'archivistica contemporanea un campo di applicazione fecondo per la comprensione di fenomeni altrimenti invisibili all'analisi tradizionale. Per la prima volta nella storia degli studi archivistici, la granularità e l'elefantiasi dei dati consentono di applicare metodologie di analisi quantitativa su scala massiva, rivelando pattern strutturali precedentemente inaccessibili agli strumenti dell'archivistica tradizionale. Se Moretti aveva concepito questa prospettiva per superare i limiti del *close reading* nell'approccio ai grandi corpora testuali, consentendo di studiare i testi attraverso la visualizzazione di pattern quantitativi e strutturali (2000), l'adozione di metodologie analoghe negli studi archivistici permette di affrontare l'iperproduzione documentale digitale, in cui diventa sempre più complesso individuare elementi rilevanti all'interno del crescente caos informativo.

L'opposizione binaria tra *close* e *distant reading* è tuttavia fuorviante, come ha argomentato Martin Mueller nel suo concetto di *scalable reading*[252]. Come ricostruisce Benjamin Krautter (2024), Mueller propone lo *scalable reading* come una sintesi felice tra approcci qualitativi e quantitativi: lo *scaling* non rappresenta una scelta tra *big* o *small data*, ma una sfida metodologica su come analizzare oggetti di studio da prospettive multiple, combinando risultati di operazioni condotte a scale diverse.

Applicata all'archivistica, questa prospettiva si innesta su una tradizione disciplinare già avvezza al cambio di scala nell'analisi documentaria. L'informatizzazione degli archivi prende avvio già negli anni Sessanta con i primi esperimenti di automazione degli strumenti di ricerca, e si consolida nel corso degli anni Ottanta con l'introduzione di database relazionali e sistemi informativi che permisero di elaborare statistiche aggregate su consistenza dei fondi, tipologie documentarie e distribuzioni cronologiche (Bunn 2016). La disponibilità di archivi nativi digitali introduce però una novità fondamentale: mai prima d'ora si è disposto di una tale mole di dati granulari derivanti dalla natura intrinseca dei materiali stessi. Ogni file digitale incorpora automaticamente metadati tecnici che documentano la propria genesi e le proprie trasformazioni, tracce involontarie dell'attività computazionale che, estratte sistematicamente e

[252] Il blog post originale di Martin Mueller, *Scalable Reading* (2012), pubblicato sul blog Scalable Reading (https://scalablereading.northwestern.edu), non è attualmente più accessibile online. Le considerazioni qui riportate sono tratte dalla ricostruzione di Krautter (2024).

modellate attraverso ontologie e *knowledge graph*, diventano oggetti di indagine aggregabili e analizzabili su larga scala, rendendo possibile la ricostruzione di pratiche d'uso, flussi di lavoro e stratificazioni temporali.

Lo *scalable reading d'archivio* si configura così come approccio metodologico che integra analisi automatizzate dei metadati con indagini qualitative mirate, permettendo di muoversi fluidamente tra vista d'insieme e analisi puntuale per delineare configurazioni e pattern all'interno di ecosistemi documentari altrimenti difficilmente gestibili per dimensioni e complessità. Questa metodologia consente di individuare ipotesi di ricerca che orientano successivamente L'indagine qualitativa verso aree di particolare interesse archivistico, storico o letterario.

In questo senso, possiamo tentare di ribaltare la prospettiva e considerare come gli archivi nativi digitali, grazie alla loro ricca componente informativa intrinseca e alla possibilità di modellarne i dati, offrano l'opportunità di trasformare una massa quantitativa potenzialmente problematica in una risorsa conoscitiva. È possibile attivare analisi quantitative che si configurano come parte integrante del valore testimoniale del digitale d'autore, aprendo a nuove possibilità di comprensione del fondo, delle pratiche creative e delle forme di autorialità (Gorini e Giagnolini 2025).

Per esemplificare in concreto le potenzialità di questo approccio, il presente capitolo presenta l'applicazione della metodologia all'Archivio Evangelisti attraverso una serie di analisi, illustrate nei capitoli seguenti, che comprendono: statistiche generali (cap. 8.1) e strutturali (cap. 8.2); la distribuzione dei tipi di *media* (cap. 8.3); individuazione dei codici hash come chiavi di esplorazione e strumenti di tracciamento dei file (cap. 8.4); l'analisi dell'attività computazionale di Evangelisti nel tempo, ovvero la creazione e la modifica dei contenuti digitali (cap. 8.5), con un focus sulle risorse legate alla produzione dei romanzi (cap. 8.6).

8.1 Statistiche generali

L'analisi strutturale dell'Archivio Evangelisti si basa su un corpus totale di 61.154.322 triple RDF che documentano in totale 78.211 `rico:Record` (rappresentazione del contenuto informativo dei file), 11.135 `rico:RecordSet` (rappresentazione del contenuto informativo delle cartelle) e 89.346 `rico:Instantiation` (rappresentazione delle istanze fisiche di file e cartelle), ciascuna associata a una specifica `prov:Location` che ne identifica il percorso nella struttura gerarchica (ossia il *file path*).

Il dataset comprende tre raggruppamenti principali: il disco rigido principale (RS1_RS1), il disco esterno (RS1_RS2) e i riversamenti dai supporti floppy disk (RS1_RS3), ciascuno caratterizzato da

configurazioni organizzative differenziate che riflettono diverse funzioni d'uso e diversi vincoli tecnologici (Tabella 8.1).

Supporto	rico:Record	rico:RecordSet
Hard Disk	56.171	9.624
Hard Disk esterno	19.810	1.394
Floppy Disk	2.230	117

Tabella 8.1. Distribuzione dei record e dei record set per tipologia di supporto.

L'estrazione dei metadati tecnici ha identificato 5.574.418 entità `bodi:TechnicalMetadata` distribuite sulle tre *directory* dell'archivio attraverso quattro strumenti di estrazione specializzati (Apache Tika, ExifTool e libreria os). L'hard disk principale concentra il 74,6% dei metadati totali (4.157.298 entità), mentre l'hard disk esterno rappresenta il 23,6% del corpus (1.316.452 metadati), i floppy disk, prevedibilmente costituiscono solo l'1,8% del totale (100.668 metadati) (Tabella 8.2).

Supporto	bodi:TechnicalMetadata (Os Library)	bodi:TechnicalMetadata (Apache Tika)	bodi:TechnicalMetadata (ExifTool)
Hard Disk	855.335	1.737.220	1.564.743
Hard Disk esterno	275.652	461.813	578.987
Floppy Disk	30.511	22.368	47.789

Tabella 8.2. Distribuzione delle entità bodi:TechnicalMetadata per strumento di estrazione e tipologia di supporto.

L'analisi della granularità tipologica dei metadati evidenzia come i diversi strumenti impiegati presentino gradi di specializzazione eterogenei (Tabella 8.3). ExifTool si distingue per l'ampiezza della varietà tipologica prodotta, che raggiunge 106.211 `bodi:TechnicalMetadataType` sull'hard disk, suggerendo una marcata incidenza di contenuti multimediali. Apache Tika, pur senza raggiungere la stessa estensione, mantiene un livello di diversificazione intermedio (tra 244 e 960 tipologie per *directory*), in coerenza con la molteplicità dei formati documentali trattati. Al contrario, la libreria os restituisce un insieme essenziale di tipologie dettate dalla natura stessa dei metadati estratti (metadati di sistema, 13 per *directory*). Questa caratterizzazione quantitativa documenta non solo la ricchezza del patrimonio digitale evangelistiano, ma anche la complessità tecnologica crescente dell'ecosistema di scrittura.

Supporto	bodi:TechnicalMetadataType (Os Library)	bodi:TechnicalMetadataType (Apache Tika)	bodi:TechnicalMetadataType (ExifTool)
Hard Disk	13	960	106.211
Hard Disk esterno	13	867	7.647
Floppy Disk	13	244	269

Tabella 8.3. Distribuzione del numero di tipologie di metadati (bodi:TechnicalMetadataType) per strumento di estrazione e tipologia di supporto.

8.2 Analisi strutturale

Il primo affondo analitico è condotto attraverso l'esame della distribuzione dei file per livello di profondità, dove il livello 0 corrisponde alla *root* di ogni supporto. La Tabella 8.4 riporta i livelli delle *directory* con il maggior numero di file, ordinati in senso decrescente e per supporto di provenienza, indicando per ciascuno di essi il grafo di appartenenza e il conteggio complessivo dei file.

Supporto	Grafo	Profondità	Conteggio file
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	12	11.640
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	9	7.814
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	10	7.545
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	11	6.068
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	3	5.882
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	6	3.376
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	15	3.328
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	14	2.955
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	7	2.419
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	13	2.106
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	8	1.206
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	17	682
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	5	347
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	4	291
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	16	262
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	2	219
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	18	18
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	1	13

Supporto	Grafo	Profondità	Conteggio file
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	10	7.528
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	4	3.067
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	11	3.062
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	3	2.931
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	9	1.086
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	5	398
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	12	398
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	7	392
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	15	349
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	17	143
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	14	137
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	13	99
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	6	83
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	8	65
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	2	43
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	18	18
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	16	9
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	1	2
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	2	1.931
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	3	298
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	4	1

Tabella 8.4. Distribuzione dei file per livello di profondità, supporto e grafo di appartenenza.

La distribuzione dei file nel disco rigido principale mostra un picco marcato alla profondità 12 (11.640 file). Volumi elevati si osservano anche alle profondità 9, 10 e 11, delineando un'area di intensa attività organizzativa compresa tra i livelli 9 e 12.

Le profondità più basse (livelli 1-4) registrano un numero molto ridotto di file, suggerendo che i primi livelli della gerarchia siano occupati prevalentemente da cartelle di contenimento. È presente una coda verso le profondità estreme (15-18), contenente poche centinaia di file, probabilmente riconducibili a sottostrutture residuali o cartelle di esportazione/archivi compressi non uniformati alla struttura principale. Rispetto alla configurazione dell'hard disk portatile, il computer di lavoro presenta una maggiore profondità (raggiungendo il picco di 18 livelli di profondità) e una ramificazione più estesa.

Nell'hard disk esterno, il massimo si colloca alla profondità 10 (7.528 file), ma la distribuzione complessiva risulta più omogenea nell'intervallo 3-12. Particolarmente rilevante è il dato del livello 4 (3.067 file), indicativo di un'organizzazione meno gerarchicamente annidata e potenzialmente più orientata a un accesso diretto ai contenuti.

La coda alle profondità più elevate è quasi assente (livelli 16-18 con valori trascurabili), confermando che la struttura è mediamente meno stratificata. Tali caratteristiche confermano, anche dal punto di vista macroscopico, come l'hard disk esterno contenga copie o backup eterogenei, integrati da più fonti e privi di una coerenza strutturale uniforme, ma comunque con una distribuzione più compatta rispetto alla copia dell'hard disk del computer fisso.

Chiaramente, la distribuzione dei file nei supporti floppy è marcatamente piatta: il massimo si registra alla profondità 2 (1.931 file), seguito da valori minori ai livelli 3 e 4. Sono completamente assenti profondità superiori a 4, un dato coerente con le limitazioni intrinseche del supporto e con la decisione dell'Associazione di utilizzare il livello 1 come semplice contenitore in rappresentanza del disco.

L'analisi della distribuzione delle cartelle per livello di profondità evidenzia strutture differenziate tra i dispositivi considerati (Tabella 8.5).

Supporto	Grafo	Profondità	Conteggio cartelle
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	9	1695
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	8	1619
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	7	1268
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	14	1049
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	10	965
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	11	725
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	12	712
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	13	557
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	5	352
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	6	279
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	2	130
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	15	85
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	16	61
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	4	51
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	3	46
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	17	18

Supporto	Grafo	Profondità	Conteggio cartelle
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	1	11
Hard Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS1	0	1
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	9	176
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	2	171
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	12	171
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	6	167
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	13	141
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	3	117
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	10	107
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	14	74
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	4	70
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	11	57
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	5	41
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	16	26
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	8	20
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	15	19
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	1	16
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	7	12
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	17	8
Hard Disk Esterno	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS2	0	1
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	1	93
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	2	22
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	0	1
Floppy Disk	http://ficlit.unibo.it/ArchivioEvangelisti/structure/RS1_RS3	3	1

Tabella 8.5. Distribuzione delle cartelle per livello di profondità, supporto e grafo di appartenenza.

Sul disco principale, il numero di cartelle cresce rapidamente fino ai livelli intermedi (7-12), con un picco di file per cartella al livello 9-12, dove si concentra la maggior parte della produzione. I livelli più bassi (1-4) contengono poche cartelle e pochissimi file, suggerendo che le *directory* principali fungono da contenitori di ordine generale. Ai livelli più profondi (15-17) si osservano poche cartelle con una minima quantità di file, probabilmente residui o sottostrutture archiviate. Il livello di profondità 14 presenta un numero relativamente elevato di cartelle (1.049), suggerendo la possibile presenza di una struttura interna altamente annidata all'interno dell'albero delle cartelle. Analizzando quali tipologie di cartelle sono

effettivamente associate a questo livello con conteggio anomalo, emerge che il livello 14 concentra principalmente backup di smartphone Samsung (2013-2014) con contenuti WhatsApp, cache di browser multipli (Chrome, Edge, Maxthon) con estensioni localizzate, e dati di utilizzo di software Microsoft. Si tratta principalmente di residui di interazioni digitali, che dimostrano come questo tipo di archivi, al di là dei contenuti creativi, possa rappresentare anche un deposito fenomenologico dell'interazione umano-digitale.

Se si intrecciano i dati di distribuzione per livello con le tipologie di file in essi contenuti, l'analisi tipologica conferma l'interpretazione: il livello 14 contiene 1.192 documenti di testo, 280 immagini vettoriali SVG e 147 documenti HTML, distribuiti in 1.049 cartelle con una densità di circa 2,8 file per cartella. Questa configurazione, caratterizzata da alta frammentazione e tipologie miste, è caratteristica delle cartelle di cache profonde del sistema operativo, probabilmente legate a processi automatici di backup, log di sistema o cache di applicazioni web. Il livello 3 presenta una configurazione anomala con 3.867 immagini JPEG, concentrazione massiva che contrasta con la distribuzione più diffusa dei livelli successivi, suggerendo una zona di accumulo multimediale distinta dalla logica organizzativa generale. La zona più densa dell'albero (livelli 7-12), dove si concentra la maggior parte del lavoro computazionale, rivela attraverso l'analisi tipologica la sua natura sistemica. Il livello 9 presenta 1.695 cartelle con una distribuzione tipologica caratterizzata da 2.611 immagini vettoriali SVG e 1.997 immagini PNG, indicative di risorse grafiche di sistema, temi Windows e icone di applicazioni, piuttosto che di contenuti personali. Ai livelli più profondi (15-17) si osservano poche cartelle con una minima quantità di file, ma l'analisi tipologica rivela una specializzazione funzionale residuale: il livello 15 concentra 1.441 documenti di testo in sole 85 cartelle (densità di 17 file per cartella), suggerendo aggregazioni automatiche di log o file di configurazione, mentre il livello 17 presenta 319 file binari in 18 cartelle, probabilmente componenti di sistema orfani.

Il livello 12 rappresenta il culmine della stratificazione sistemica con 712 cartelle contenenti 9.194 file binari generici, una densità di quasi 13 file binari per cartella che documenta la presenza del cuore del sistema operativo: librerie DLL, componenti Windows, file di registro e cache di sistema profondamente annidati.

Sull'hard disk esterno, la distribuzione delle cartelle è più uniforme tra i livelli 2-12, con densità di file per cartella più bassa rispetto al disco principale, coerente con una struttura meno rigidamente organizzata e potenzialmente frutto di copie o backup parziali. Sul disco portatile, la distribuzione delle cartelle è più uniforme tra i livelli 2-12, con densità di file per cartella più bassa rispetto al disco principale, coerente con una struttura meno rigidamente organizzata e frutto di copie o backup parziali.

L'analisi tipologica conferma questa interpretazione rivelando una strategia di selezione consapevole: il livello 3 presenta 1.382 immagini JPEG e 250 documenti Word Legacy in 117 cartelle, indicando una densità controllata (circa 14 file per cartella) che contrasta con l'accumulo massivo del disco principale. Il livello 10, pur raggiungendo il picco quantitativo con 4.369 immagini JPEG, mantiene una distribuzione tipologica orientata ai contenuti personali con una ragguardevole presenza di video MPEG (118) e AVI (48), documentando una funzione archivistica multimediale piuttosto che sistemica. La quasi totale assenza di file binari di sistema e la prevalenza di formati documentari di testo (Word Legacy, JPEG, PDF) confermano ulteriormente che l'hard disk esterno rappresenta un backup selettivo e curatoriale che esclude il "rumore" tecnologico del sistema operativo.

8.3 Considerazioni sui codici hash

L'analisi dell'Archivio Evangelisti dal punto di vista delle impronte crittografiche ha individuato 78.211 entità di tipo `premis:Fixity` presenti nella *knowledge base*, corrispondenti ad altrettanti codici hash SHA-256 associati a ciascun file del corpus. Di questi, 51.714 rappresentano hash unici, mentre 9.080 sono condivisi da almeno due file. Questo approccio metodologico ha permesso di individuare pattern di duplicazione che vanno oltre il valore conservativo o la semplice ridondanza tecnica, acquisendo valore documentario per la ricostruzione dell'ambiente digitale dell'autore. In questo senso, l'analisi delle impronte crittografiche rivela piste di indagine per comprendere le strategie di conservazione e organizzazione adottate da Evangelisti, così come l'evoluzione tecnologica del suo ecosistema operativo.

L'esame delle occorrenze più frequenti ha evidenziato al primo posto 1.816 file che condividono lo stesso codice hash SHA-256, corrispondente al valore universale dei file vuoti (zero byte). Questo specifico hash (e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855) rappresenta una costante informatica riconosciuta da tutti i sistemi e applicazioni, costituendo un identificatore affidabile per l'individuazione di file privi di contenuto (Scientific Computing 2024).

L'esame delle denominazioni dei file privi di contenuto rivela grande eterogeneità, ma orientata verso componenti dell'ecosistema Windows e applicazioni web. La tipologia predominante comprende file di sistema Windows, rappresentati da elementi di configurazione del desktop come *desktop.ini*, che preservano metadati di personalizzazione dell'interfaccia utente con marcatura temporale specifica (Microsoft 2024). Secondo la documentazione ufficiale Microsoft, questi file contengono configurazioni per la personalizzazione delle cartelle attraverso sezioni `[.ShellClassInfo]` e sono automaticamente generati durante l'inizializzazione di cartelle di sistema (Microsoft 2024).

Parallelamente, emergono identificatori di sistema caratterizzati da Globally Unique Identifiers (GUID) utilizzati per garantire univocità nell'identificazione di componenti software e hardware all'interno dell'architettura Windows (Microsoft Developer Documentation 2024b). Troviamo anche i file di lock hardware e driver, esemplificati da denominazioni come *C4C-172F-41B9-91B8-7F0C3B1E9021_VEN_1002&DEV_699F&SUBSYS_841E&REV_C7.lock*, che indicano meccanismi di controllo per dispositivi hardware specifici, implementati attraverso il Windows Driver Framework per prevenire accessi concorrenti alle risorse hardware (Microsoft Developer Documentation 2024a).

Evidente anche la presenza di strutture di database e cache dalla presenza di file denominati *CURRENT* e *MANIFEST-000001*, caratteristici delle architetture database LevelDB[253] e RocksDB[254] utilizzate rispettivamente da Google e da Facebook per la gestione di dati strutturati ad alta performance (Dean et al. 2024; Facebook 2024). Questi si accompagnano a file di indicizzazione generici (*index*) e configurazioni in formato JSON (*messages.json*).

Un'ulteriore dimensione emersa dall'analisi riguarda la ricorrenza componenti di interfaccia applicativa e web, rappresentati da file HTML come *previewsItem.html* e file di testo strutturato come *mainItem.txt*, oppure file di immagine per elementi interattivi come *btnNews_on.jpg*, *btnShowroom_on.jpg*.

Di particolare interesse archivistico è l'identificazione di file in formato RTF e DOC privi di contenuto ma inequivocabilmente collegati alla produzione di Valerio Evangelisti. Il fatto che siano file vuoti potrebbe suggerire un loro ruolo da *placeholder* oppure di residui di processi di migrazione, copia-e-incolla o backup incompleti.

Il file fin qui presi in esame, che condividono lo stesso codice hash (e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855), sono tutti privi di contenuto e si potrebbe ipotizzarne lo scarto. A questi file, si uniscono anche 5276 cartelle vuote, di cui 4905 nella copia dell'hard disk principale, 363 nell'hard disk esterno e 8 nel riversamento dei floppy. In termini di rappresentazione, in tal caso, eliminando la copia dal supporto, sarebbe opportuno eliminare anche la relativa istanza di `rico:Instantiation`, ma lasciare l'istanza di `rico:Record` o `rico:RecordSet` come traccia della precedente esistenza del file o della cartella, arricchita di metadati che ne comunichino il ruolo e associata anche ad una `rico:Activity` specifica a testimonianza dello scarto. Per il momento, tuttavia, in assenza di indicazioni specifiche sulla destinazione finale dei materiali nativi digitali dell'Archivio Evangelisti in termini di conservazione, si è preferito limitarsi a fotografare

[253] https://github.com/google/leveldb.
[254] https://github.com/facebook/rocksdb.

l'esistente. L'ecosistema di rappresentazione, in ogni caso, è abbastanza flessibile per documentare qualsiasi attività sui materiali a partire da questo quadro iniziale.

Nella classifica dei file con più occorrenze, al secondo posto troviamo 549 file che condividono l'hash SHA-256 "1d76e2b20fd0d1d8336b3146da5bf9bc2dfd6a9634b4c60952604f312d483a0e", tutti identificati come file *desktop.ini*. La dimensione del file è di 298 byte per ogni istanza, mentre varia esclusivamente la denominazione: alla label "desktop" si associa una marcatura temporale specifica, ad esempio *desktop (2015_11_24 13_46_43 UTC).ini*.

L'analisi dei *file path* rivela una localizzazione uniforme all'interno della *directory* dell'hard disk del computer, */FileHistory/Valerio/VALERIO/Data/C/Users/Valerio/SkyDrive/*, che identifica l'origine di questi artefatti: il sistema Windows FileHistory, implementato per la prima volta in Windows 8 come soluzione di backup automatico per i file utente (Windows Team 2012). Questa collocazione, combinata con la presenza di SkyDrive (ora OneDrive) (Microsoft Italia 2014), suggerisce un sistema ibrido di sincronizzazione cloud e backup locale che operava sull'ambiente di lavoro di Valerio Evangelisti.

L'analisi temporale rivela una distribuzione di questi file che abbraccia un anno, dal 22 novembre 2014 al 24 novembre 2015, suggerendo un'attivazione periodica di un meccanismo di backup automatico. La dimensione costante di 298 byte per tutti i file conferma che il contenuto del file "desktop.ini" è rimasto inalterato durante questo periodo, verosimilmente riconducibile a una configurazione desktop stabile nel tempo.

L'integrazione tra FileHistory e SkyDrive evidenzia una strategia consapevole di preservazione digitale che suggerisce abitudini di lavoro metodiche e un ambiente digitale organizzato per una continuità operativa. Benché tecnicamente superflui per la loro ridondanza, dal punto di vista archivistico questi 549 file costituiscono una forma di documentazione automatica che traccia involontariamente la persistenza dell'ambiente di lavoro di Evangelisti, acquisendo valore testimoniale delle pratiche personali di conservazione digitale e della disciplina nell'utilizzo degli strumenti informatici da parte dell'autore, offrendo una prospettiva inedita sulle sue modalità operative. Inoltre, benché possegga un codice hash diverso, il rinvenimento del file *desktop (2015_11_28 02_38_10 UTC).ini* con *file path* */FileHistory/Valerio/VALERIO/Data/C/Users/Valerio/OneDrive/desktop (2015_11_28 02_38_10 UTC).ini* suggerisce una transizione tecnologica piuttosto che l'abbandono del sistema di backup. La presenza di "OneDrive" invece di "SkyDrive" in questo percorso, datato 28 novembre 2015, è infatti immediatamente successiva al *rebranding* e all'aggiornamento del servizio Microsoft (Microsoft Italia 2014) e potrebbe quindi indicare un adattamento del sistema di backup al nuovo servizio cloud.

Questa evoluzione trova conferma nell'emergere di un secondo ramo localizzato nella cartella */altro disco/Nicol/nicol/AppData/Local/Microsoft/OneDrive/*, che documenta la continuità operativa del sistema OneDrive sotto quello che sembrerebbe essere un diverso profilo utente. L'analisi di questo secondo ramo rivela una certa persistenza temporale che si estende fino al 2022, pur presentando alcune anomalie cronologiche in termini di continuità.

Al terzo posto nella classificazione dei file con maggiori duplicazioni emerge un pattern che amplia l'analisi dei sistemi di backup cloud operanti nell'ambiente digitale esaminato. I 182 file che condividono l'hash SHA-256 "4bf5122f344554c53bde2ebb8cd2b7e3d1600ad631c385a5d7cce23c7785459a" rivelano l'esistenza del sistema Dropbox, essendo tutti posizionati all'interno del percorso `/altro disco/Nicol/nicol/AppData/Local/Packages/C27EB4BA.DropboxOEM_xbfy0k16fey96/AC/BackgroundTransferApi/`. La dimensione uniforme di 1 byte per tutti i 182 file suggerisce si tratti di file di metadati o di stato, caratteristici dei sistemi di sincronizzazione cloud per tracciare operazioni di trasferimento, anche se la natura specifica di questi file richiederebbe analisi aggiuntive per essere determinata con certezza. Dropbox, d'altro canto, era sicuramente in uso sull'altro ramo della *directory* dell'hard disk del pc, data la presenza della cartella "Dropbox" con *file path* */FileHistory/Valerio/VALERIO/Data/C/Users/Valerio/Dropbox*.

Al quarto e quinto posto della classifica dei file più ricorrenti emergono rispettivamente 161 file *CURRENT* (16 byte) e 98 file *MANIFEST-000001* (41 byte), entrambi componenti strutturali del sistema database LevelDB. Come per i precedenti pattern analizzati, questi file costituiscono tracce dell'infrastruttura tecnologica piuttosto che documenti della produzione intellettuale di Evangelisti. La loro natura puramente sistemica li colloca in una categoria di "documentazione involontaria" dell'ambiente digitale, preziosa per ricostruire il contesto operativo ma priva di connessione diretta con l'attività creativa dell'autore.

Al di là della classifica per frequenza, l'indagine sulle occorrenze dello stesso codice hash evidenzia come nel fondo esistano file contenutisticamente identici ma con denominazioni diverse. Non si registrano variazioni nelle dimensioni, nei formati o nelle date di creazione e modifica del contenuto, mentre può variare sistematicamente la data associata alla `rico:Instantiation`, ossia la data registrata dal *file system*. Questa variazione testimonia le operazioni di spostamento, copia o riorganizzazione avvenute in fasi successive alla creazione originaria. Questo semplice dato di fatto evidenzia ancora una volta la complessità ontologica dell'oggetto digitale nell'archivistica contemporanea. L'informazione si articola necessariamente su multiple dimensioni simultanee, richiedendo strumenti concettuali adeguati

alle esigenze di rappresentazione della loro natura stratificata, alle quali cercano di rispondere modelli come BoDi.

8.4 Tipologie di file

La classificazione MIME (Multipurpose Internet Mail Extensions) standardizzata dalla Internet Assigned Numbers Authority (IANA)[255], fornisce un sistema di classificazione formale dei di file che permette di categorizzare il corpus documentario secondo criteri uniformi e internazionalmente condivisi (vedi capitolo 7.2). L'analisi della composizione tipologica dell'Archivio Evangelisti attraverso i *media type* estratti dai metadati tecnici consente di delineare il profilo tecnologico e operativo dell'ambiente digitale dell'autore, individuando 60 tipologie MIME Type distinte distribuite su un totale di 57.015 file classificati (Tabella 8.6).

MIME Type	Hard Disk	Hard Disk Esterno	Floppy Disk	Totale
image/jpeg	10.581	7.153	44	17.778
text/rtf	4.629	4.656	34	9.319
text/plain	3.966	1.576	662	6.204
image/png	4.587	346	0	4.933
image/svg+xml	3.461	1	0	3.462
application/msword	986	994	853	2.833
application/xml	231	1.420	0	2.731
text/html	456	1.427	10	1.893
application/json	1.434	11	0	1.445
application/octet-stream	817	179	14	1.010
application/pdf	558	215	0	773
application/x-gzip	750	1	0	751
image/vnd.fpx	219	454	6	679
image/gif	275	207	25	507
image/bmp	292	166	0	458
application/unknown	191	5	153	349

[255] https://www.iana.org/assignments/media-types/media-types.xhtml.

MIME Type	Hard Disk	Hard Disk Esterno	Floppy Disk	Totale
application/zip	263	50	18	331
audio/mpeg	161	108	0	269
application/rdf+xml	125	125	0	250
video/mpeg	0	127	6	133
font/woff	128	0	0	128
image/webp	108	0	0	108
application/x-font-ttf	64	7	0	71
application/vnd.openxmlformats-officedocument.wordprocessingml.document	38	22	0	60
video/x-msvideo	0	58	0	58
font/woff2	57	0	0	57
image/x-icon	40	8	1	49
application/epub+zip	30	16	0	46
image/vnd.djvu	43	0	0	43
video/mp4	15	7	5	27
video/x-ms-wmv	0	27	0	27
application/x-shockwave-flash	25	1	0	26
application/x-7z-compressed	25	1	0	26
image/tiff	6	5	11	22
audio/ogg	11	11	0	22
application/vnd.ms-officetheme	21	0	0	21
image/x-cursor	17	0	2	19
application/vnd.oasis.opendocument.text	8	6	0	14
image/x-jps	11	0	0	11
application/vnd.ms-excel	4	4	2	10
application/x-rar-compressed	2	6	0	8
audio/x-wav	4	0	1	5

MIME Type	Hard Disk	Hard Disk Esterno	Floppy Disk	Totale
application/vnd.adobe.air-application-installer-package+zip	3	2	0	5
application/x-bittorrent	3	2	0	5
application/vnd.ms-powerpoint	2	2	0	4
application/vnd.openxmlformats-officedocument.wordprocessingml.template	4	0	0	4
audio/x-pn-realaudio	2	2	0	4
audio/x-matroska	3	0	0	3
application/x-mobipocket-ebook	3	0	0	3
video/webm	3	0	0	3
image/pcx	0	0	2	2
application/vnd.ms-word.template.macroEnabledTemplate	2	0	0	2
application/x-iso9660-image	2	0	0	2
application/vnd.iccprofile	2	0	0	2
video/quicktime	2	0	0	2
audio/mp4	1	1	0	2
video/x-m4v	2	0	0	2
application/postscript	1	1	0	2
application/ResEdit	0	0	1	1
application/bzip2	1	0	0	1

Tabella 8.6. Distribuzione media type per supporto.

La distribuzione evidenzia una netta predominanza di contenuti visuali e testuali, che costituiscono congiuntamente oltre il 70% del corpus totale, confermando la natura dell'archivio come *repository* di un'attività intellettuale che integra sistematicamente la dimensione testuale con quella iconografica.

Per facilitare l'interpretazione dell'articolazione tipologica, i 60 *media type* identificati sono stati aggregati in 10 macrocategorie sulla base della classificazione dei media (Figura 8.1).

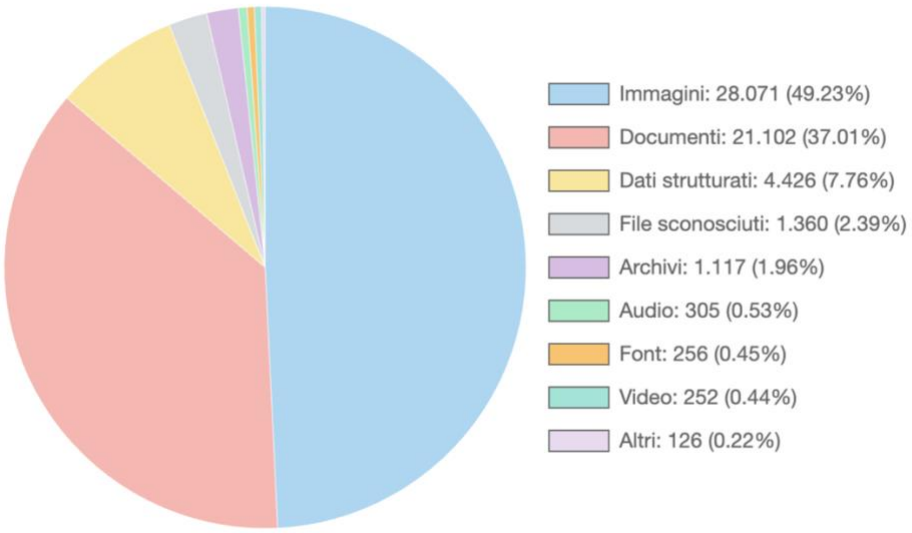


Figura 8.1. Distribuzione delle macro-tipologie documentarie individuate nell'Archivio Evangelisti.

Le immagini (49,23%) emergono come la componente quantitativamente dominante, con 28.071 file distribuiti su 13 *media type* distinti. All'interno di questa macrocategoria, il formato JPEG rappresenta da solo il 63,3% delle immagini totali (17.778 file), seguito da PNG (17,6%, 4.933 file) e SVG (12,3%, 3.462 file). La prevalenza del formato JPEG e la presenza massiccia di PNG suggeriscono una componente documentaria importante legata alla raccolta di materiali visivi, accumulati per finalità di ricerca ma anche come raccolta di fotografie personali. La presenza massiva di SVG, formato vettoriale caratteristico delle risorse grafiche dei sistemi operativi moderni, indica una forte componente sistemica costituita da icone ed elementi di interfaccia. I media type minori includono `image/vnd.fpx` (FlashPix, 679 file), un formato Kodak per immagini multi-risoluzione ormai obsoleto; `image/gif` (507 file), standard per animazioni e grafica web; `image/bmp` (458 file), formato bitmap non compresso tipico dell'ambiente Windows; e formati più specializzati come WebP (108 file), DjVu (43 file, specifico per documenti scansionati), TIFF (22 file), e formati residuali come x-cursor (19 file) e x-icon (49 file) per elementi di interfaccia.

Nella categoria documenti sono stati inclusi tutti i *media type* tendenzialmente riconducibili a risorse testuali. Compongono il 37,01% dei materiali con 21.102. L'analisi della distribuzione interna rivela una predominanza del formato RTF (Rich Text Format) con 9.319 file (44,2% dei documenti), seguito da `text/plain` con 6.204 file (29,4%) e `application/msword` con 2.833 file (13,4%). La prevalenza di RTF rispetto ai formati Microsoft Word (con soli 60 file nel formato moderno DOCX) potrebbe riflettere l'uso prolungato di versioni più datate di software di elaborazione testi che utilizzavano RTF come

formato predefinito, convenzioni e accordi con gli editor o semplicemente un'abitudine personale. La presenza di 1.893 file HTML (8,9% dei documenti) suggerisce una componente importante legata al salvataggio di contenuti web. I file PDF costituiscono solo il 3,7% dei documenti (773 file), una percentuale relativamente contenuta rispetto al totale della produzione. Infine, i formati aperti (OpenDocument con 14 file) e i template Word (4 file standard e 2 con macro) rappresentano componenti residuali dell'ecosistema documentario.

Il raggruppamento dei dati strutturati (7.76%, 4.426 file) aggrega formati caratterizzati da strutture sintattiche formali destinate primariamente all'interscambio tra sistemi piuttosto che alla lettura umana diretta. La predominanza di `application/xml` (2.731 file, 61,7% della categoria) e `application/json` (1.445 file, 32,6%) indica l'utilizzo intensivo di formati di configurazione software, export di database e risorse per applicazioni web.

La categoria Archivi (1,96%, 1.117 file) documenta pratiche sistematiche di compressione e archiviazione dei dati. La netta prevalenza del formato `application/x-gzip` (751 file, 67,2%) indica l'adozione di standard tipici degli ambienti Unix/Linux, verosimilmente in relazione a procedure di backup automatizzate, distribuzioni software o esportazioni. Il formato ZIP (331 file, 29,6%) rappresenta lo standard di riferimento per la compressione multipiattaforma, compatibile con la maggior parte dei sistemi operativi. La presenza di formati meno diffusi, quali 7Z (26 file, ad alta efficienza di compressione), RAR (8 file, formato proprietario WinRAR) e BZip2 (1 file, associato a sistemi Unix), suggerisce l'impiego differenziato di strumenti di compressione in funzione di specifiche esigenze tecniche o operative.

Le categorie Audio (0.53%, 305 file) e Video (0.44%, 252 file) rappresentano componenti minoritarie ma documentano un arco temporale esteso di evoluzione tecnologica. Per l'audio, il formato dominante è MPEG/MP3 (269 file, 88.2%), seguito da formati meno diffusi come Ogg Vorbis (22 file), WAV (5 file), RealAudio (4 file, legacy), Matroska Audio (3 file) e MP4 Audio (2 file). La distribuzione dei file video mostra la coesistenza di formati appartenenti a diverse generazioni tecnologiche: standard storici come MPEG (133 file, 52,8%), AVI (58 file, 23,0%) e WMV (27 file, 10,7%) si affiancano a formati più recenti come MP4 (27 file, 10,7%), QuickTime, M4V e WebM.

Una serie di file residuali (0,22%, 126 file) aggrega formati con percentuali individuali inferiori allo 0,1%: fogli di calcolo (`application/vnd.ms-excel`: 10 file), presentazioni (`application/vnd.ms-powerpoint`: 4 file; `application/vnd.ms-officetheme`: 21 file), e-book (`application/epub+zip`: 46 file; `application/x-mobipocket-ebook`: 3 file), file di sistema (`application/vnd.iccprofile`: 2 file

per profili colore ICC; `application/x-iso9660-image`: 2 file per immagini disco ISO), applicazioni (`application/x-shockwave-flash`: 26 file Adobe Flash; `application/vnd.adobe.air-application-installer-package+zip`: 5 file Adobe AIR), protocolli di rete (`application/x-bittorrent`: 5 file torrent), documenti di stampa (`application/postscript`: 2 file PostScript), e template con macro (`application/vnd.ms-word.template.macroEnabledTemplate`: 2 file).

Infine, la categoria file sconosciuti (2,38%, 1.360 file) aggrega tre tipologie di media indicative di incertezza classificatoria: `application/octet-stream` (1.010 file, 74,3% della categoria), classificazione generica assegnata a file binari quando il sistema non riesce a determinare il formato specifico; `application/unknown` (349 file, 25,7%), etichetta esplicitamente indicativa di fallimento del processo di identificazione; `application/ResEdit` (1 file), riferimento a un editor di risorse Macintosh legacy. La predominanza di octet-stream suggerisce la presenza di file corrotti, formati proprietari obsoleti senza signature riconoscibili, o componenti di sistema binari non standardizzati. L'alta percentuale di file sconosciuti nei floppy disk (153 su 2.230 file totali, 6,9% contro il 2,4% complessivo) conferma l'ipotesi di problematiche legate all'obsolescenza del supporto e dei relativi formati.

L'analisi della distribuzione tipologica differenziata per supporto di memorizzazione rivela strategie organizzative e pattern di utilizzo divergenti tra l'hard disk principale, l'hard disk esterno e i floppy disk. Sul disco rigido principale, la distribuzione tipologica riflette la complessità di un sistema operativo attivo con stratificazioni sistemiche massicce. Le immagini JPEG, pur rappresentando la categoria singola più numerosa (10.581 file, 18,8% del disco), sono controbilanciate da una forte presenza di file sistemici: `image/svg+xml` (3.461 file, 6,2%), `image/png` (4.587 file, 8,2%) e `application/json` (1.434 file, 2,6%) documentano l'infrastruttura tecnologica del sistema operativo Windows e delle applicazioni installate. La numerosa presenza di `application/x-gzip` (750 file) suggerisce attività di archiviazione e backup automatizzati, mentre i documenti di testo (text/rtf: 4.629, `text/plain`: 3.966, `application/msword`: 986) costituiscono, almeno in parte, il nucleo della produzione intellettuale, rappresentando complessivamente il 20,4% del disco principale.

L'hard disk esterno (19.810 file totali) presenta una distribuzione tipologica differente che, anche da questo punto di vista, conferma una marcata selezione curatoriale predominanza di immagini JPEG (7.153 file, 36,1% del disco esterno) e documenti RTF (4.656 file, 23,5%) suggerisce una funzione archivistica orientata alla preservazione selettiva dei materiali di lavoro. La quasi totale assenza di file sistemici (solo 1 file SVG, 346 PNG contro i 4.587 del disco principale, 11 JSON contro i 1.434) conferma la natura curatoriale del backup, che esclude sistematicamente il "rumore" tecnologico dell'infrastruttura operativa.

I floppy disk (2.230 file totali) presentano una distribuzione estremamente coerente con l'uso storico del supporto per la memorizzazione di documenti di lavoro: la predominanza assoluta di documenti di testo (`text/plain`: 662 file, 29,7%; `application/msword`: 853 file, 38,2%; `text/rtf`: 34 file, 1,5%, per un totale del 69,4% del supporto) documenta l'utilizzo prioritario del supporto per la conservazione della produzione letteraria e saggistica nelle fasi precedenti all'adozione degli hard disk. La presenza di 153 file `application/unknown` (6,9% contro il 2,4% medio dell'intero archivio) riflette, probabilmente, la presenza di formati proprietari legacy non più riconoscibili dai sistemi contemporanei.

L'analisi basata sui *media type*, pur fornendo un quadro macroscopico efficace della composizione tipologica dell'archivio, presenta alcuni limiti metodologici che suggeriscono sviluppi futuri della ricerca. In primo luogo, la classificazione MIME opera a un livello di dettaglio limitato: il tipo `application/msword`, ad esempio, non distingue tra diverse versioni del formato Word (come DOC 97-2003, DOC 6.0/95, DOC 2.0), ciascuna caratterizzata da specifiche proprietà tecniche e gradi differenti di obsolescenza. In secondo luogo, l'identificazione MIME si basa primariamente su euristiche (estensione file, *magic numbers* nei primi byte) che possono risultare ambigue o fallire completamente per file corrotti, legacy o con metadati compromessi, come evidenziato dall'alta percentuale di file classificati come `application/octet-stream` e `application/unknown`. Infine, la classificazione dei media non cattura informazioni importanti per la preservazione e la descrizione, come le dipendenze software specifiche e le versioni di formato.

Per superare questi limiti, come prospettiva futura si propone di integrare l'analisi dei *media type* con l'identificazione dei formati tramite DROID, strumento sviluppato da The National Archives UK che utilizza il registro PRONOM per distinguere con precisione formato e versione. La sua applicazione permetterebbe di discriminare, ad esempio, tra diverse versioni di formati Microsoft Office, RTF e JPEG e di identificare i file attualmente classificati come "sconosciuti", rivelando formati legacy, componenti software obsoleti o file danneggiati.

Come ulteriore sviluppo, si potrebbero integrare le informazioni estratte da DROID con una lettura incrociata dei registri della Library of Congress sui formati adatti alla preservazione (Library of Congress 2025), al fine di supportare valutazioni più accurate e ipotesi di migrazione dei formati a rischio. L'integrazione di queste fonti permetterebbe, da un lato, di migliorare la granularità dell'informazione e, dall'altro, di guidare decisioni operative di preservazione digitale, differenziando materiali stabili da materiali a rischio e orientando strategie di migrazione verso formati più sostenibili.

8.5 Distribuzione temporale delle attività

L'analisi delle date di creazione e modifica associate ai `rico:Record` provenienti dall'hard disk principale di Evangelisti consente di delineare l'andamento temporale della sua attività attraverso due prospettive complementari: il corpus totale e la documentazione specificamente attribuibile all'autore[256].

Nell'intervallo temporale compreso tra il 1990 e il 2022, l'archivio restituisce un totale di 6.888 file dotati di data di creazione[257]. Fra questi, focalizzando l'analisi sui documenti il cui creatore o ultimo autore di modifica risulta essere Evangelisti, emergono 5.120 file, corrispondenti al 74,35% del totale.

Entrambi i dataset mostrano una convergenza significativa nell'identificazione dei periodi di massima produttività. La distribuzione temporale del corpus totale evidenzia una concentrazione netta a partire dagli anni Duemila, che costituiscono il decennio più produttivo (3.582 file), seguiti dagli anni Dieci (2.511 file), mentre gli anni Novanta restano marginali (198 file) (Figura 8.2).

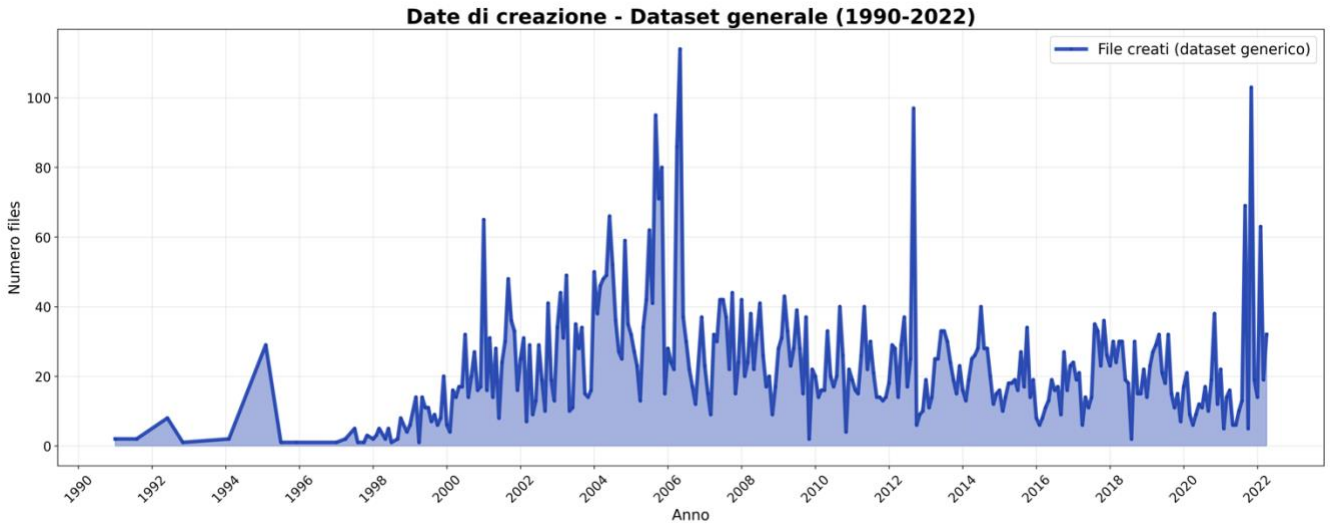


Figura 8.2. Distribuzione temporale dei file basata sulle date di creazione.

[256] Per definire il dataset autoriale sono stati selezionati i record con metadati `dc:creator` o `meta-LastAuthor` uguali a "Evangelisti", "Valerio" o "Valerio Evangelisti". Come osservato nell'indagine sui fondi PAD (Gorini e Giagnolini 2025), questa selezione non consente di identificare con certezza l'intera produzione dell'autore: è probabile che alcuni materiali scritti o elaborati da Evangelisti non siano stati tracciati con questi metadati, così come alcuni file contrassegnati con questi valori potrebbero non essere stati fisicamente prodotti da lui. Pur con queste limitazioni, la selezione permette di isolare un sottoinsieme di file attribuibili ad alta probabilità all'autore, utile per analisi temporali e strutturali della sua attività digitale.

[257] Per garantire la coerenza dei risultati, sono state escluse le date anomale, ossia quelle anteriori agli anni Ottanta e quelle successive al 2022: le prime, verosimilmente imputabili a calcoli errati del *timestamp*, richiederebbero una più approfondita indagine filologica; le seconde possono invece essere ricondotte sia ad anomalie informatiche sia a interventi postumi di terzi, come nel caso dei documenti datati al 2023.

Il dataset autoriale conferma sostanzialmente questa periodizzazione, pur con alcune specificità. I documenti di Evangelisti come autore mostrano una progressiva intensificazione a partire dal 2000, con una distribuzione che rispecchia quella del corpus generale ma con una maggiore concentrazione negli anni Duemila e una presenza più limitata negli anni Dieci (Figura 8.3).

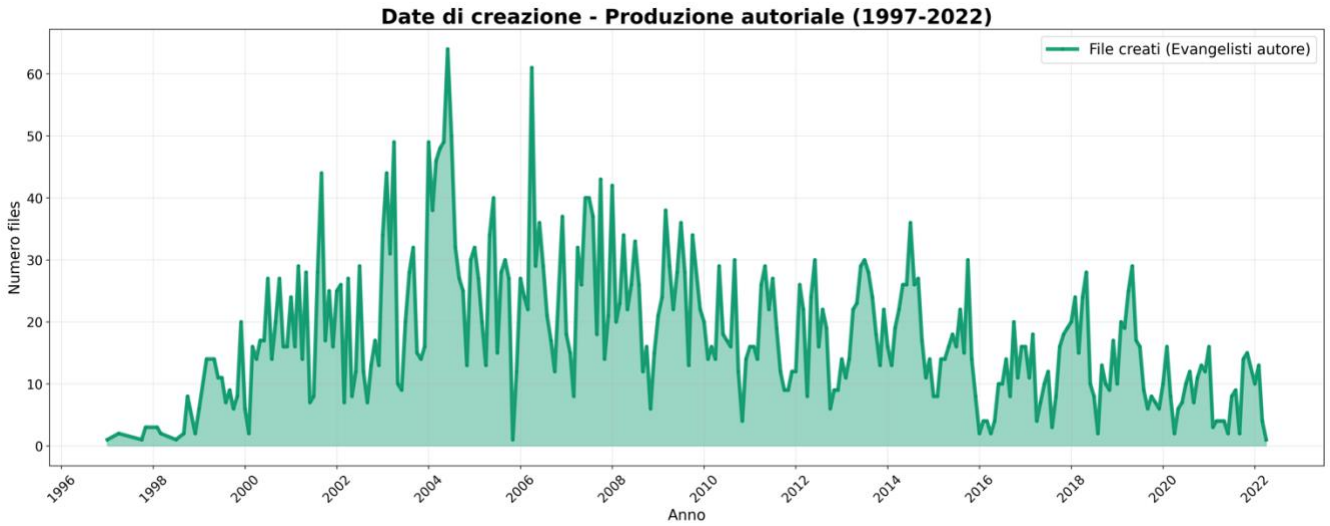


Figura 8.3. Distribuzione temporale dei file creati da Valerio Evangelisti basata sulle date di creazione.

Nel dataset generale, il periodo 2020-2022 conta 597 file, un dato che, se rapportato alla breve durata di soli due anni, si configura come una media piuttosto elevata. Tale circostanza è probabilmente legata al fatto che l'hard disk analizzato apparteneva a un computer in uso negli ultimi anni di vita dell'autore: è quindi comprensibile che vi si concentri una porzione corposa di documenti recenti, anche non necessariamente correlati al processo creativo. Al tempo stesso, risulta importante osservare la presenza di materiali risalenti ai primi anni Duemila e, in misura minore, persino agli anni Novanta. Tale dato testimonia un'attenzione consapevole al recupero e alla conservazione della propria produzione, resa evidente dai trasferimenti di documenti attraverso diversi dispositivi informatici.

L'analisi dei file datati agli anni Novanta mostra, in particolare, il riversamento sul computer di lavoro di articoli, lettere e appunti, ma non delle stesure dei primi romanzi, che risultano invece conservate su floppy disk. In alcuni casi, la copia situata sull'hard disk presenta nel nome del file una data recente fra parentesi (ad esempio *Article pour Le Monde Diplomatique (2020_02_07 16_14_24 UTC).rtf*), spesso coincidente con la data di ultima modifica della cartella che lo contiene. Questo elemento induce a ipotizzare che tali marcature temporali siano state generate automaticamente dal sistema di lettura dei floppy disk utilizzato da Evangelisti nelle fasi di riversamento oppure, più probabilmente, da un sistema di backup.

L’analisi dei picchi di produzione consente di isolare momenti di intensificazione attraverso una doppia prospettiva che rivela tanto l’ecosistema digitale complessivo quanto i momenti di creatività autoriale (Figura 8.4).

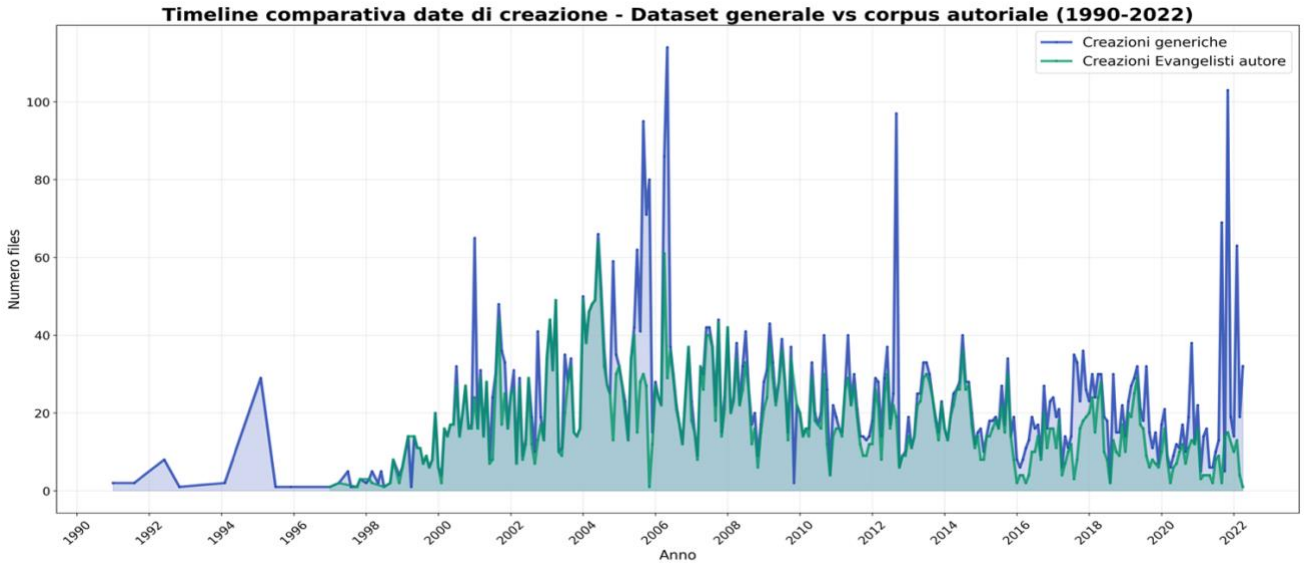


Figura 8.4. Distribuzione temporale dei file dell’Archivio Valerio Evangelisti: confronto tra il corpus totale e il dataset attribuibile all’autore sulla base delle date di creazione.

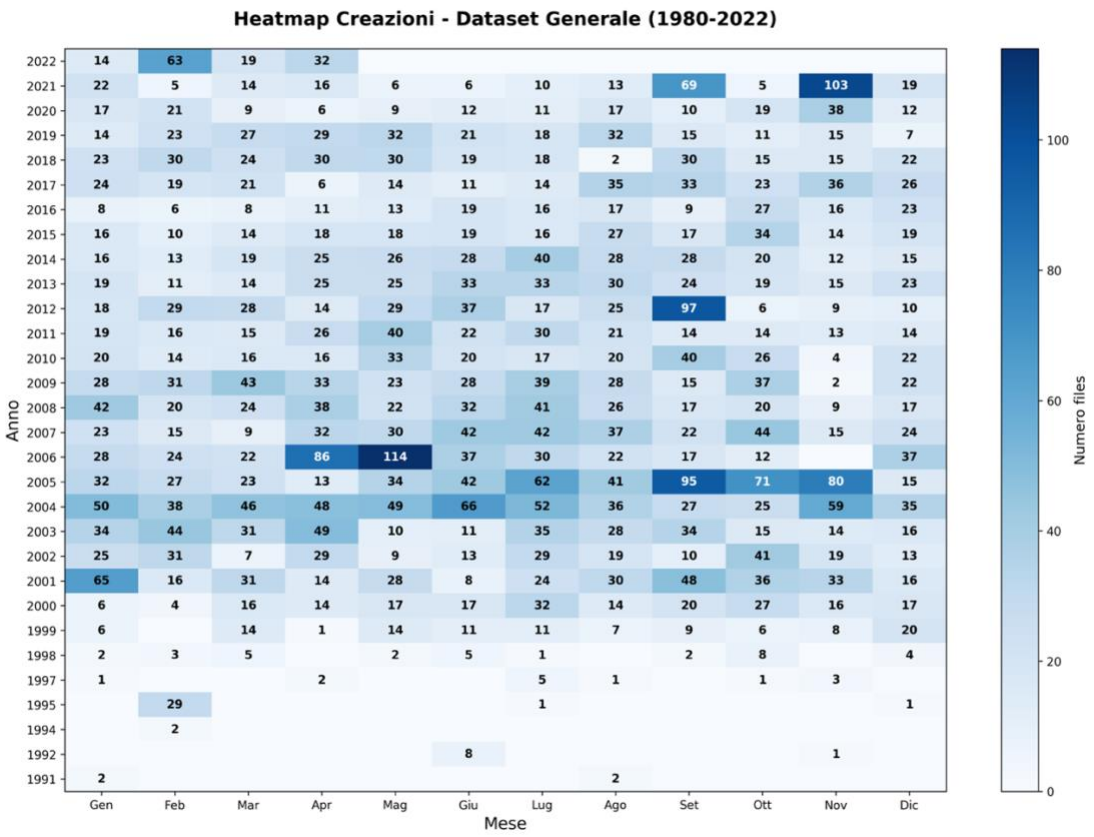


Figura 8.5. Heatmap della distribuzione temporale dei file dell'Archivio Valerio Evangelisti basata sulle date di creazione, con dettaglio mensile.

Nel dataset generale, il massimo assoluto si registra nel maggio 2006 (114 file creati), seguito da novembre 2021 (103 file) e da settembre 2012 (97 file). Altri periodi di particolare densità si collocano fra il 2005 e il 2006, confermando una fase di attività particolarmente intensa con valori che oscillano stabilmente sopra le 80 unità mensili (Figura 8.5).

Il maggio 2006 trova conferma anche nella documentazione autoriale, seppur con intensità ridotta (29 file), testimoniando che questo momento di massima produttività coincide effettivamente con l'attività diretta dell'autore. Tuttavia, emergono picchi autoriali specifici che arricchiscono il quadro: settembre 2005 (30 file), aprile 2006 (61 file) e, più tardi, febbraio 2022 (13 file). La concentrazione nel biennio 2005-2006 risulta specificamente marcata nel dataset autoriale, confermando che si tratta di un periodo di straordinaria fertilità creativa piuttosto che di semplice attività di archiviazione (Figura 8.6).

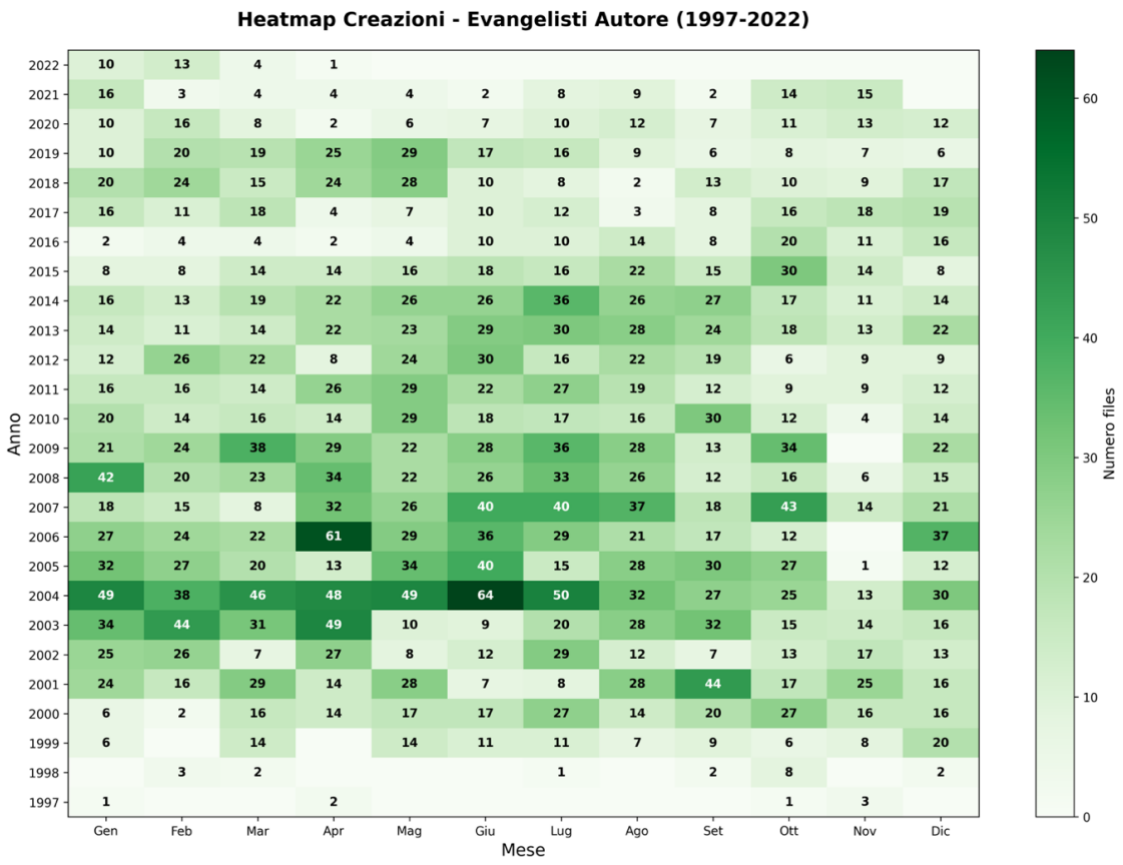


Figura 8.6. Heatmap della distribuzione temporale dei file attribuibili a Valerio Evangelisti, basata su creazione e ultima modifica, con dettaglio mensile.

Altri periodi di particolare densità autoriale si collocano nel 2004 (471 file totali nell'anno) e nel 2003 (321 file). Al contrario, i picchi di novembre 2021 e settembre 2012, rilevanti nel dataset generale, mostrano una presenza autoriale limitata (rispettivamente 15 e 19 file), suggerendo che questi momenti di intensa attività computazionale fossero legati prevalentemente a operazioni di gestione tecnica, backup o semplice utilizzo quotidiano, soprattutto per il periodo più prossimo alla morte.

Si configura dunque una duplice polarizzazione cronologica che assume significati diversi nei due dataset: da un lato, la metà degli anni Duemila emerge come periodo di convergenza tra attività digitale generale e produzione autoriale; dall'altro, il ritorno di produttività negli anni immediatamente precedenti alla morte si manifesta prevalentemente come attività di gestione quotidiana nel dataset generale, mentre mantiene una presenza autoriale minore nel dataset specifico.

Parallelamente, l'analisi delle date di modifica restituisce un quadro sostanzialmente coerente con le date di creazione, pur rivelando interessanti divergenze tra corpus generale e documentazione autoriale. Il

corpus totale conta 1.806 file modificati[258], distribuiti anch'essi in prevalenza negli anni Duemila (1.116 file), con picchi concentrati fra il 2001 e il 2006 (Figura 8.7). Anche in questo caso, il maggio 2006 emerge come il mese di maggiore intensità, con 85 file modificati, seguito da novembre 2005 (78 file) e settembre 2005 (65 file) (Figura 8.8).

Tuttavia, l'analisi comparativa con il dataset autoriale rivela dinamiche più complesse. Mentre il corpus generale presenta 1.806 file modificati, la documentazione autoriale ne conta solo 283, concentrati principalmente nel periodo 2000-2007. Questa discrepanza, con il dataset autoriale che rappresenta solo il 15,7% delle modifiche totali, suggerisce che gran parte dell'attività di revisione documentata nell'archivio non sia direttamente riconducibile all'intervento dell'autore (Figure 8.9 e 8.10).

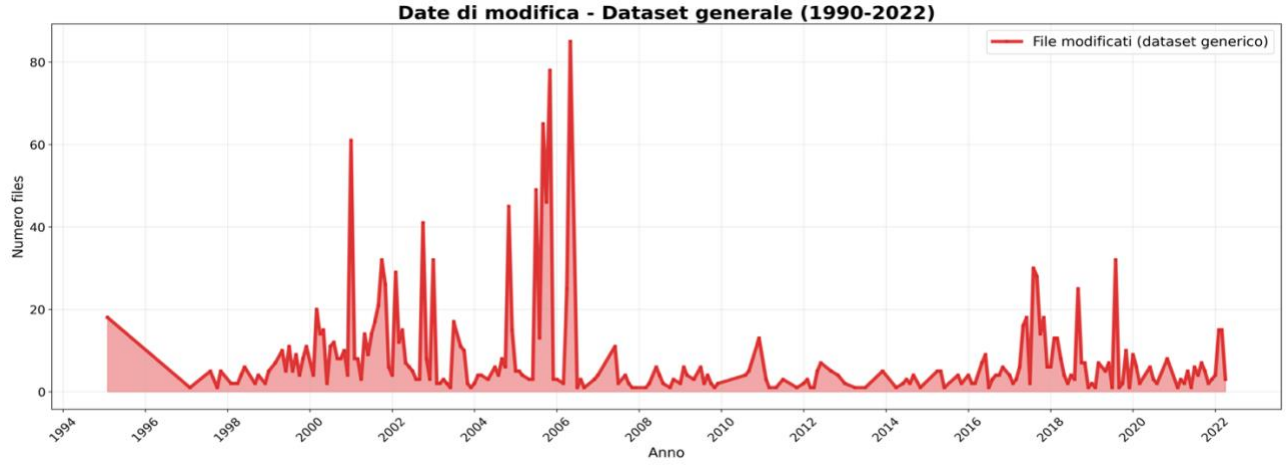


Figura 8.7. Distribuzione temporale dei file basata sulle data di modifica.

[258] La discrepanza fra il numero di date di creazione e il numero di date di modifica è dettata dal fatto che non tutti i formati di file supportano modifiche e dunque la registrazione di tale metadato.

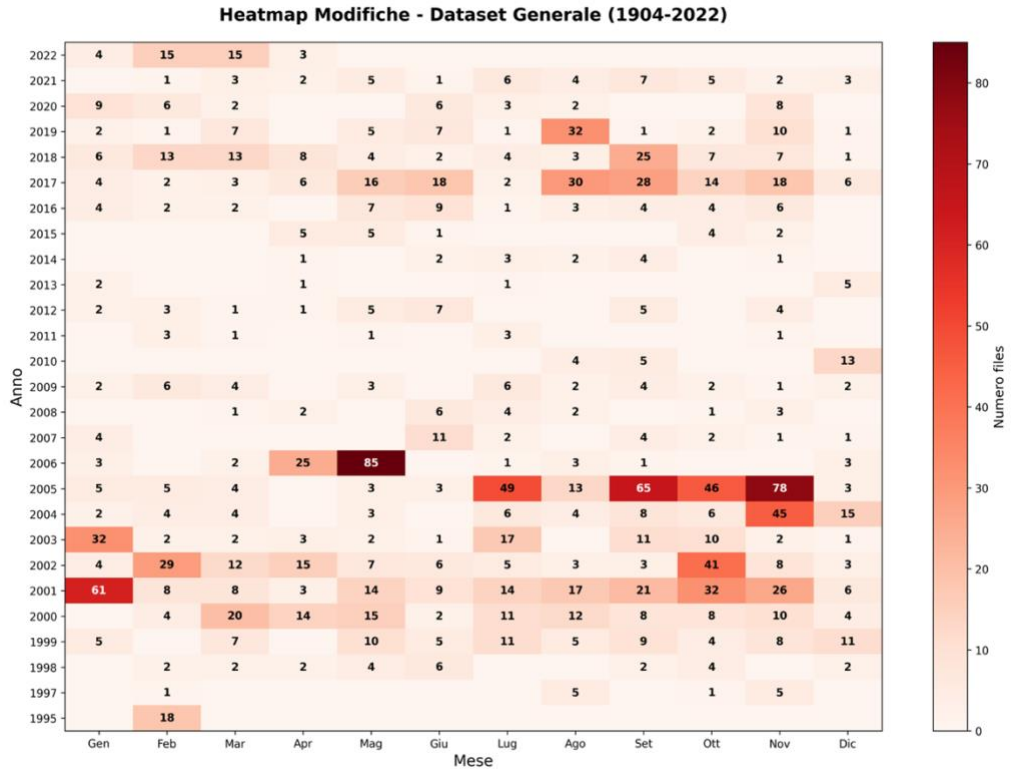


Figura 8.8. Heatmap della distribuzione temporale dei file dell'Archivio Valerio Evangelisti basata sulle date di creazione, con dettaglio mensile.

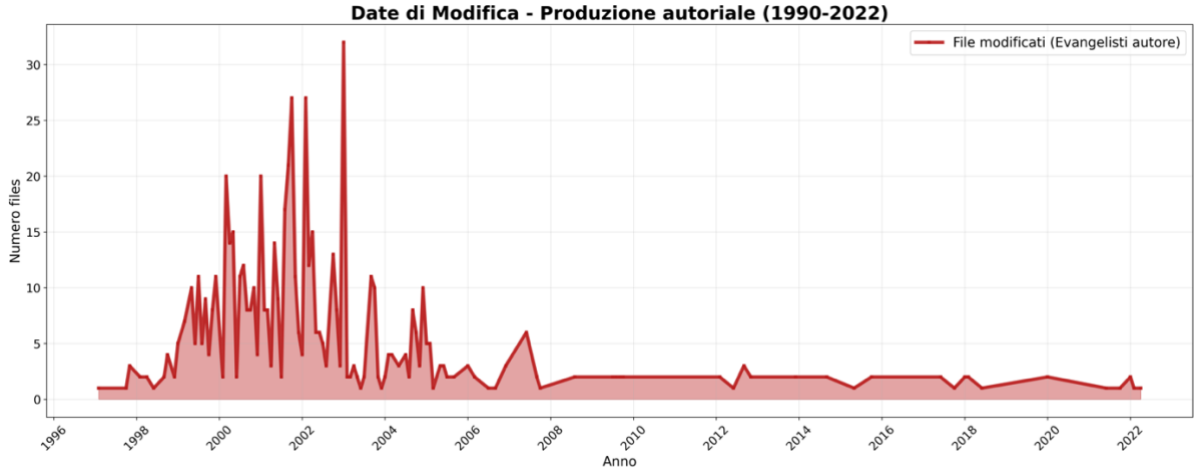


Figura 8.9. Distribuzione temporale dei file creati da Valerio Evangelisti basata sulle date di modifica.

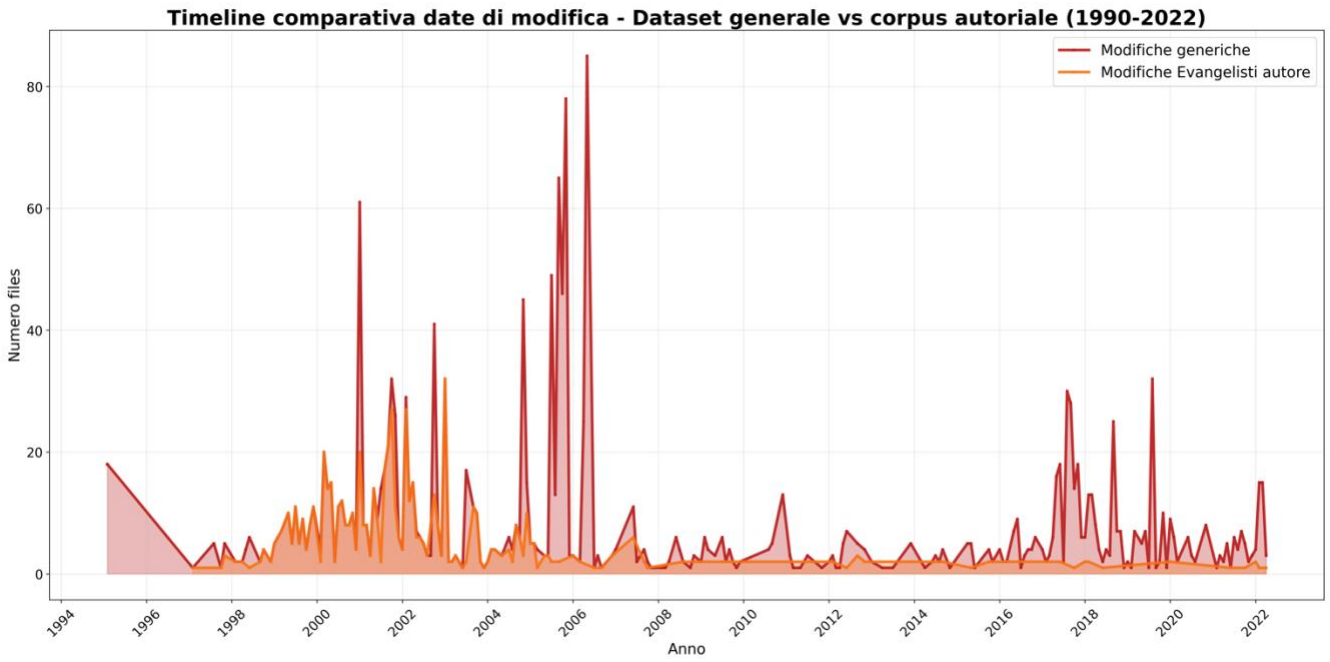


Figura 8.10. Distribuzione temporale dei file dell'Archivio Valerio Evangelisti: confronto tra il corpus totale e il dataset attribuibile all'autore sulla base delle date di modifica.

Il maggio 2006, pur emergendo come picco massimo nel corpus generale (85 file modificati), mostra una presenza autoriale minima nel dataset specifico. Al contrario, i picchi di modifica autoriale si concentrano in periodi diversi e precedenti: marzo 2000 (20 file), febbraio 2002 (27 file) e gennaio 2001 (20 file), indicando che i processi di revisione dell'autore seguivano tempistiche differenti rispetto all'attività generale dell'ambiente (Figura 8.11).

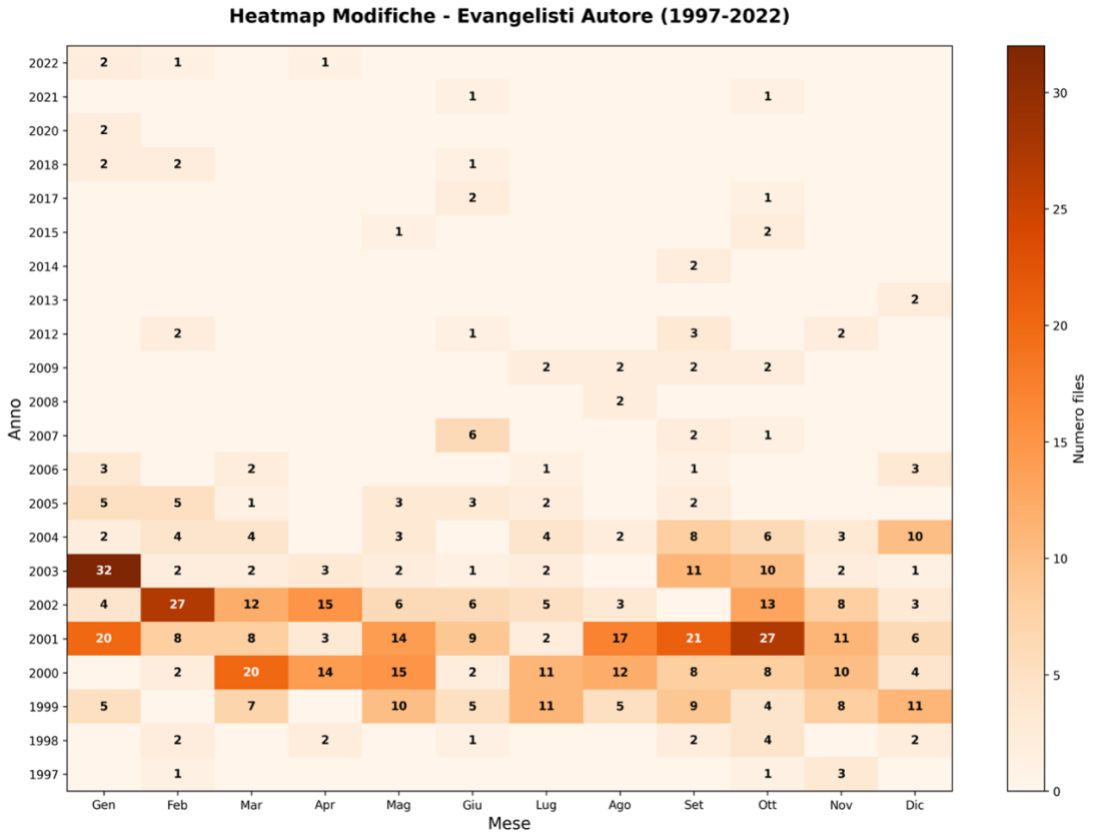


Figura 8.11. Heatmap della distribuzione temporale dei file dell'Archivio Valerio Evangelisti basata sulle date di modifica, con dettaglio mensile.

Il delta fra il numero di modifiche e di creazioni si amplia nel periodo 2008-2015, un fenomeno che nel dataset autoriale risulta ancora più marcato, con una presenza di modifiche estremamente ridotta in questi anni. Questo pattern suggerisce che, dopo il periodo di intensa creatività del biennio 2005-2006, l'attività di revisione autoriale diretta si sia notevolmente ridotta.

Tali evidenze quantitative, se messe a confronto con la produzione letteraria di Evangelisti, mostrano come il periodo di massima intensità computazionale coincida con una fase di grande fertilità editoriale. Nel biennio 2005-2006, infatti, l'autore pubblica i primi due volumi del *Ciclo messicano* (*Il collare di fuoco*, 2005; *Il collare spezzato*, 2006), oltre al racconto *La sala dei giganti* (2006). La sovrapposizione tra picco digitale e editoriale suggerisce che l'accresciuta attività computazionale non si limiti a riflettere un momento di intensa scrittura, ma includa anche operazioni complementari di consultazione di materiali, revisione, editing e preparazione editoriale.

La sovrapposizione tra i dati del corpus totale e del dataset autoriale conferma questa correlazione, rivelando però la natura stratificata dell'attività computazionale. L'analisi rivela che dei 964 documenti creati nel corpus generale durante il periodo 2005-2006, una porzione importante (circa 279 file) è

riconducibile direttamente all'attività autoriale di Evangelisti. Di questi, 234 corrispondono effettivamente a stesure de *Il collare di fuoco*, *Il collare spezzato* e *La sala dei giganti*, accompagnati da una ventina di interviste e recensioni collegate.

Allo stesso nucleo autoriale si affiancano testi di opere che saranno pubblicate negli anni immediatamente successivi, quali il racconto *Controinsurrezione* (in *Controinsurrezioni*, Mondadori, 2008, con A. Moresco), "*Siamo gli autonomi, siamo i più duri…*" (in S. Bianchi e L. Caminiti, a c. di, *Gli autonomi: le storie, le lotte, le teorie*, DeriveApprodi, 2007), *I fratelli della costa* (in A. D. Altieri, a c. di, *Anime nere*, Mondadori, 2007) e *Wilhelm Reich contro l'America* (in R. Polese, a c. di, *Il complotto: teoria, pratica, invenzione*, Guanda, 2007).

La presenza di picchi meno marcati nel dataset autoriale rispetto al corpus generale suggerisce che quest'ultimo include anche materiali complementari di consultazione, editing e preparazione editoriale non direttamente attribuibili ad Evangelisti. Accanto ai documenti testuali, infatti, il corpus in questo lasso temporale comprende anche consistenti album fotografici relativi a viaggi a Barcellona, Gijón e Puerto Escondido che non solo testimoniano pratiche personali di archiviazione, ma rappresentano viaggi che sono stati fonti di ispirazione per la scrittura del Ciclo dei Pirati. Il picco massimo di produttività del maggio 2006 coincide infatti con la redazione degli ultimi capitoli de *Il collare spezzato*, ma anche con l'importazione sul computer delle fotografie del viaggio a Puerto Escondido, località in cui Evangelisti possedeva un'abitazione. Infine, i dati relativi alle modifiche (85 nel maggio 2006 e 78 nel novembre 2005) confermano l'esistenza di un processo di elaborazione intensiva e pressoché simultanea, concentrato soprattutto nelle cartelle dedicate ai romanzi.

In generale, si nota una correlazione notevole fra i picchi di elementi creati o modificati e la parabola editoriale, a testimonianza anche della rapidità del processo che dall'atto creativo portava alla pubblicazione. Gli anni Duemila (3.582 file nel corpus totale) coincidono con la pubblicazione di opere fondamentali come *Il Castello di Eymerich* (2001), *Black Flag* (2002), *Mater Terribilis* (2002), l'avvio del *Ciclo Americano* con *Antracite* (2003) e *Noi Saremo Tutto* (2004). Si tratta del periodo di massima diversificazione narrativa dell'autore, che sviluppa contemporaneamente il ciclo di Eymerich, il Ciclo del Metallo Urlante e il Ciclo Americano, un'intensità creativa confermata dalla presenza di 792 file d'autore negli anni 2003-2004, testimonianza diretta della produttività dell'autore in questa fase cruciale della sua carriera.

L'analisi fin qui condotta ha evidenziato come la distinzione tra corpus totale e dataset autoriale possa inquadrare con maggiore precisione i momenti di creatività dell'autore rispetto al portato generale del

dataset. Tuttavia, è necessario esplicitare che i criteri metodologici adottati per questa distinzione hanno alcune limitazioni inevitabili legate alla natura stessa del digitale d'autore. Infatti, per individuare il dataset autoriale, sono stati selezionati quei documenti che esplicitamente riportassero un riferimento al nome e/o cognome di Evangelisti nei metadati `dc:creator`. Tale criterio, pur consentendo di isolare buona parte della produzione direttamente attribuibile all'autore (come dimostrato dalla correlazione tra i picchi computazionali del 2005-2006 e la pubblicazione del Ciclo messicano) presenta limitazioni significative che richiedono una riflessione critica sulle problematiche dell'attribuzione autoriale negli archivi digitali.

La presenza di altri appellativi nel valore del metadato `dc:creator`, infatti, è molto comune e può derivare dall'utilizzo di dispositivi condivisi o istituzionali, da ambienti di scrittura collaborativa, dalla possibilità di modificare manualmente l'autore e altri metadati nelle impostazioni dei software di videoscrittura, dal download e dallo scambio di materiali, dal trasferimento e salvataggio di documenti nel tempo. La logica con cui si sedimentano i documenti nativi digitali in questo non differisce dagli archivi analogici, ma il contesto digitale implica circostanze che, in mancanza delle tracce paleografiche e diplomatistiche proprie dei supporti cartacei tradizionali, complicano l'attribuzione autoriale e pongono sfide inedite per la descrizione archivistica e la ricerca filologica.

I metadati di tipo `dc:creator` che differiscono dal nome dell'autore possono rivelare, ad esempio, relazioni con editor, familiari o colleghi, generando intrecci di autorialità complessi. Una ricerca condotta con Adele Gorini sui fondi di Silvia Avallone e Paolo Di Paolo, conservati dal Centro Manoscritti di Pavia nel contesto del progetto Pavia Archivi Digitali, ha evidenziato dinamiche emblematiche in questo senso (Gorini e Giagnolini 2025). Nel fondo Avallone, per esempio, emergono tracce della collaborazione con Ritanna Armeni per il *Dizionario Femminile*, dove documenti inizialmente inviati da Armeni sono stati riutilizzati come modelli per nuovi testi, implicando la presenza della scrittrice come autrice in contesti svincolati dalla sua documentata collaborazione. Nel fondo Di Paolo si alternano casi in cui testi dell'autore risultano attribuiti ad altri e viceversa, suggerendo utilizzi condivisi di computer o coinvolgimenti nel processo creativo non immediatamente evidenti o noti. Questo quadro introduce, fra l'altro, complicazioni rilevanti nel contesto di opere inedite, dove distinguere tra diversi autori in assenza di testi editi di riferimento diventa ancor più difficoltoso.

Ulteriori elementi di complessità emergono dai campi `LastModifiedBy`[259] o `Company`[260], che possono suggerire relazioni editoriali o istituzionali, e dalla possibilità di effettuare modifiche manuali in Microsoft Word, che consente agli utenti di personalizzare parte dei metadati associati al documento.

Nel caso specifico dell'Archivio Evangelisti, è molto probabile ci siano elementi generati direttamente dall'autore ma non catturati dal filtro applicato: ad esempio, esistono alcune etichette corrispondenti a nomi di uffici potenzialmente riconducibili al suo impiego presso la pubblica amministrazione che potrebbero essere ulteriormente indagate. Una categoria particolare è rappresentata dai materiali di studio in cui il campo `dc:creator` riporta nomi di personaggi storici: filtrando i documenti nativi digitali contemporanei potrebbe stupire ritrovarsi nomi quali Margaret Peeke, Ramón de Luanco o François Tommy Perrens. Ma questi file, materiali di studio provenienti da *digital libraries*, riflettono i metadati di catalogazione, non di generazione del documento. Catalogazione che è dunque corretta dal punto di vista del contenuto, perché associa il testo al suo autore originario, ma non riflette informazioni su chi abbia effettivamente creato quello specifico file. Gli autori del `rico:Record` e della `rico:Instantiation` potrebbero dunque essere diversificati per veicolare questa informazione.

La selezione dei soli documenti recanti il nome di Valerio Evangelisti, pur risultando riduttiva, rappresenta il criterio più immediatamente praticabile per circoscrivere l'insieme documentario di riferimento attraverso la sola analisi computazionale dei metadati. Si tratta di rinunciare, almeno in un primo momento, alla granularità dell'analisi ravvicinata per guadagnare in estensione, con la consapevolezza che il valore euristico di questo livello di analisi dipende dalla sua integrazione con un esame puntuale dei casi significativi che emergono dall'aggregazione. L'attribuzione dell'autorialità si configura, infatti, come operazione particolarmente complessa anche nel contesto del nativo digitale, richiedendo necessariamente l'integrazione di un accurato studio contenutistico e contestuale, in assenza delle componenti diplomatistiche e paleografiche tradizionali. In questa prospettiva, il *distant reading* potrebbe offrire supporto attraverso sperimentazioni di stilometria computazionale (Daelemans 2013; Savoy 2020) da condurre, tuttavia, con estrema cautela in considerazione delle implicazioni relative al diritto d'autore sui materiali.

[259] https://learn.microsoft.com/en-us/dotnet/api/system.io.packaging.packageproperties.lastmodifiedby?view=windowsdesktop-9.0#system-io-packaging-packageproperties-lastmodifiedby.
[260] https://learn.microsoft.com/en-us/dotnet/api/documentformat.openxml.extendedproperties.company?view=openxml-3.0.1.

8.6 Distribuzione dei materiali relativi ai romanzi

Se analizziamo le tre *directory* dell'archivio concentrandoci specificamente sui romanzi, è evidente come Valerio Evangelisti abbia sviluppato una struttura organizzativa coerente, generalmente dedicando una cartella per ciascun romanzo, all'interno della quale i documenti risultano frequentemente articolati secondo la suddivisione in capitoli. L'analisi quantitativa della distribuzione di file e cartelle correlate ai romanzi di Evangelisti consta di 6.073 entità, corrispondenti a file o cartelle correlate ad un'opera o ad un ciclo o trilogia. La distribuzione complessiva dei materiali mostra una presenza rilevante sia sull'hard disk esterno (2.852 elementi, 46,96%) sia su quello principale (2.592 elementi, 42,68%). La cartella dedicata ai floppy disk raccoglie invece 629 elementi, pari al 10,36% del patrimonio riferibile ai romanzi. La vicinanza dei valori nei due hard disk lascia intravedere pratiche di duplicazione, confermando strategie di backup. Questa ipotesi è rafforzata dall'elevato numero di file con codice hash identico, indizio di una ridondanza almeno in parte intenzionale.

L'esame della distribuzione della correlazione alle opere (tabella 8.7) e ai cicli e trilogie (tabella 8.8) rivela l'esistenza di diverse tipologie di archiviazione.

Opera	**Elementi Hard Disk**	**Elementi Hard Disk Esterno**	**Elementi Floppy Disk**	**Totale**
1849. I guerrieri della libertà	143 100%	/	/	143
Antracite	21 50%	21 50%	/	42
Black Flag	60 50%	60 50%	/	120
Cartagena	124 48,63%	131 51,37%	/	255
Cherudek	19 11,31%	20 11,90%	129 76,79%	168
Eymerich risorge	127 68,65%	58 31,35%	/	185
Gli anni del coltello	104 100%	/	/	104
Gocce Nere	1 100%	/	/	1
Il castello di Eymerich	22 29,33%	22 29,33%	31 41,34%	75
Il collare di fuoco	158	158	/	316

	50%	50%		
Il collare spezzato	110 50%	110 50%	/	220
Il corpo e il sangue di Eymerich	16 22,86%	16 22,86%	38 54,28%	70
Il fantasma di Eymerich	168 100%	/	/	168
Il mistero dell'inquisitore	24 28,91%	24 28,91%	35 42,18%	83
Il Sole dell'Avvenire. Chi ha del ferro ha del pane	183 50,97%	176 49,03%	/	359
Il Sole dell'Avvenire. Nella notte ci guidano le stelle	176 14,83%	1011 85,17%	/	1187
Il Sole dell'Avvenire. Vivere lavorando o morire combattendo	156 50,32%	154 49,68%	/	310
La Sala dei Giganti	10 50%	10 50%	/	20
La fredda guerra dei mondi	43 100%	/	/	43
La furia di Eymerich	37 45,68%	43 53,09%	1 1,23%	81
La luce di Orione	3 42,86%	4 57,14%	/	7
Le catene di Eymerich	17 15,89%	16 14,95%	74 69,16%	107
Magus. Il romanzo di Nostradamus. Vol. 1: Il Presagio	22 30,55%	20 27,78%	30 41,67%	72
Magus. Il romanzo di Nostradamus. Vol. 2: L'Inganno	23 23,47%	25 25,51%	50 51,02%	98
Magus. Il romanzo di Nostradamus. Vol. 3: L'Abisso	15 11,03%	15 11,03%	106 74,94%	136
Mater Terribilis	53	53	/	106

	50%	50%		
Metallo Urlante	16 32%	17 34%	17 34%	50
Nicolas Eymerich, inquisitore	49 53,85%	41 45,05%	1 1,10%	91
Noi saremo tutto	99 50,77%	96 49,23%	/	195
One Big Union	107 49,54%	109 50,46%	/	216
Picatrix. La scala per l'inferno	16 11,76%	16 11,76%	104 76,48%	136
Rex tremendae maiestatis	146 48,67%	154 51,33%	/	300
Tortuga	121 65,05%	65 34,95%	/	186
Veracruz	118 49,79%	119 50,21%	/	237

Tabella 8.7. Distribuzione dei file e delle cartelle correlate ai romanzi di Valerio Evangelisti nei diversi supporti di memorizzazione (hard disk principale, hard disk esterno e floppy disk). I valori percentuali indicano la proporzione di ciascun supporto rispetto al totale dei materiali associati a ogni opera.

Ciclo/Trilogia	**Numero opere del Ciclo/Trilogia**	**Elementi Hard Disk**	**Elementi Hard Disk Esterno**	**Elementi Floppy Disk**	**Totale**
Ciclo Il Sole dell'Avvenire	3	542 28,42%	1365 71,58%	/	1907
Ciclo dei Pirati	3	364 53,45%	317 46,55%	/	681
Ciclo di Eymerich	14	740 43,71%	510 30,12%	443 26,17%	1693
Ciclo di Pantera	3	97 45,54%	99 46,48%	17 7,98%	213
Ciclo Messicano	2	268 50%	268 50%	/	536
Trilogia Americana	3	228 50,22%	226 49,78%	/	454
Trilogia di Magus	3	88 24,11%	91 24,93%	186 50,96%	365

Tabella 8.8. Distribuzione dei file e delle cartelle riferibili ai cicli e alle trilogie di Valerio Evangelisti nei diversi supporti di memorizzazione. I valori percentuali indicano la proporzione di ciascun supporto rispetto al totale dei materiali associati a ogni ciclo o trilogia.

La prima tipologia di archiviazione comprende dodici opere caratterizzate da una distribuzione pressoché equilibrata tra l'hard disk principale e quello esterno, senza alcuna presenza di materiali su floppy disk. Si tratta di *Antracite*, *Black Flag*, *Cartagena*, *Il collare di fuoco*, *Il collare spezzato*, *Il Sole dell'Avvenire. Chi ha del ferro ha del pane*, *La Sala dei Giganti* e *Mater Terribilis*. Una seconda tipologia mostra invece una ripartizione analoga tra i due supporti principali, ma con una limitata presenza di file conservati su floppy disk, come avviene per *La furia di Eymerich* e *Nicolas Eymerich, inquisitore*. La terza tipologia, al contrario, si distingue per una considerevole incidenza di materiali archiviati su floppy disk, evidente nei casi di *Picatrix. La scala per l'inferno*, *Magus. Il romanzo di Nostradamus. Vol. 3: L'Abisso*, *Le catene di Eymerich* e *Cherudek*. Infine, alcune opere presentano una distribuzione più articolata e diffusa fra tutti e tre i supporti di memorizzazione, tra cui *Metallo Urlante*, *Il castello di Eymerich*, *Il corpo e il sangue di Eymerich*, *Il mistero dell'inquisitore* e i primi due volumi di *Magus. Il romanzo di Nostradamus* (*Il Presagio* e *L'Inganno*).

Il *Ciclo di Eymerich* (1.693 elementi totali) e la *Trilogia di Magus* (365 elementi) emergono come i nuclei documentari più consistenti e gli unici a conservare una presenza rilevante di materiali su floppy disk (rispettivamente 443 e 186 elementi). Si tratta dei progetti narrativi di più ampia estensione temporale e complessità strutturale, nei quali le diverse fasi di scrittura e revisione si sono sviluppate attraverso successive transizioni tecnologiche. Le copie, i trasferimenti e le versioni multiple testimoniano la stratificazione dei processi di lavoro e ne restituiscono la dimensione evolutiva. Le opere singole, al contrario, mostrano una distribuzione più lineare e concentrata, suggerendo una gestione unitaria e temporalmente circoscritta dei materiali. La varietà dei modelli di archiviazione osservati riflette dunque una relazione diretta tra pratiche autoriali, tecnologie impiegate e durata dei progetti narrativi, configurando l'archivio come una forma di autorappresentazione operativa e materiale del processo creativo.

9. Mediare il grafo: accesso e visualizzazione

Nell'ecosistema digitale, la descrizione archivistica non riguarda più solo la rappresentazione dei contenuti, ma anche il modo in cui i ricercatori accedono, esplorano e interpretano i dati (Langdon 2016). In questo senso, la descrizione offre la possibilità di mediare tra il grafo di conoscenza e le pratiche di ricerca, traducendone la complessità relazionale in forme di accesso comprensibili e navigabili anche da utenti non specialisti.

Il capitolo precedente è incentrato sulle possibilità analitiche della modellazione ontologica e sulle strategie di *distant* e *scalable reading* come modalità di esplorazione aggregata dei dati; in questo capitolo il focus si sposta sulle modalità di fruizione e lettura dei dati orientate all'usabilità. L'obiettivo è mostrare come un grafo possa essere reso comprensibile tramite interfacce e visualizzazioni più familiari, senza perdere coerenza concettuale e granularità semantica.

In questa prospettiva, il capitolo presenta le potenzialità di un'applicazione implementata attraverso ResearchSpace, piattaforma *open source* per la gestione e la pubblicazione di *knowledge graph* del patrimonio culturale. Vengono illustrate l'architettura e l'implementazione della piattaforma per la consultazione dei dati dell'Archivio Valerio Evangelisti, mostrando le principali funzionalità, insieme alle prime sperimentazioni di visualizzazione che evidenziano come i dati possano essere resi accessibili e facilmente esplorabili anche da utenti non specialisti.

9.1 La piattaforma sperimentale

Nel contesto di questa ricerca si è resa necessaria l'individuazione di un ambiente in grado di consentire la pubblicazione, l'esplorazione e l'interrogazione di dati descrittivi in formato RDF, garantendo al contempo la separazione tra la struttura dei dati e la loro modalità di presentazione. In questo senso, la gestione standard dei dati tramite endpoint SPARQL e la loro visualizzazione attraverso un'interfaccia dedicata rappresentavano requisiti tecnici essenziali per la selezione dell'ambiente da adottare.

Tra le piattaforme disponibili per la gestione e l'esplorazione dei dati RDF, è stata scelta ResearchSpace [261], in fase di testing presso il Digital Humanities Advanced Research Centre (/DH.ARC)[262] dell'Università di Bologna e pienamente conforme ai requisiti tecnici individuati. Questa scelta ha consentito al progetto dedicato all'Archivio Valerio Evangelisti di partecipare attivamente alla

[261] https://researchspace.org/.

[262] https://centri.unibo.it/dharc/en.

sperimentazione del /DH.ARC, valutandone al contempo l'efficacia rispetto agli obiettivi di rappresentazione e fruizione dei dati.

ResearchSpace è una piattaforma *open source* per la gestione, l'esplorazione e la pubblicazione di *knowledge graph* dedicati al patrimonio culturale. Sviluppata presso il British Museum con il supporto della Andrew W. Mellon Foundation e in partnership con Metaphacts[263], è definita dai suoi sviluppatori come una piattaforma «to help establish a community of researchers, where their underlying activities are framed by data sharing, active engagement in formal arguments, and semantic publishing» (Oldman e Tanase 2018, 325). Dal 2023, ResearchSpace è primariamente sviluppata e mantenuta da Kartography CIC[264], un'impresa sociale non-profit registrata come Community Interest Company in Inghilterra e Galles, ma presenta una forte componente di sviluppo collaborativo tramite il *repository* GitHub pubblico [265] , attraverso cui la comunità può interagire, proporre miglioramenti e contribuire all'evoluzione della piattaforma, favorendo una partecipazione attiva e condivisa.

L'architettura della piattaforma si basa sulle tecnologie del Web Semantico, utilizzando RDF per la rappresentazione dei dati e SPARQL come linguaggio di interrogazione[266]. ResearchSpace propone una concezione di *knowledge graph* intesa come «a continually changing informational structure that mediates between a human, the world and a computer. The graph itself is ontologically based and enhanced by human epistemology» (Oldman e Tanase 2018, 330).

Il sistema è strutturato in moduli (*templates*) integrabili e configurabili, ciascuno dei quali offre funzionalità specializzate che possono essere adattate alle esigenze dei singoli progetti, permettendo di combinare strumenti di analisi, visualizzazione e curatela in un ambiente personalizzabile. Ogni *template* è costruito attorno a una o più query SPARQL che recuperano i dati da visualizzare, che possono essere definite e modificate in base agli obiettivi del progetto. Sopra questo livello di interrogazione, il *template* specifica le modalità di presentazione dei dati estratti, consentendo un'ampia personalizzazione sia

[263] https://metaphacts.com/.

[264] https://kartography.org/index.html.

[265] https://github.com/researchspace.

[266] Il sistema integra di default Blazegraph come *triplestore backend* per l'archiviazione dei dati, ma è possibile configurare il collegamento anche con altri *triplestore*. L'interfaccia è costruita sulla *metaphactory knowledge graph platform*, che permette personalizzazione ed estensibilità dell'interazione con il database grafico attraverso l'uso di standard aperti. Il sistema utilizza template HTML5, React Components e Handlebars per la creazione di interfacce personalizzabili e integra strumenti esterni tra cui OntoDia per la visualizzazione di grafi, MIRADOR per immagini IIIF, e l'editor Ory per la creazione di narrative. ResearchSpace può essere implementata in diverse modalità: come istanza unica centralizzata oppure come sistema di più istanze confederate, che permettono di creare, condividere e collegare informazioni secondo diversi gradi e modelli di decentralizzazione (Oldman e Tanase 2018).

funzionale, nella logica degli obiettivi di visualizzazione e interazione, sia grafica. Tra questi, *Semantic Search* e *Semantic Forms* permettono di effettuare ricerche per contesto e per faccette; i *Knowledge Patterns* forniscono schemi ontologici riutilizzabili per aggiungere, recuperare o modificare relazioni complesse tra entità, mentre le *Knowledge Maps* consentono di visualizzare in forma di network connessioni tra attori, luoghi, eventi, oggetti e concetti. Altri moduli avanzati includono: l'*Image Viewer*, per il confronto e l'annotazione di immagini ad alta risoluzione; le *Semantic Narratives*, che combinano testo, visualizzazioni, tabelle e mappe di conoscenza aggiornandosi automaticamente al variare dei dati; le timeline dinamiche, utili per visualizzare eventi e contenuti temporali del dataset; dati geografici con Open Street Map (Kartography CIC 2023). I *template* sono configurabili in base alle esigenze dei progetti e la piattaforma consente anche la creazione di moduli completamente nuovi, progettati ad hoc per rispondere a specifici obiettivi di ricerca. Una volta sviluppati e testati, tali moduli possono essere condivisi con la comunità ResearchSpace, permettendo il riuso e l'adattamento a contesti diversi.

ResearchSpace è stata impiegata in molteplici contesti di ricerca, conservazione e gestione del patrimonio culturale, dimostrando flessibilità nell'adattarsi a domini eterogenei. La piattaforma è stata adottata da network culturali e consorzi, come LINCS[267] e Pharos[268], facilitando la collaborazione tra istituzioni internazionali nella condivisione di informazioni su oggetti, immagini e metadati. Nel settore archeologico e storico, progetti come *Amara West*[269] e *Deir el-Medina*[270] hanno sfruttato ResearchSpace per trasformare database di scavo in ambienti di ricerca semantica, in cui dati spaziali, materiali ed etnografici sono correlati e interrogati in modo integrato. Per quanto riguarda le collezioni museali, istituzioni come il British Museum e i National Archives UK hanno utilizzato la piattaforma per gestire dati di conservazione, *provenance* e ricerca (The National Archives 2021). In ambito artistico e bibliografico, ResearchSpace ha supportato progetti come *Late Hokusai*[271], *Hokusai: The Great Picture Book of Everything*[272], *Sphaera*[273], e la corrispondenza di Belle da Costa Greene con Bernard Berenson curata da Villa I Tatti[274], nonché iniziative come *VeNiss*[275], *RePIM - Repertorio della Poesia Italiana in Musica, 1500-1700*[276] e Florentia *Illustrata* (Andreose 2025), dove i dati storici e geografici sono stati

[267] https://lincsproject.ca/.
[268] https://pharosartresearch.org/.
[269] https://amara-west.researchspace.org/.
[270] https://deir-el-medina-dev.kartography.net/resource/rsp:ThinkingFrames.
[271] https://latehokusai.researchspace.org/resource/rsp:Start.
[272] https://hokusai-great-picture-book-everything.researchspace.org/resource/rsp:Start.
[273] https://db.sphaera.mpiwg-berlin.mpg.de/resource/Start.
[274] https://bellegreene.itatti.harvard.edu/resource/Start.
[275] https://veniss.eu/.
[276] https://repim.itatti.harvard.edu/resource/repim:formSearch.

modellati per analisi comparative e visualizzazioni del territorio. Complessivamente, questi esempi mostrano come la piattaforma consenta la gestione di *knowledge graph* complessi, la creazione di ambienti collaborativi e la fruizione di dati complessi attraverso interfacce intuitive, promuovendo un approccio multidisciplinare e partecipativo alla ricerca digitale nel patrimonio culturale.

ResearchSpace è stata sviluppata e testato adottando CIDOC CRM come *framework* ontologico di riferimento per la modellazione dei dati. Questa ontologia *event-based*, originariamente concepita per il dominio museale, è stata scelta perché consente di esprimere un elevato grado di granularità semantica e di strutturare le informazioni in un contesto ricco di relazioni. CIDOC CRM fornisce infatti un modello capace di integrare dati eterogenei e variabili, preservandone la specificità (Oldman e Tanase 2018, 327, 333).

In questo contesto, il caso di studio dell'Archivio Valerio Evangelisti ha rappresentato un'occasione concreta per mettere alla prova la flessibilità della piattaforma. Infatti, oltre alla finalità primaria di presentazione e accesso ai dati archivistici, il progetto ha consentito di sperimentare la prima implementazione di ResearchSpace basata su un'ontologia di natura archivistica, BoDi, consentendo, di fatto, di valutare la capacità della piattaforma di supportare RiC-O e verificarne la compatibilità con *framework* ontologici diversi da CIDOC CRM, per il quale il sistema è stato originariamente progettato.

9.2 Prime sperimentazioni di visualizzazione

L'interfaccia per l'accesso ai dati della partizione digitale dell'Archivio Valerio Evangelisti in ResearchSpace è stata progettata come primo esperimento per tradurre visualmente i cinque requisiti del modello concettuale adottato, convertendo principi astratti in strumenti di navigazione operativi[277]. Dal punto di vista metodologico, l'elemento essenziale risiede nell'adozione del *file system* come metafora di rappresentazione dei materiali archivistici. Tale scelta ha costituito il punto di partenza per la definizione di una struttura di riferimento capace di rendere comprensibile l'organizzazione interna dei contenuti digitali. L'obiettivo è stato dunque quello di proporre, come base prototipale, una visualizzazione ad albero espandibile in profondità, in grado di restituire la complessità delle relazioni semantiche sottostanti e di supportare percorsi di esplorazione alternativi rispetto alla sola dimensione

[277] La demo dell'applicazione è accessibile al link: http://evangelisti.dharc.unibo.it. Attualmente l'applicazione permette di visualizzare solo una porzione del dataset completo. L'intero corpus è in fase di revisione da parte dell'Associazione Valerio Evangelisti – "Il Sol dell'Avvenire". Il dataset sarà pubblicato progressivamente, con nuove porzioni del grafo che verranno rese disponibili non appena approvate. Il repository GitHub dedicato all'applicazione è disponibile al seguente indirizzo: https://github.com/LuciaGiagnolini12/ValerioEvangelisti_Project.

verticale. L'approccio combina la familiarità delle strutture gerarchiche tradizionali con il potenziale conoscitivo offerto dalle tecnologie semantiche.

All'interno dell'ecosistema ResearchSpace, il modulo *Semantic Tree*[278] costituisce lo strumento predefinito per la rappresentazione di relazioni gerarchiche tra entità RDF. Il suo funzionamento si basa su una o più query SPARQL che recuperano l'insieme delle risorse e delle loro relazioni, costruendo una struttura ad albero in cui ciascun nodo corrisponde a un'entità del grafo e i rami rappresentano le relazioni di appartenenza o dipendenza logica tra esse. Il *Semantic Tree basic* offre una visualizzazione di tipo statico: un'unica query iniziale recupera l'intero grafo delle relazioni gerarchiche, che viene poi reso disponibile all'utente attraverso funzioni di espansione e compressione dei singoli nodi. Questo approccio, sebbene efficace per insiemi di dati di dimensioni contenute o per ontologie con profondità limitata, presenta limiti significativi quando applicato a strutture archivistiche di grande complessità o estensione, come nel caso dell'Archivio Evangelisti. Inoltre, l'impostazione di base non consente di gestire in modo efficiente scenari in cui l'albero deve essere ricostruito dinamicamente in funzione di criteri di ricerca o filtraggio: ogni modifica della vista comporta il ricaricamento integrale della gerarchia, con la conseguente perdita del contesto di navigazione e dello stato di espansione dei nodi.

Per rispondere a queste esigenze strutturali e per adattare lo strumento alla complessità specifica del fondo Evangelisti, è stata sviluppata una versione estesa del modulo[279], denominata *Semantic Tree Advanced* (Grillo et al. 2026). La principale novità del modulo è il caricamento progressivo e contestuale dei dati. Invece di recuperare l'intero grafo delle relazioni in un'unica query, all'avvio vengono visualizzati solo la radice dell'albero e il primo livello di discendenza. Quando l'utente decide di espandere un nodo, il sistema interroga il database per recuperarne specificamente i nodi figli attraverso l'implementazione di un meccanismo di caricamento asincrono e progressivo. Ogni sottoalbero viene così generato esclusivamente in risposta all'interazione dell'utente, ottimizzando i tempi di elaborazione e consentendo la gestione di dataset di dimensioni considerevoli. Questa funzionalità ha superato i limiti del *Semantic Tree*, che non era in grado di sostenere il carico derivante dal caricamento simultaneo dell'intera struttura[280]. La selezione di un singolo elemento dell'albero consente di accedere a una vista di dettaglio della risorsa, nella quale vengono visualizzati le informazioni descrittive associate e i metadati completi.

[278] https://documentation.researchspace.org/resource/Help:SemanticTree.

[279] Lo sviluppo del nuovo modulo si deve a Remo Grillo, dottorando del XL ciclo del corso "Patrimonio culturale nell'ecosistema digitale" dell'Università di Bologna nell'ambito delle attività formative extra curriculari previste dal corso.

[280] Una volta espanso, il sistema conserva in memoria i dati del ramo dell'albero visualizzato, consentendo di velocizzare le operazioni nelle successive consultazioni.

Parallelamente al caricamento progressivo dei dati, *Semantic Tree Advanced* introduce una funzione di ricerca integrata che amplia le modalità di esplorazione dell'archivio. A differenza delle ricerche tradizionali, che restituiscono un elenco di risultati separato dalla struttura, questa ricerca opera direttamente sulla struttura ad albero: non produce una lista, ma filtra e riorganizza la gerarchia stessa, mettendo in evidenza solo i nodi pertinenti (Figura 9.1). Quando l'utente inserisce un termine nella barra di ricerca, il sistema genera una query SPARQL che seleziona i nodi corrispondenti e ricostruisce dinamicamente l'albero, visualizzando i risultati nel loro contesto archivistico. In questo modo è possibile identificare sia i documenti che soddisfano i criteri di ricerca sia la loro collocazione all'interno della struttura del fondo. La visualizzazione conserva quindi la gerarchia originaria dell'archivio e consente una consultazione continua, nella quale i risultati vengono evidenziati direttamente nella vista senza modificare la logica complessiva della navigazione.

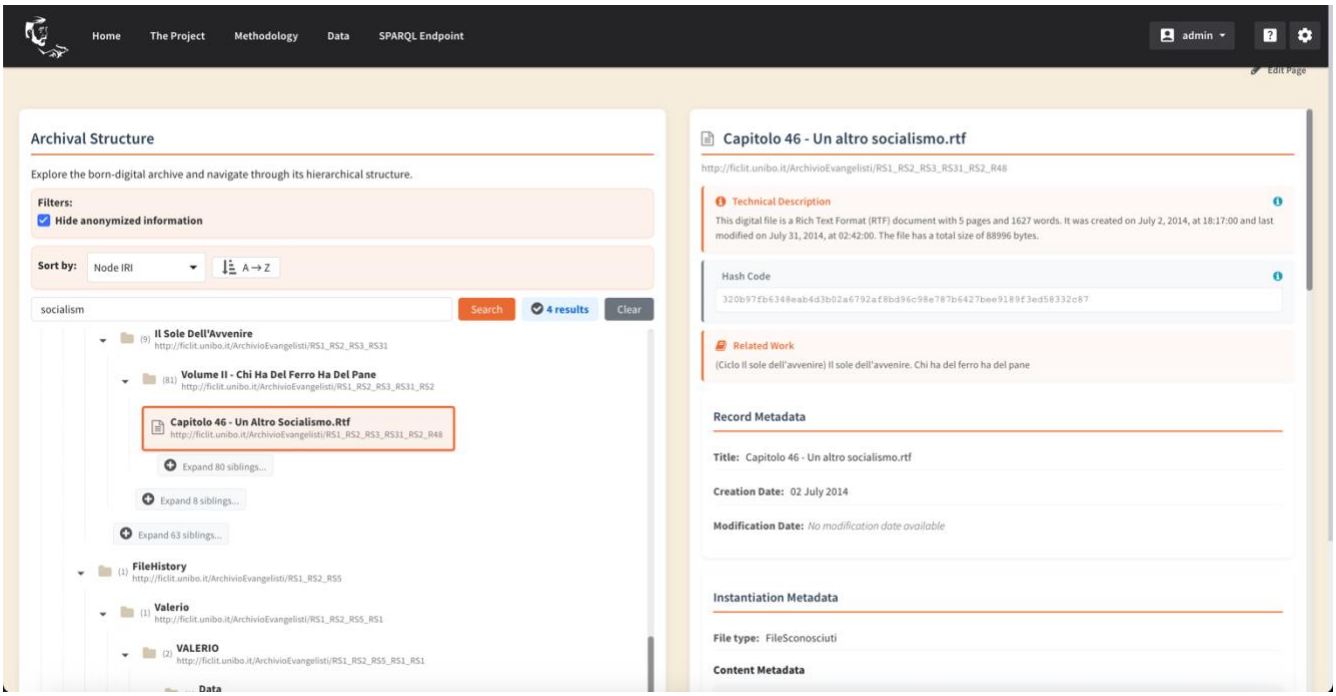


Figura 9.1. Esempio di visualizzazione di un risultato di ricerca contestualizzato all'interno della struttura gerarchica.

Accanto alla funzione di ricerca, il modulo integra un sistema di filtri reversibili e combinabili, volto a modulare la quantità e la tipologia dei nodi dell'albero visualizzati. I nodi che non soddisfano la condizione vengono temporaneamente nascosti, permettendo di evidenziare le informazioni di interesse calate nel contesto strutturale. Nel progetto Evangelisti, i filtri sono stati testati per escludere dall'albero i materiali anonimizzati per motivi di privacy. Tali record vengono comunque conservati nella struttura con metadati di base (dimensione, tipologia e date di creazione e modifica), come specificato al capitolo

7.5, permettendo analisi quantitative tramite query SPARQL. Il resto delle informazioni rimane anonimizzato e, per gli utenti non interessati a elaborazioni quantitative, la loro visualizzazione può essere disattivata tramite filtro, riducendo il rumore informativo.

Il modulo integra anche la funzione ordinamento che permette di riorganizzare la visualizzazione dei nodi secondo criteri definiti dall'utente. Attraverso un menu a tendina è possibile selezionare il campo su cui basare l'ordinamento (titolo, data di creazione, IRI, numero di elementi discendenti) e invertirne la direzione tra crescente e decrescente. L'operazione viene applicata ricorsivamente a tutti i livelli della gerarchia: all'interno di ciascun nodo padre, i nodi figli vengono riordinati secondo il criterio selezionato, preservando le relazioni di appartenenza ma modificando la sequenza di presentazione. In questo modo è possibile modulare la visualizzazione dei contenuti senza alterare la struttura logica del fondo.

Accanto alla possibilità di ordinare gli elementi dell'albero, il modulo ha implementato una funzionalità di particolare rilevanza, ossia il sistema dei nodi correlati, che consente di superare i vincoli della navigazione strettamente gerarchica aprendo percorsi di esplorazione trasversali. Al passaggio del mouse su ciascun nodo compare un'icona che, una volta selezionata, propone un menu di opzioni di correlazione configurabili in base alle caratteristiche del fondo come, ad esempio, elementi con la stessa data di creazione, creati con lo stesso software, aventi lo stesso codice hash (e dunque formalmente duplicati), oppure correlati al altri per ragioni funzionali o tematiche. Ogni criterio di correlazione è associato a una specifica query SPARQL che utilizza l'IRI del nodo come parametro di ricerca, interrogando il database per identificare le entità collegate attraverso relazioni semantiche non immediatamente visibili nella struttura ad albero. I nodi correlati individuati vengono automaticamente evidenziati all'interno della vista, permettendo all'utente di riconoscere connessioni che attraversano trasversalmente l'organizzazione archivistica ma in essa contestualizzata. Questa funzionalità si rivela essenziale per valorizzare la ricchezza del modello semantico sottostante: mentre la visualizzazione gerarchica riflette l'ordinamento fisico-logico del fondo, i collegamenti trasversali restituiscono la complessità delle relazioni tematiche, temporali o contestuali che legano i documenti tra loro. Nel caso dell'Archivio Evangelisti, ad esempio, è possibile navigare da un manoscritto verso altri materiali che fanno riferimento alla stessa opera di Evangelisti, che sono stati prodotti nello stesso periodo, o che condividono lo stesso hash code indipendentemente dalla loro collocazione nella struttura originaria. Vista la natura del dataset e il suo carattere così recente, non è stato purtroppo possibile valorizzare in questa chiave le relazioni interpersonali a causa della presenza di dati personali; tuttavia, la funzionalità dei collegamenti trasversali si rivela particolarmente interessante anche da questo punto di vista, soprattutto nel caso degli archivi storici in cui questi vincoli non sussistono più. Il sistema di nodi correlati rappresenta, dunque, un ponte

tra la dimensione verticale dell'ordinamento gerarchico e quella reticolare degli altri contesti, contribuendo a trasformare l'interfaccia in uno strumento di ricerca attivo che supporta la scoperta e l'interpretazione critica delle fonti.

L'insieme di questi strumenti, pur non costituendo singolarmente soluzioni tecnologiche inedite, configura un ambiente di consultazione che estende notevolmente le potenzialità del modulo preesistente. L'adattamento del *Semantic Tree*, originariamente concepito per strutture gerarchiche di limitata complessità, alle esigenze specifiche della consultazione archivistica rappresenta un primo risultato operativo nella direzione di un'interfaccia capace di coniugare familiarità d'uso e profondità semantica.

Il *Semantic Tree Advanced*, applicato alla visualizzazione dei dati della partizione digitale dell'Archivio Valerio Evangelisti, ha permesso di visualizzare i dati sottolineando la rappresentazione dei cinque requisiti definiti al capitolo 6.1: la rappresentazione dei materiali digitali nelle loro componenti fisiche, logiche e contenutistiche; la documentazione dell'integrità dei file; la valorizzazione e conservazione dei metadati nativi; la tracciabilità della *provenance* processuale e curatoriale; e la restituzione dei molteplici contesti, gerarchici e trasversali.

La struttura gerarchica adottata rispecchia l'organizzazione del *file system*, utilizzando questa metafora come strumento di rappresentazione dei materiali nativi digitali. Tale approccio consente di preservare e rendere intellegibile il contesto di provenienza dei dati, riproducendone almeno nelle linee generali la configurazione del *file system*. Parallelamente, la funzione di navigazione per nodi correlati permette di superare la dimensione strettamente verticale della gerarchia, aprendo percorsi di esplorazione che restituiscono le relazioni semantiche e contestuali che legano trasversalmente i materiali tra loro.

Dal punto di vista della struttura informativa, l'albero individua al livello più alto un raggruppamento generale dei materiali nativi digitali, che funge da contenitore per le tre principali tipologie di supporti documentati: i contenuti del computer principale utilizzato da Evangelisti, quelli dell'hard drive esterno e i riversamenti dei materiali conservati su floppy disk (Figura 9.2). Ciascuno di questi tre raggruppamenti replica l'organizzazione dei relativi supporti secondo le modalità descritte nei capitoli 6 e 7. Per il computer principale e per l'hard disk esterno, la visualizzazione riproduce la struttura delle cartelle e delle sottocartelle presenti sui dispositivi come pervenuti in ADLab. Anche per la collezione dei floppy disk viene rappresentata la struttura delle cartelle così come giunti in ADLab, con la differenza sostanziale che tale organizzazione non riflette la configurazione originaria dei supporti, ma costituisce una struttura generata dall'Associazione Valerio Evangelisti in seguito al versamento dei materiali. In questo caso, il primo livello rappresenta il raggruppamento complessivo della collezione, mentre le

sottocartelle immediatamente discendenti corrispondono virtualmente ciascuna a un singolo floppy disk, inteso come unità fisica contenente materiali.

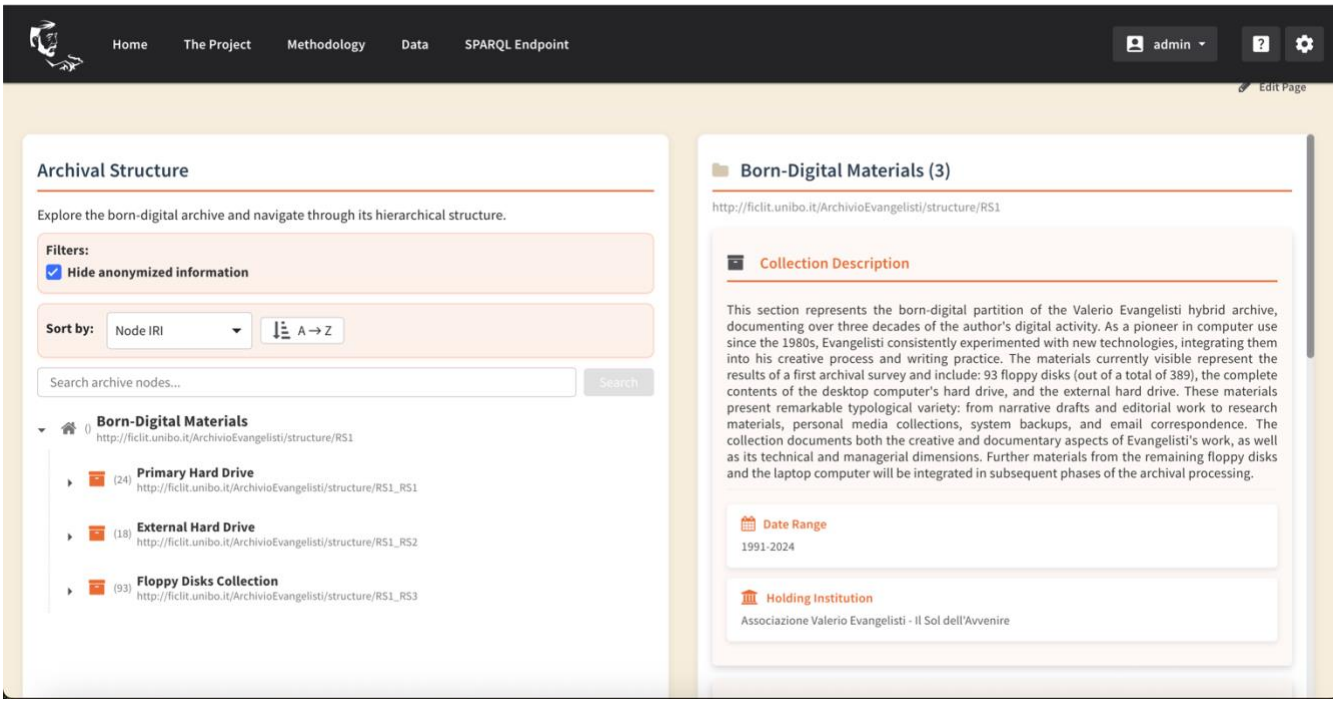


Figura 9.2. Struttura informativa dei materiali nativi digitali di Evangelisti, con distinzione tra computer principale, hard drive esterno e riversamenti da floppy disk.

Espandendo ciascun livello dell'albero è possibile navigare la gerarchia dei contenuti dei vari supporti; la selezione di un nodo consente di visualizzare, nella scheda a destra, la finestra di dettaglio che ne riporta la descrizione completa, includendo informazioni di *provenance* dei dati a diversi livelli di granularità. Per i tre raggruppamenti principali (*hard drives* e collezione di floppy disk), la scheda di dettaglio presenta una panoramica generale delle informazioni, con una descrizione sommaria dei contenuti, il numero di elementi, le dimensioni totali della raccolta e gli estremi cronologici. A questa si affianca la ricostruzione della storia archivistica dei materiali, intesa come documentazione dei passaggi di ambiente da supporto a supporto secondo quanto descritto nel capitolo 7.4. Ulteriori dettagli sulle modalità di migrazione, sulle tipologie di supporti coinvolti e sulla supervisione delle operazioni sono accessibili attraverso i box informativi associati (Figure 9.3, 9.4).

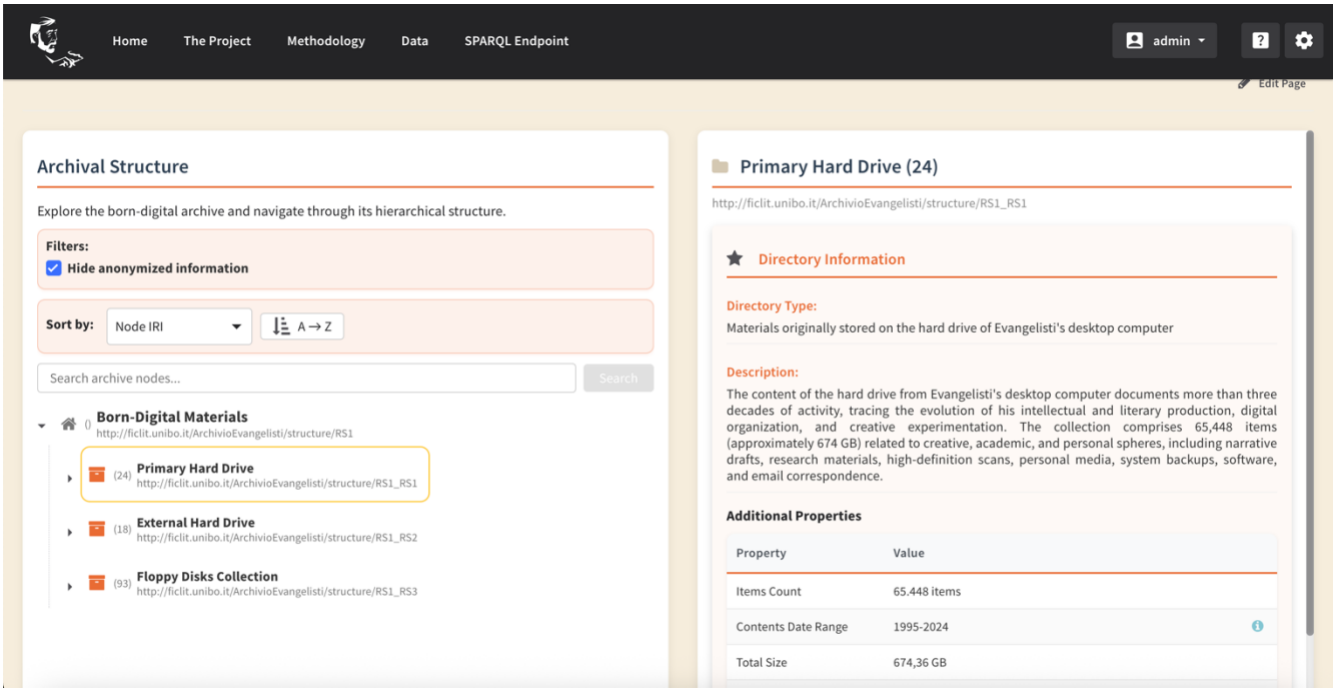


Figura 9.3. Struttura ad albero dei materiali digitali nativi di Evangelisti e finestra di dettaglio sui materiali dell'hard drive del computer di Evangelisti.

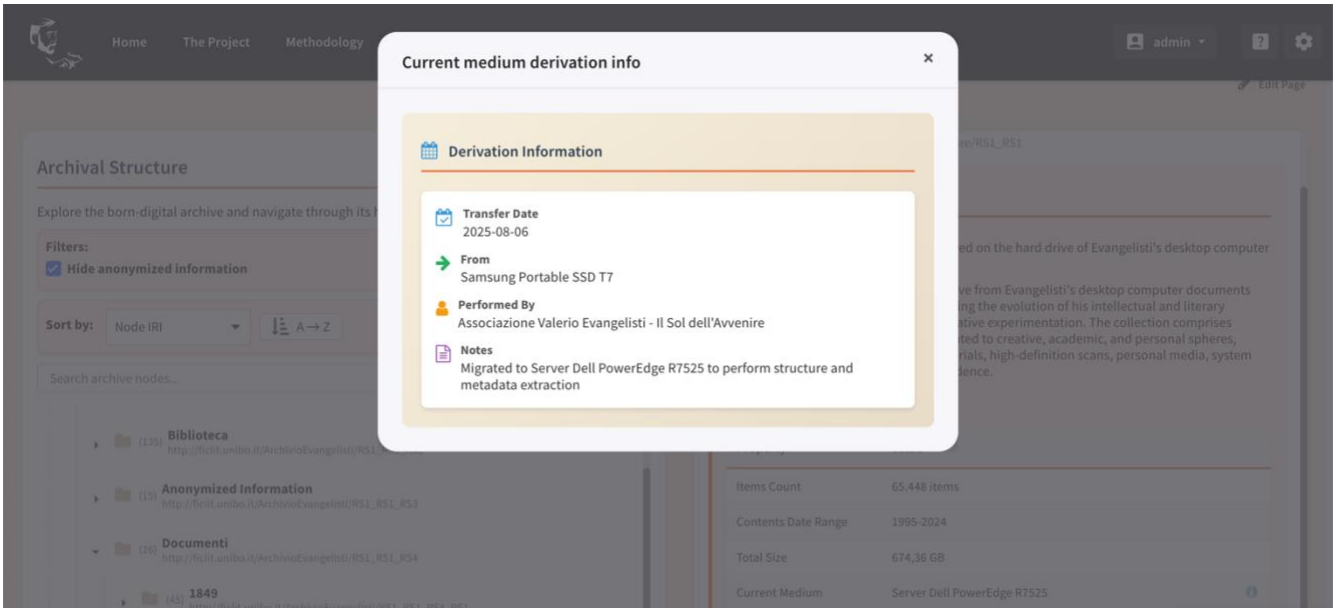


Figura 9.4. Visualizzazione delle informazioni di provenance e dei processi di migrazione tra supporti.

Per i nodi corrispondenti a cartelle e file, la scheda descrittiva visualizza una descrizione di anteprima generata tramite LLM (si veda il capitolo 7.3), informazioni sull'integrità del file (codice hash e relativa *provenance* di calcolo), l'insieme completo dei metadati tecnici estratti e, se presente, il riferimento all'opera di Evangelisti corrispondente. Anche in questo caso, dettagli sulle modalità e la data di

estrazione, sul software utilizzato e sulla supervisione delle operazioni sono consultabili attraverso i box informativi (Figura 9.5, 9.6, 9.7).

Infine, il passaggio del mouse sui nodi dell'albero evidenzia le possibili relazioni trasversali con altri materiali presenti nell'archivio: è possibile selezionare tutti i documenti o le cartelle che condividono lo stesso codice hash, lo stesso titolo, la stessa data di creazione o che risultano collegati alla medesima opera letteraria.

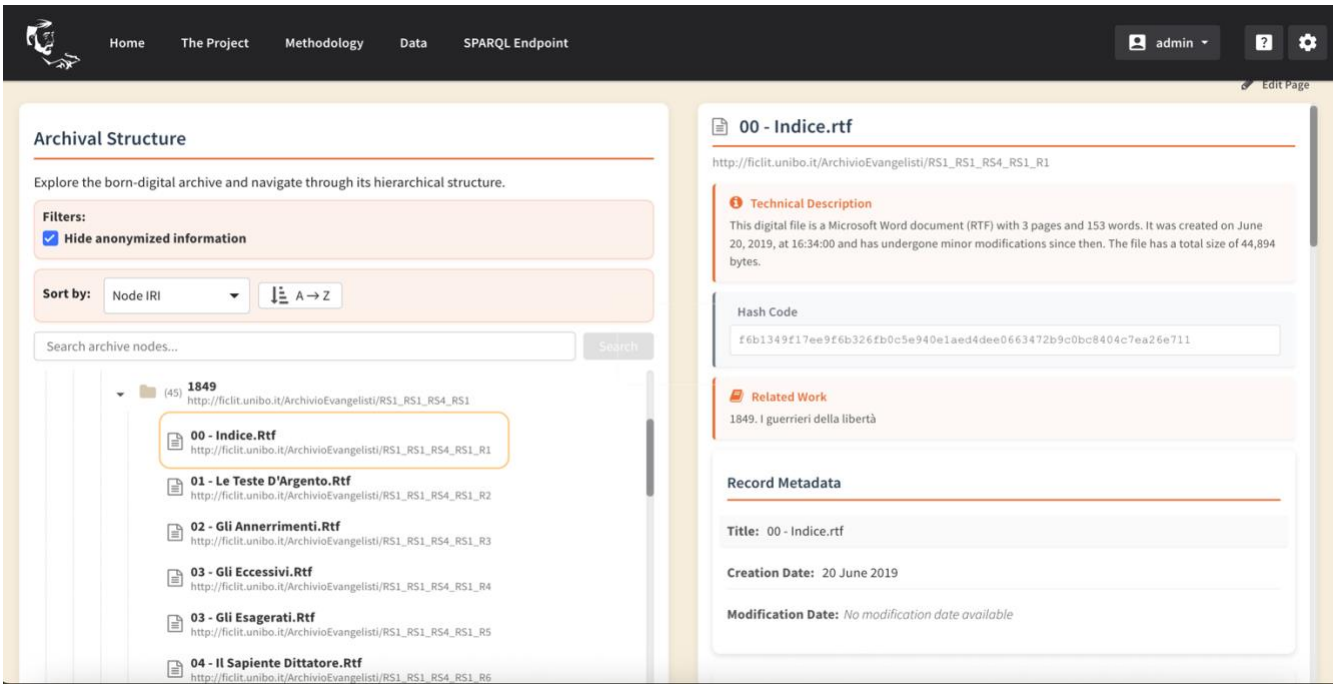


Figura 9.5. Scheda descrittiva del file "00-Indice.Rtf" con focus su informazioni descrittive di anteprima, riferimento al codice hash e all'opera correlata.

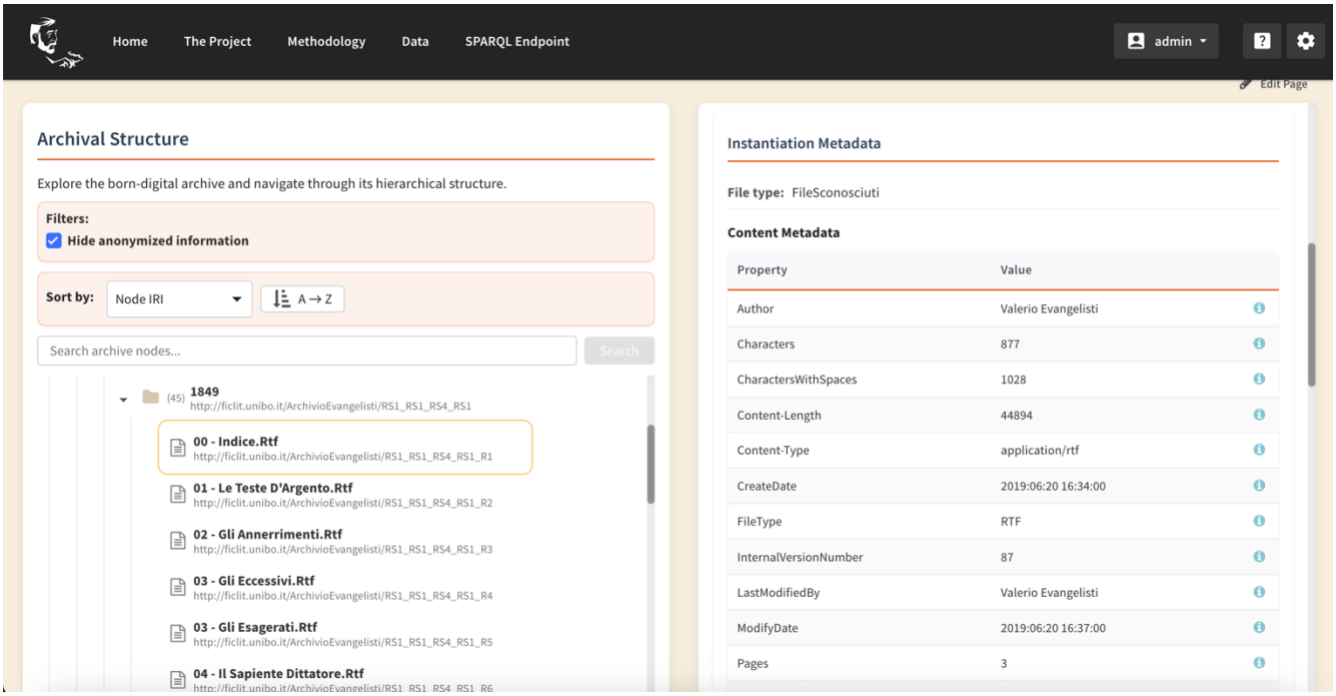


Figura 9.6. Scheda descrittiva del file "00-Indice.Rtf" con focus sui metadati tecnici.

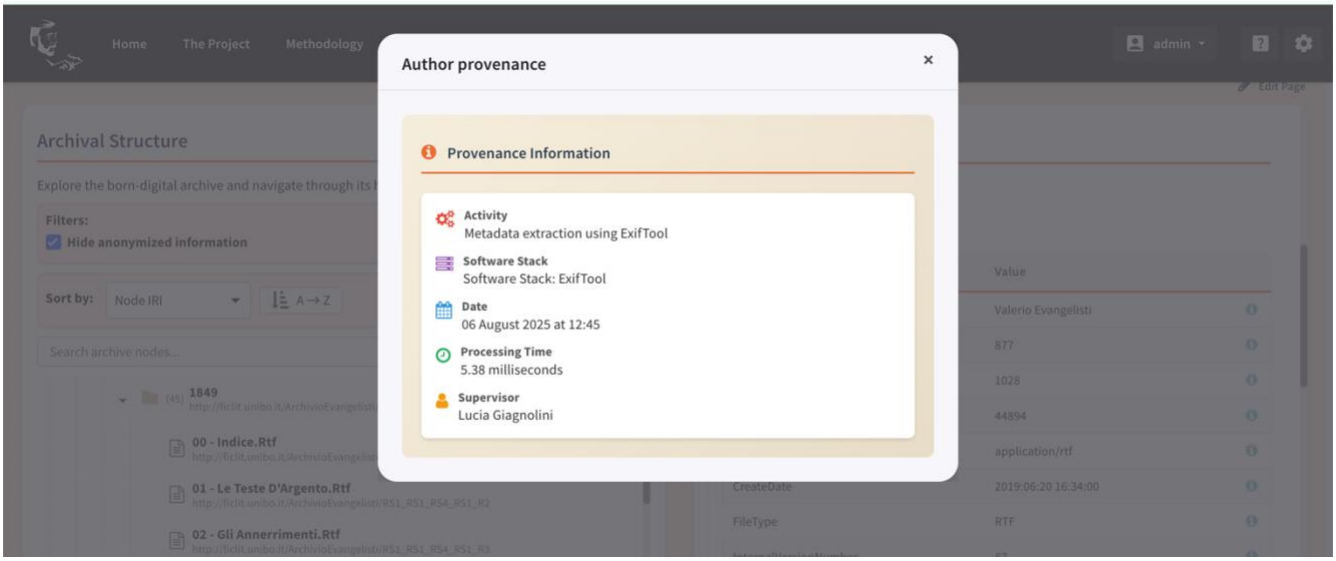


Figura 9.7. Informazioni di provenance relative al processo di estrazione dei metadati dal file "00 - Indice.rtf".

L'attuale visualizzazione costituisce soltanto una prima esplorazione delle potenzialità semantiche ed esplorative sia dei dati, sia della piattaforma, e necessita di ulteriori approfondimenti. Dal punto di vista tecnico, gli sviluppi futuri del sistema prevedono l'implementazione di un modulo di ricerca avanzata che permetta *query* più articolate attraverso la combinazione di criteri multipli, nonché la possibilità di visualizzare i risultati anche in formato lista, alternativo alla vista ad albero. Un ulteriore obiettivo riguarderà la gestione simultanea di più finestre di dettaglio dei metadati, mantenendo attiva la

corrispondenza visiva con la posizione delle risorse nell'albero gerarchico, al fine di supportare operazioni di confronto e analisi comparativa tra documenti. In una prospettiva più ampia, le attività si concentreranno sulla valorizzazione del dataset e della sua modellazione concettuale tramite l'implementazione di viste di accesso ai dati diversificate e tematiche, complementari alla visualizzazione gerarchica. La sperimentazione proseguirà nel corso del prossimo anno attraverso il completamento di questi obiettivi e mediante la curatela sistematica del dataset, con l'arricchimento delle informazioni contestuali e l'inclusione della descrizione dei materiali analogici del fondo. Tuttavia, già in questo stadio iniziale, il lavoro condotto ha permesso di dare conto delle potenzialità offerte dalla visualizzazione di un grafo di conoscenza al di là della rappresentazione canonica a network, spesso poco leggibile e difficilmente gestibile in contesti di consultazione archivistica. Il punto fondamentale risiede nella natura dei dati: essendo strutturati secondo un modello semantico formale ed espressi in RDF, non sono vincolati a un'unica forma di rappresentazione, ma possono essere interrogati e visualizzati attraverso modalità differenti in funzione degli obiettivi conoscitivi e delle esigenze di consultazione. Tramite ResearchSpace, la medesima base informativa può generare, infatti, viste alternative – gerarchiche, reticolari, cronologiche, tematiche – senza necessità di duplicazione o trasformazione dei dati sottostanti, poiché sono le query SPARQL a determinare quale porzione del grafo e quali relazioni rendere visibili in ciascun contesto. ResearchSpace, pur nella fase ancora iniziale della sperimentazione, si è dimostrato un ambiente particolarmente idoneo alla presentazione e alla valorizzazione di queste informazioni, offrendo una base standard per la gestione dei dati e la possibilità di personalizzare le interfacce secondo le specifiche esigenze del progetto.

Conclusioni

La presente ricerca ha esaminato il tema della rappresentazione degli archivi nativi digitali d'autore, con l'obiettivo di comprendere e descrivere la complessità di questi fondi in un contesto in cui la descrizione archivistica tradizionale sta evolvendo verso forme di modellazione a semantica esplicita. Attraverso un approccio metodologico integrato che ha combinato indagine fenomenologica, modellazione ontologica e sperimentazione applicativa sul caso di Valerio Evangelisti, lo studio ha fornito contributi sia sul piano teorico sia su quello applicativo, dimostrando la possibilità di integrare metodologie tradizionali con strumenti tecnologici innovativi.

Lo studio si è sviluppato attorno a quattro domande di ricerca principali (RQ1-RQ4), che hanno orientato l'intero percorso di dottorato: dalla raccolta e analisi dei dati alla modellazione, dall'applicazione del modello al caso di studio con tecniche di automazione fino alla rappresentazione dei risultati, consentendo di ottenere conoscenze, strumenti operativi e approcci metodologici potenzialmente replicabili in contesti analoghi.

RQ1 - Fenomenologia: come si configura un archivio d'autore contemporaneo?

Per rispondere a questa domanda, la ricerca ha sviluppato un approccio metodologico articolato su tre livelli complementari: la ricognizione dello stato dell'arte, l'indagine empirica su un campione di autori contemporanei e l'approfondimento sull'Archivio Valerio Evangelisti come caso di studio.

La definizione del concetto di "digitale d'autore" e la mappatura dei progetti e delle linee di ricerca sviluppate sul tema hanno fornito il quadro teorico di riferimento, permettendo di identificare il problema specifico della descrizione archivistica. L'analisi delle interviste a cinquanta finalisti dei Premi Strega e Campiello (1985-2024) ha evidenziato come gli autori contemporanei si configurino quali soggetti produttori d'archivio impegnati nello sviluppo di pratiche documentarie complesse e multiformi, articolate in una molteplicità di ambienti digitali. Emergono pattern gestionali ricorrenti, strategie di autorappresentazione e nuove forme di "volontà d'archivio", da cui derivano archivi distribuiti e stratificati, composti da materiali eterogenei accumulati nel tempo e accompagnati da una densità di metadati senza precedenti. Pur persistendo in minima parte il cartaceo, il digitale si impone come centro delle pratiche creative, con modalità diversificate di gestione, backup e trasmissione dei materiali.

Particolarmente problematica risulta la questione della trasmissione del patrimonio documentale digitale: la maggior parte degli autori appare perplessa di fronte alla possibilità di predisporre accortezze per garantire l'accessibilità futura dei propri materiali, soprattutto in assenza di chiare regolamentazioni. Si

riscontra, inoltre, una scarsa percezione del valore documentario del proprio archivio, confermando quanto emerso dalla letteratura internazionale sulle pratiche di *Personal Information Management* (PIM) e *Personal Digital Archiving* (PDA). Il ruolo degli autori si dimostra centrale per la futura preservazione dei loro archivi, soprattutto alla luce degli ostacoli informatici e legali che circoscrivono i margini di intervento delle istituzioni culturali. Tuttavia, tali margini non escludono un ruolo attivo delle istituzioni, che possono agire come attori di mediazione tra la produzione autoriale e le pratiche di conservazione, favorendo una riflessione condivisa sull'importanza del digitale d'autore. Le istituzioni culturali hanno infatti uno spazio d'azione importante nell'incentivare la collaborazione tra autori, eredi e archivi. Il loro compito non si esaurisce nella sola conservazione del materiale versato: comprendere i processi di creazione e le strategie adottate dagli autori consente di orientare le politiche di conservazione verso la ricezione e il recupero del digitale d'autore, migliorando la capacità descrittiva e interpretativa degli archivi.

La scelta di concentrare la raccolta dei dati sui finalisti dei premi Strega e Campiello, pur metodologicamente motivata, esclude una parte considerevole della produzione letteraria contemporanea, in particolare quella legata a generi e forme narrative meno canoniche. Nella composizione anagrafica del dataset, prevalentemente superiore ai quarant'anni, sono inoltre sottorappresentate le fasce d'età più avanzate e quelle più giovani. Per future ricerche sarebbe auspicabile un ampliamento del campione per includere una maggiore diversità, sia in termini di numero che di profilo degli autori, consentendo di esplorare una gamma più ampia di approcci e strategie creative e riflettendo in modo più completo le complessità del panorama autoriale contemporaneo.

In questa prospettiva, l'approfondimento sull'Archivio Valerio Evangelisti ha svolto una funzione parzialmente compensativa rispetto ai limiti del campione principale: l'autore, estraneo al circuito dei premi letterari *mainstream* e attivo nell'ambito della narrativa di genere, offre infatti una prospettiva complementare e utile per ampliare lo spettro delle pratiche autoriali indagate. L'analisi del suo archivio ha consentito di verificare sul campo le dinamiche emerse dalle interviste e di tradurre le osservazioni teoriche in sfide operative per la descrizione archivistica, offrendo un'esemplificazione concreta della complessità fenomenologica individuata dall'indagine generale.

RQ2 - Modellazione: come rappresentare il digitale d'autore?

La modellazione è stata sviluppata a partire dai requisiti individuati attraverso lo studio morfologico e fenomenologico del digitale d'autore. L'analisi delle sue peculiarità ha evidenziato la necessità di

distinguere tra caratteristiche intrinseche ed estrinseche dei documenti e di ripensare concetti fondamentali come metadati, *provenance* e contesto. Da questa riflessione sono emersi cinque requisiti principali per una rappresentazione a semantica esplicita del digitale d'autore: la necessità di individuare la stratificazione fisica, logica e concettuale dei file; di integrare misure crittografiche di integrità; di adottare un approccio scalabile ai metadati tecnici; di documentare la *provenance* e i molteplici livelli contestuali.

È seguita un'analisi dei principali modelli esistenti nel campo della descrizione archivistica (RiC-O, ArCo, ArchOnto e PREMIS) dalla quale la scelta è ricaduta su RiC-O, per la sua capacità di rappresentare reti complesse di relazioni tra entità archivistiche, agenti e contesti, e per il suo ruolo di standard internazionale emergente promosso dall'ICA. Sebbene RiC-O offra un solido quadro concettuale, non copre pienamente le esigenze descrittive del digitale d'autore. Per colmare questa lacuna è stata sviluppata BoDi (Born-Digital Ontology), un'estensione di RiC-O che integra elementi da PREMIS, PROV-O e LRMoo, arricchita da nuove classi e proprietà specifiche. BoDi consente di rappresentare archivi d'autore complessi, riconciliando standard di preservazione e di descrizione e trasformando i metadati tecnici in parte integrante della narrazione archivistica.

Il modello è stato validato formalmente attraverso diversi livelli di controllo: validazione sintattica RDF, verifica semantica tramite *reasoner*, analisi di conformità ontologica e sperimentazione sul caso Evangelisti.

La modellazione dell'Archivio Evangelisti ha consentito di esplorare la rappresentazione delle risorse archivistiche e del loro ecosistema metadatale, tracciando la *provenance* dei processi e abilitando percorsi di ricerca multipli di tipo gerarchico, tematico, cronologico, tipologico, relazionale, processuale e contestuale. La ricerca ha dimostrato come RiC-O, esteso con BoDi, costituisca un *framework* promettente per la rappresentazione degli archivi nativi digitali d'autore, risultando capace di coniugare rigore formale, interoperabilità e granularità descrittiva.

Pur confermando la validità concettuale e tecnica del modello, la sperimentazione ha evidenziato alcuni limiti. La validazione è stata condotta su un singolo caso di studio, sebbene complesso e articolato, circostanza che riduce la possibilità di generalizzare pienamente i risultati e di verificarne l'applicabilità ad archivi di diversa natura o scala. La presenza di materiali e dati personali ha inoltre limitato la possibilità di testare la rappresentazione delle relazioni interpersonali e di alcune dinamiche contestuali, riducendo la possibilità di indagare aspetti rilevanti dei processi creativi.

La validazione su ulteriori casi di studio sarà cruciale per verificare la generalizzabilità di BoDi, consentendo di affinare progressivamente l'ontologia e di individuare eventuali necessità di miglioramenti ed estensioni. L'integrazione dei materiali analogici nella modellazione permetterà, inoltre, di rappresentare archivi ibridi nella loro interezza, documentando le relazioni tra le componenti e offrendo una visione più completa del fondo.

RQ3 - Automazione: in che modo le proprietà intrinseche del digitale possono essere sfruttate per automatizzare e facilitare i processi di descrizione archivistica?

La sperimentazione sull'Archivio Valerio Evangelisti ha confermato la fattibilità operativa di una pipeline automatizzata in grado di trasformare il portato informativo di materiali digitali eterogenei in grafi RDF modellati su BoDi. Il *workflow* sviluppato si articola in cinque fasi integrate: sistematizzazione dei dati, validazione e arricchimento *rule-based*, integrazione di conoscenza specialistica e modelli generativi, ricostruzione del contesto originario e anonimizzazione dei dati personali.

L'applicazione della pipeline ha permesso di elaborare 2 TB di materiali distribuiti su supporti differenti, generando oltre sessanta milioni di triple e documentando più di 78.000 file e oltre 11.000 cartelle con metadati, relazioni e contesti di provenienza. Questo risultato evidenzia come l'automazione possa potenziare la scalabilità e la sistematicità dei processi descrittivi, liberando l'intervento umano dalle operazioni ripetitive e consentendo agli archivisti di concentrarsi sull'analisi contenutistica e contestuale dei materiali.

Al contempo, la sperimentazione ha messo in luce diversi limiti: l'elaborazione automatica di grandi moli di dati richiede competenze avanzate e infrastrutture hardware adeguate, condizioni che possono costituire una barriera per la diffusione del metodo in contesti istituzionali con risorse limitate. Inoltre, la validazione su un singolo caso di studio non consente di verificare pienamente né la robustezza del modello né la replicabilità della pipeline in archivi di diversa natura o scala.

La pipeline presenta margini significativi di sviluppo, che potranno essere esplorati in future implementazioni, ad esempio attraverso l'integrazione di funzionalità avanzate di estrazione testuale e di *Named Entity Recognition*, il perfezionamento dei moduli di normalizzazione e il potenziamento dei controlli di qualità sui metadati, inclusi eventuali allineamenti con altri modelli. Su un orizzonte temporale più ampio, un'implementazione standardizzata e documentata di procedure automatizzate come quella presentata potrebbe contribuire a definire buone pratiche per la descrizione degli archivi nativi digitali, combinando in modo operativo competenze di archivistica e informatica umanistica.

RQ4 - Visualizzazione: come presentare e analizzare i dati ottenuti?

L'analisi e la visualizzazione dei dati dell'Archivio Valerio Evangelisti sono state sviluppate secondo due modalità complementari: operazioni di *distant reading* basate su query SPARQL e rappresentazione della *knowledge base* tramite ResearchSpace.

Le operazioni di *distant reading* hanno permesso di individuare pattern complessi e relazioni trasversali difficilmente rilevabili con metodi convenzionali. Questo approccio valorizza l'analisi dei metadati e delle strutture archivistiche, trasformando una massa quantitativa potenzialmente ingestibile in una risorsa conoscitiva. Le analisi hanno incluso: statistiche generali e strutturali; distribuzione delle tipologie di media; identificazione dei codici hash come chiavi di esplorazione del fondo; monitoraggio dell'attività di Evangelisti nel tempo, con particolare attenzione alla produzione dei romanzi. Queste operazioni non solo hanno consentito di analizzare in profondità l'archivio, ma hanno anche fornito lenti inusuali di lettura, evidenziando configurazioni, tendenze e dinamiche altrimenti difficili da cogliere, dimostrando le potenzialità del *distant reading* d'archivio. Le analisi quantitative condotte rappresentano, tuttavia, solo una parte delle possibili interrogazioni del dataset; ulteriori approfondimenti potrebbero rivelare pattern e relazioni non ancora esplorati, ampliando la comprensione dei processi creativi e delle pratiche documentarie dell'autore.

Parallelamente, l'adozione di ResearchSpace ha dimostrato la possibilità di rendere fruibili i dati anche a utenti non specialisti, preservandone il portato informativo. La piattaforma consente di visualizzare un dataset a grafo in molteplici forme, non limitandosi alla rappresentazione canonica in nodi e archi. La sperimentazione si è concentrata, ad esempio, sulla possibilità di tradurre il grafo in forme gerarchiche più tradizionali, integrandole con prospettive trasversali e contestuali e garantendo il tracciamento della *provenance*. Pur trattandosi di primi esperimenti, e sebbene l'applicazione sia ancora in piena fase di sviluppo, il caso di studio risulta sufficiente per dimostrare come le visualizzazioni possano diventare strumenti per esplorare il potenziale semantico dei LOD e per avvicinare il linguaggio tecnico a un'esperienza d'uso intuitiva. L'implementazione, inoltre, costituisce il primo caso documentato in cui ResearchSpace è stato utilizzato con ontologie di stampo archivistico (BoDi e RiC-O) alternative a CIDOC CRM, confermando la flessibilità del software e la sua capacità di supportare *framework* ontologici differenti.

Complessivamente, i risultati di questo studio evidenziano come la convergenza dei quattro elementi chiave individuati dalle domande di ricerca possa costituire un approccio sistematico e integrato per affrontare le sfide poste dalla descrizione del digitale d'autore. In primo luogo, l'individuazione delle esigenze di rappresentazione consente di definire criteri funzionali e coerenti per la modellazione dei

dati. In secondo luogo, la modellazione mediante LOD e ontologie garantisce coerenza, interoperabilità e granularità nella rappresentazione delle relazioni e dei contesti. A ciò si aggiunge l'automazione dei processi, che permette l'estrazione sistematica di strutture, metadati e contesti, riducendo l'intervento manuale e aumentando la scalabilità delle operazioni. Infine, la progettazione delle modalità di restituzione contempla sia la possibilità di effettuare interrogazioni dirette sui dati, sia la definizione di interfacce e visualizzazioni web finalizzate a rendere i dati archivistici più accessibili anche a utenti non specialisti.

L'integrazione di questi elementi dà origine a un quadro metodologico articolato, attraverso cui il digitale d'autore può essere osservato, interpretato e compreso nella sua complessità. È tuttavia importante sottolineare che questa ricerca si configura come un tentativo esplorativo, volto a offrire una prospettiva di analisi e una metodologia di lavoro in un ambito ancora in fase di definizione, senza ambire a esaurire le molteplici sfaccettature della rappresentazione del digitale d'autore. Il percorso delineato rappresenta dunque un tassello all'interno di un discorso estremamente ampio e in continua evoluzione, che invita a ulteriori approfondimenti teorici e sperimentazioni applicative, capaci di ampliare, mettere in discussione e arricchire il quadro qui proposto.

Bibliografia

Tutti i collegamenti ipertestuali contenuti nel presente documento sono stati verificati e risultavano attivi alla data del 25 ottobre 2025

Acker, Amelia. 2021. «Emulation Practices for Software Preservation in Libraries, Archives, and Museums». *Journal of the Association for Information Science and Technology* 72 (9): 1148–60. https://doi.org/10.1002/asi.24482.

Agenzia per l'Italia Digitale (AgID). 2021. *Linee guida sulla formazione, gestione e conservazione di documenti informatici*. Agenzia per l'Italia Digitale. https://www.agid.gov.it/sites/default/files/repository_files/linee_guida_sul_documento_informatico.pdf.

AIMS. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*. https://escholarship.org/uc/item/1031p8xq.

Aissaoui, Khalid, Hafsa Ait idar, Hicham Belhadaoui, e Mounir Rifi. 2017. «Survey on data remanence in Cloud Computing environment». *2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 1–4. https://doi.org/10.1109/WITS.2017.7934624.

Alexander, Benjamin. 2015. «4 The Salman Rushdie Archive and the Re-Imagining of a Philological E-volution». In *Texts, Transmissions, Receptions*, a cura di André Lardinois, Sophie Levie, Hans Hoeken, e Christoph Lüthy. BRILL. https://doi.org/10.1163/9789004270848_006.

Alfier, Alessandro. 2017. «La classificazione archivistica: nuovi scenari d'uso tra web semantico e traditio degli esemplari digitali». JLIS.It: Italian Journal of Library, Archives and Information Science. Rivista Italiana Di Biblioteconomia, Archivistica e Scienza Dell'informazione. *JLIS.it: Italian Journal of Library, Archives and Information Science. Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione* 8 (2): 34–51.

Ali, Irfan, e Nosheen Fatima Warraich. 2020. «The relationship between mobile self-efficacy and mobile-based personal information management practices: A systematic review». *Library Hi Tech* 39 (1): 126–43. https://doi.org/10.1108/LHT-06-2019-0116.

Ali, Irfan, e Nosheen Fatima Warraich. 2021. «Personal information management through ubiquitous devices: Students' mobile self-efficacy and PIM practices». *Journal of Librarianship and Information Science* 54 (2): 174–87. https://doi.org/10.1177/0961000621992821.

Ali, Irfan, e Nosheen Fatima Warraich. 2022. «Modeling the Process of Personal Digital Archiving through Ubiquitous and Desktop Devices: A Systematic Review». *Journal of Librarianship and Information Science* 54 (1): 132–43. https://doi.org/10.1177/0961000621996410.

Ali, Irfan, e Nosheen Fatima Warraich. 2023. «Impact of personal innovativeness, perceived smartphone ease of use and mobile self-efficacy on smartphone-based personal information management practices». *The Electronic Library* 41 (4): 419–37. https://doi.org/10.1108/EL-12-2022-0262.

Allegrezza, Stefano. 2020. «Biblioteche e archivi personali in ambiente digitale: le sfide che si profilano all'orizzonte». In *Il privilegio della parola scritta: Gestione, conservazione e valorizzazione di carte e libri di persona*, a cura di Giovanni Di Domenico e Fiammetta Sabba. Associazione Italiana Biblioteche.

Allegrezza, Stefano. 2021. «Il problema dell'eredità digitale nella trasmissione di archivi e biblioteche personali». *Bibliothecae.it* 10 (1): 1. https://doi.org/10.6092/issn.2283-9364/13074.

Allegrezza, Stefano. 2024. «Il riversamento di formato elettronico tra standard internazionali e Linee guida sulla formazione, gestione e conservazione di documenti informatici dell'AgID». *DigItalia* 19 (1): 36–62. https://doi.org/10.36181/digitalia-00093.

Allegrezza, Stefano, Federico Boschetti, Emmanuela Carbé, et al. 2025. «Mapping for Understanding: the ALDiNa Project». *DigiCAM25: Born-Digital Collections, Archives and Memory* (London, UK), aprile.

Allegrezza, Stefano, e Luca Gorgolini, a c. di. 2016. *Gli archivi di persona nell'era digitale: il caso dell'archivio di Massimo Vannucci*. Percorsi. Convegno di studi «Gli archivi di persona nell'era digitale, il caso dell'archivio dell'on. Massimo Vannucci». Società editrice Il mulino.

Alon, Lilach, e Rafi Nachmias. 2020. «Gaps between Actual and Ideal Personal Information Management Behavior». *Computers in Human Behavior* 107 (giugno): 106292. https://doi.org/10.1016/j.chb.2020.106292.

Andreose, Erica. 2025. «Florentia Illustrata Knowledge Graph». Università di Bologna, gennaio. https://amsacta.unibo.it/id/eprint/8236/.

Apache Software Foundation. 2012. *Metadata Roadmap*. Apache Tika Project. https://cwiki.apache.org/confluence/display/TIKA/MetadataRoadmap.

Apache Software Foundation. 2024. *Apache Tika: A content analysis toolkit*. Versione 3.2.3. Apache Software Foundation, released. https://tika.apache.org/.

Athayde, Manaíra Aires, e Rejane Cristina Rocha. 2022. «Um arquivo para a literatura digital brasileira e algumas questões concretas». In *RE-AUTO-META ARQUIVO: Formas e Transformações do Arquivo*, 1ª ed., a cura di Manuel Portela e Daniela Côrtes Maduro, vol. 4. Fundação Fernando Pessoa.

«Avventura, non solo cinema: Nicolas Eymerich tra videogame e letteratura». 2013. *Cinefilia Ritrovata*, dicembre 17. https://www.cinefiliaritrovata.it/avventura-non-solo-cinema-nicolas-eymerich-tra-videogame-e-letteratura/.

Baader, Franz, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, e Peter F. Patel-Schneider, a c. di. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.

Babaei Giglou, Hamed, Jennifer D'Souza, e Sören Auer. 2023. «LLMs4OL: Large Language Models for Ontology Learning». In *The Semantic Web – ISWC 2023*, a cura di Terry R. Payne, Valentina Presutti, Guilin Qi, et al. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47240-4_22.

Bagley, Philip R. 1968. *Extension of programming language concepts*. University City Science Center.

Bailey, Jefferson. 2013. «Disrespect Des Fonds: Rethinking Arrangement and Description in Born-Digital Archives». *Archive Journal*, giugno. http://dev.archivejournal.net/?p=4722.

Bak, Greg. 2024. «Digital Provenance». *Archival Science* 24 (4): 847–69. https://doi.org/10.1007/s10502-024-09462-w.

Balkibayeva, Z. 2024. «Methods of Extracting and Analyzing Metadata for Evidentiary Purposes». *Uzbek Journal of Law and Digital Policy* 2 (5): 31–44. https://doi.org/10.59022/ujldp.233.

Banek, M., B. Vrdoljak, e A. M. Tjoa. 2008. «Word Sense Disambiguation as the Primary Step of Ontology Integration». In *Database and Expert Systems Applications*, a cura di S. S. Bhowmick, J. Küng, e R. Wagner, vol. 5181. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-540-85654-2_8.

Barr, Debra. 1987. «The fonds concept in the Working Group on Archival Descriptive Standards report». *Archivaria* 25: 163–70.

Barrera-Gomez, Julianna, e Ricky Erway. 2013. *Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house*. OCLC Research. https://doi.org/10.25333/C3F92C.

Barrero Junior, R. C. 2024. «Uma vida contada por computadores: a experiência com a organização de fontes nato-digitais no arquivo de Luiza Erundina». In *Arquivos pessoais: experiências no CPDOC*, 1ª ed., a cura di Celso Castro, Martina Spohr, e Thais Blank. FGV.

Bastian, Jeannette A. 2020. *Community Archives, Community Spaces : Heritage, Memory and Identity*. 1st ed.. Facet Publishing, UK.

Beaman, John. 2020. *Minimum Preservation Tool (MPT) - Digital Preservation Coalition*. luglio 29. https://www.dpconline.org/blog/minimum-preservation-tool-mpt.

Becker, Devin, e Collier Nogues. 2012. «Saving-Over, Over-Saving, and the Future Mess of Writers' Digital Archives: A Survey Report on the Personal Digital Archiving Practices of Emerging Writers». *The American Archivist* 75 (2): 482–513. https://doi.org/10.17723/aarc.75.2.t024180533382067.

Belhajjame, Khalid, James Cheney, David Corsar, et al. 2013. «PROV-O: The PROV Ontology». https://www.w3.org/TR/prov-o/.

Bellekens, Xavier, Greig Paul, James M. Irvine, et al. 2015. «Data remanence and digital forensic investigation for CUDA Graphics Processing Units». *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 1345–50. https://doi.org/10.1109/INM.2015.7140493.

Berners-Lee, Tim, James Hendler, e Ora Lassila. 2001. «The Semantic Web». *Scientific American* 284 (5): 34–43.

Bez, R., E. Camerlenghi, A. Modelli, e A. Visconti. 2003. «Introduction to flash memory». *Proceedings of the IEEE* 91 (4): 489–502. https://doi.org/10.1109/JPROC.2003.811702.

Biasiori, Lucio. 2018. «Prefazione alle Lettere». In *Tutte le opere. Secondo l'edizione di Mario Martelli (1971)*, a cura di Mario Martelli e Pier Davide Accendere. Bompiani.

Bizer, Christian, Tom Heath, e Tim Berners-Lee. 2011. «Linked Data: The Story So Far». In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, a cura di Amit Sheth. IGI Global.

Blanchette, Jean-François. 2011. «A Material History of Bits». *Journal of the American Society for Information Science and Technology* 62 (6): 1042–57. https://doi.org/10.1002/asi.21542.

Bodleian Libraries and Gardens e John Rylands Library. 2007. «PARADIGM Workbook on Digital Private Papers». University of Oxford and University of Manchester. https://wayback.archive-it.org/org-467/20170930070055/http://www.paradigm.ac.uk/.

Bolter, Jay David. 2001. *Writing Space : Computers, Hypertext, and the Remediation of Print / Jay David Bolter*. In *Writing Space Computers, Hypertext, and the Remediation of Print*, 2. ed. Lawrence Erlbaum Associates publishers.

Bolter, Jay David, e Michael Joyce. 1987. «Hypertext and creative writing». *Proceedings of the ACM Conference on Hypertext* (New York, NY, USA), HYPERTEXT '87, 41–50. https://doi.org/10.1145/317426.317431.

Borgo, Stefano, Roberta Ferrario, Aldo Gangemi, et al. 2022. «DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering». *Applied Ontology* 17 (1): 45–69. https://doi.org/10.3233/AO-210259.

Born-Digital Description – Joint Processing Guidelines. 2023. https://sites.harvard.edu/joint-processing-guidelines/description/born-digital-description/.

Bortolotti, Gherardo. 2008. «Blog e letteratura». *NAZIONE INDIANA*, dicembre 3. https://www.nazioneindiana.com/2008/12/03/blog-e-letteratura/.

Bouma, Jelle, Hugo Jonker, Vincent Van Der Meer, e Eddy Van Den Aker. 2023. «Reconstructing Timelines: From NTFS Timestamps to File Histories». *Proceedings of the 18th International Conference on Availability, Reliability and Security*, agosto 29, 1–9. https://doi.org/10.1145/3600160.3605027.

Braida, Lodovica, e Alberto Cadioli, a c. di. 2011. *Collezionismo librario e biblioteche d'autore: viaggio negli archivi culturali*. Vol. 5. Quaderni di APICE. Skira.

Brown, Tom B., Benjamin Mann, Nick Ryder, et al. 2020. «Language Models are Few-Shot Learners». *Advances in Neural Information Processing Systems 33*, 1877–901.

Bunn, Jenny. 2016. «Archival description and automation: a brief history of going digital». *Archives and Records* 37 (1): 65–78. https://doi.org/10.1080/23257962.2016.1145577.

Bunn, Jenny. 2021. *Born Digital Archive Cataloguing and Description*. Digital Preservation Coalition. https://doi.org/10.7207/twgn21-05.

Bureau of Canadian Archivists, Planning Committee on Descriptive Standards. 2008. *Rules for Archival Description (RAD)*. Revised July 2008. Canadian Council of Archives. https://archivescanada.ca/wp-content/uploads/2022/08/RADComplete_July2008.pdf.

Busby, Helen. 2024. «Dream big and ‘do different’: Digital Preservation Comes to Norfolk. - Digital Preservation Coalition». agosto 19. https://www.dpconline.org/blog/blog-helen-busby-24.

Cacopardi, Irene. 2023. «Internet e letteratura: la fine di un idillio?» *ENTHYMEMA*, fasc. 30 (gennaio): 105–17. https://doi.org/10.54103/2037-2426/19553.

Caimi, Heiko H. 2013a. «Intervista a Valerio Evangelisti, Parte II». Le Interviste. *Inkroci Magazine*. https://www.inkroci.it/racconti-brevi/interviste-a-scrittori-famosi/valerio-evangelisti-intervista.html.

Caimi, Heiko H. 2013b. «Intervista Valerio Evangelisti, Dall’Inquisizione alla Tortuga». Le Interviste. *Inkroci Magazine*. https://www.inkroci.it/racconti-brevi/interviste-a-scrittori-famosi/intervista-valerio-evangelisti.html.

Calvo, Marco, Gino Roncaglia, Fabio Ciotti, e Marco A. Zela. 1996. *Internet ’96: manuale per l’uso della rete*. I Robinson. Laterza.

Cannelli, B., e M. Musso. 2022. «Social Media as Part of Personal Digital Archives: Exploring Users’ Practices and Service Providers’ Policies Regarding the Preservation of Digital Memories». *Archival Science* 22: 259–83.

Canon. s.d. «Understanding EXIF and Metadata». Canon Georgia. https://www.canon.ge/pro/infobank/all-about-exif/.

Capussotti, Enrica. 2001. «Cyberpunk italiano: una comunità in rete (1997)». *Quaderni di Sociologia*, fasc. 26/27 (dicembre): 91–113. https://doi.org/10.4000/qds.1595.

Carbé, Emmanuela. 2023. *Digitale d’autore. Macchine, archivi, letterature*. Firenze University Press - USiena Press.

Carbé, Emmanuela. 2025. «Towards a critical understanding of born-digital literary archives: insights from Franco Fortini’s Collection». In *The intangible papers: authorial philology and born-digital texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino.

carmillaonline. 2004. «CARMILLA: Appello italiano per la liberazione di Cesare Battisti: le prime 1.500 firme». https://www.carmillaonline.com/archives/1500_firmatari.html.

carmillaonline. 2022. «Da oggi siamo un po’ più soli, Valerio ci ha lasciati». Carmilla on line, aprile 19. https://www.carmillaonline.com/2022/04/19/valerio-ci-ha-lasciato/.

Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, et al. 2019. «ArCo: The Italian Cultural Heritage Knowledge Graph». In *The Semantic Web – ISWC 2019*, a cura di Chiara Ghidini, Olaf Hartig, Maria Maleshkova, et al. Springer International Publishing.

Carroll, Laura, Erika Farr, Peter Hornsby, e Ben Ranker. 2011. «A Comprehensive Approach to Born-Digital Archives». *Archivaria*, dicembre 2, 61–92.

Casadei, Alberto. 2015. «Letteratura e web. Enciclopedia Italiana - IX Appendice (2015)». Treccani, Treccani. https://www.treccani.it/enciclopedia/letteratura-e-web_%28Enciclopedia-Italiana%29/.

Caswell, Michelle. 2021. «Dusting for fingerprints: Introducing feminist standpoint appraisal». *Journal of Critical Library and Information Studies* 3 (2).

Cesana, Roberta, e Fabio Desideri. 2025. «Archivi di carta, digitali, ibridi. Il Centro Apice e l'Archivio Contemporaneo "Alessandro Bonsanti"». In *Il futuro della memoria. Dove come cosa salvare*, a cura di Giuseppe Antonelli, Paola Italia, e Giacomo Papi. Faam. Fondazione Arnoldo e Alberto Mondadori. Fondazione Mondadori.

Charles, Vassilis Tzouvaras, e Antoine Isaac. 2020. *Innovating Metadata Aggregation in Europeana via Linked Data*. Europeana Pro.

Charmaz, K. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage Publications.

Clavaud, Florence. 2023. «Transform into Extension of CIDOC CRM». https://github.com/ICA-EGAD/RiC-O/issues/50#issuecomment-1508932062.

Clavaud, Florence, Thomas Francart, e Pauline Charbonnier. 2023. «RiC-O Converter: A Software to Convert EAC-CPF and EAD 2002 XML Files to RDF Datasets Conforming to Records in Contexts Ontology». *Journal on Computing and Cultural Heritage* 16 (3): 42:1-42:13. https://doi.org/10.1145/3583592.

Clavaud, Florence, e Tobias Wildi. 2021. «ICA Records in Contexts-Ontology (RiC-O): A Semantic Framework for Describing Archival Resources». *Linked Archives 2021: Proceedings of Linked Archives International Workshop 2021*, 79–92.

Clemens, Alison, Matthew Gorham, Jonathan Manton, Cate Peebles, e Jessica Quagliaroli. 2020. «Yale University Library Research Guides: Born Digital Archival Description Guidelines: Home». https://guides.library.yale.edu/c.php?g=934566&p=6736587.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, e Julia Noordegraaf. 2021. «Archives and AI: An Overview of Current Debates and Future Perspectives». *Journal on Computing and Cultural Heritage* 15 (1): 4:1-4:15. https://doi.org/10.1145/3479010.

Coleman, Gabriella. 2013. *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton University Press.

Colón-Marrero, Elena, e Allison Hughes. 2015. «Toni Morrison's Born-Digital Material». agosto 26. https://blogs.princeton.edu/mudd/2015/08/toni-morrisons-born-digital-material/.

Conrad, Eric, Seth Misenar, e Joshua Feldman. 2010. «Chapter 10 - Domain 9: Operations security». In *CISSP Study Guide*, a cura di Eric Conrad, Seth Misenar, e Joshua Feldman. Syngress. https://doi.org/10.1016/B978-1-59749-563-9.00010-X.

Cook, Terry. 1993. «The concept of the archival fonds in the post-custodial era: theory, problems and solutions». *Archivaria*, 24–37.

Copeland, Andrea J. 2011. «Analysis of Public Library Users' Digital Preservation Practices». *Journal of the American Society for Information Science and Technology* 62 (7): 1288–300. https://doi.org/10.1002/asi.21553.

Corriere della Sera. 2009. «Lista comunista e anticapitalista - Elezioni europee 2009». RCS Quotidiani S.p.A., maggio 21. https://www.corriere.it/politica/elezioni_2009/lista_comunista_anticapitalista.shtml.

Crasson, Aurèle, Laurent Alonso, Jean-Louis Lebrave, e Jeremy Pedrazzi. 2025. «Derrida Hexadécimal: The Forensic Genesis of *Le Toucher, Jean-Luc Nancy*». In *The Intangible Papers. Authorial Philology and Born-Digital Texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino. https://doi.org/10.1401/9788815414328/c6.

Cunningham, Adrian. 1999. «Waiting for the ghost train: strategies for managing electronic personal records before it is too late». *Archival Issues* 24 (1): 55–64.

Curino, Luciano. 1983. «Scusi, lei lo scriverebbe un romanzo con il computer?» *Tuttolibri*, giugno 5.

Daelemans, Walter. 2013. «Explanation in Computational Stylometry». In *Computational Linguistics and Intelligent Text Processing*, a cura di Alexander Gelbukh. Springer Berlin Heidelberg.

Daines, J. Gordon. 2013. «Processing Digital Records and Manuscripts». In *Archival Arrangement and Description*, a cura di Christopher J. Prom e Thomas J. Frusciano. Society of American Archivists.

Damiani, Concetta. 2022. «Archival Description and Conceptual Transversality». *JLIS.It* 13 (3): 3. https://doi.org/10.36253/jlis.it-485.

Dang, Quynh. 2012. *Recommendation for Applications Using Approved Hash Algorithms*. NIST Special Publication Nos. 800-107 Revision 1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-107r1.

Daquino, Marilena. 2021. «Linked Open Data Native Cataloguing and Archival Description». *JLIS.It* 12 (3): 3. https://doi.org/10.4403/jlis.it-12703.

Daquino, Marilena, Francesco Mambelli, Silvio Peroni, e Fabio Vitali. 2016. «Representing the Zeri Photo Archive Using CIDOC-CRM and FRBRoo».

Daquino, Marilena, e Francesca Tomasi. 2015. «Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects». In *Metadata and Semantics Research*, a cura di Emmanouel Garoufallou, Richard J. Hartley, e Panorea Gaitanou. Communications in Computer and Information Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37.

De Coulon, Baptiste. 2024. «Déploiement de la norme Records in Contexts pour la gestion des collections de la Fondation SAPA». *Revue électronique suisse de science de l'information (RESSI)*, fasc. 24. https://doi.org/10.55790/journals/ressi.2024.e1511.

Dean, Jeff, Sanjay Ghemawat, e Google. 2024. «LevelDB: A Fast Key-Value Storage Library». https://github.com/google/leveldb.

Di Corinto, Arturo. 2017. «Hacker e BBS, Centri Sociali, Reti Civiche: Chi (Prima Di Internet) Ha Diffuso La Telematica…». *Medium*, febbraio 1. https://arturodicorinto.medium.com/hacker-e-bbs-centri-sociali-reti-civiche-chi-prima-di-internet-ha-diffuso-la-telematica-e2e509741594.

Di Corinto, Arturo, e Tommaso Tozzi. 2002. *Hacktivism: la libertà nelle maglie della rete*. Indagini. Manifestolibri.

Di Marcantonio, Giorgia. 2023. «From Record to Data. New Purposes for Archival Description Processes». *JLIS.It*, pubblicazione online ad accesso anticipato, maggio 22. https://doi.org/10.36253/jlis.it-549.

Di Marcantonio, Giorgia. 2024. «Intelligenza artificiale, Large Language Models (LLMs) e Retrieval-Augmented Generation (RAG). Nuovi strumenti per l'accesso alle risorse archivistiche e bibliografiche». *Bibliothecae.it* 13 (1): 1. https://doi.org/10.6092/issn.2283-9364/19982.

Digital Preservation Coalition. 2024. *Digital Preservation Handbook*. 3rd ed. https://www.dpconline.org/handbook.

Dilley, R. M. 2002. «The Problem of Context in Social and Cultural Anthropology». *Language & Communication* 22 (ottobre): 437–56. https://doi.org/10.1016/S0271-5309(02)00019-8.

Donig, Simon, Markus Eckl, Sebastian Gassner, e Malte Rehbein. 2023. «Web archive analytics: Blind spots and silences in distant readings of the archived web». *Digital Scholarship in the Humanities* 38 (3): 1033–48. https://doi.org/10.1093/llc/fqad014.

Douglas, Jennifer. 2010. «Origins: Evolving Ideas about the Principle of Provenance». In *Currents of Archival Thinking*, a cura di Heather MacNeil e Terry Eastwood. Libraries Unlimited.

Dryden, Jean. 1995. «Archival Description of Electronic Records: An Examination of Current Practices». *Archivaria* 40 (ottobre): 99–108.

Duranti, Luciana. 1997. «The Archival Bond». *Archives and Museum Informatics* 11 (3): 213–18. https://doi.org/10.1023/A:1009025127463.

Duranti, Luciana, e Randy Preston. 2008. *Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*. Associazione Nazionale Archivistica Italiana.

ECMA International. 2016. *Office Open XML File Formats*. ECMA-376. 5th ed. ECMA International. https://ecma-international.org/publications-and-standards/standards/ecma-376/.

Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, e others. 2024. «From local to global: A graph rag approach to query-focused summarization». *arXiv preprint arXiv:2404.16130*.

Edwards, Brenna. 2025. «Bits and Terabytes: Born-Digital Archives at the Harry Ransom Center». In *The Intangible Papers: Authorial Philology and Born-Digital Texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino. https://doi.org/10.1401/9788815414328/c4.

Ennen, Inga, Daniel Kappe, Thomas Rempel, Claudia Glenske, e Andreas Hütten. 2016. «Giant Magnetoresistance: Basic Concepts, Microstructure, Magnetic Interactions and Applications». *Sensors (Basel)* 16 (6): 904. https://doi.org/10.3390/s16060904.

Erway, Ricky. 2010. *Defining "Born Digital": An Essay*. OCLC Research.

Erway, Ricky. 2012. *You've Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media*. OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-06.pdf.

Evangelisti, Valerio. 2001a. *Alla periferia di Alphaville: interventi sulla paraletteratura*. L'ancora del Mediterraneo.

Evangelisti, Valerio. 2001b. «R: [Eymerich] ritorno da genova (ot)». luglio 23.

Evangelisti, Valerio. 2008. «Le origini della mailing list (era [ML Eymerich] Apocrifo Eymerichiano)». dicembre 17.

Experts Group on Archival Description. 2016. *Records in Contexts: A Conceptual Model for Archival Description. Consultation Draft v. 01*. Conceptual Model. International Council on Archives.

Facebook. 2024. «RocksDB: A Persistent Key-Value Store for Fast Storage Environments». https://github.com/facebook/rocksdb.

Feliciati, Pierluigi. 2021. «Archives in a Graph. The Records in Contexts Ontology within the Framework of Standards and Practices of Archival Description». *JLIS.It* 12 (1): 92–101. https://doi.org/10.4403/jlis.it-12675.

Fernando, Vihara. 2021. «Cyber Forensics Tools: A Review on Mechanism and Emerging Challenges». *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 1–7. https://doi.org/10.1109/NTMS49979.2021.9432641.

Ferretti, Gian Carlo. 1985. «La scrittura elettronica». In *Pubblico 1985. Produzione letteraria e mercato culturale*, a cura di Vittorio Spinazzola. Milano Libri Edizioni.

Fiormonte, Domenico. 1996. «Il Computer e la Scrittura. Forme e Limiti di un Influsso». In *Lingua Letteratura Computer*, a cura di M. Ricciardi. Bollati Boringhieri.

Fiormonte, Domenico. 1997. «Antologia (e Archeologia) della Scrittura Elettronica: Tre Tappe di un Processo in Corso». In *Modi di Scrivere. Tecnologie e Pratiche della Scrittura dal Manoscritto al CD-ROM*, a cura di Claudio Leonardi, Marcello Morelli, e Francesco Santi. Centro di Studi sull'Alto Medioevo.

Fiormonte, Domenico. 2003. *Scrittura e Filologia nell'Era Digitale*. Bollati Boringhieri.

Forlani, Francesco. 2018. «Intervista a Valerio Evangelisti». ottobre 22. https://www.doppiozero.com/intervista-valerio-evangelisti.

Forstrom, Michael. 2009. «Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library». *The American Archivist* 72 (2): 460–77.

FORTH e CRM SIG. 2016. *Definition of the CRMdig: An Extension of CIDOC-CRM to Support Provenance Metadata*. https://www.cidoc-crm.org/crmdig/sites/default/files/CRMdig_v3.2.1.pdf.

Francart, Thomas. 2024. «RiC-O Converter». https://archivesnationalesfr.github.io/rico-converter/en/.

Furner, Jonathan. 2020. «Definitions of "Metadata": A Brief Survey of International Standards». *Journal of the Association for Information Science and Technology* 71 (6): E33–42. https://doi.org/10.1002/asi.24295.

Gambaro, F. 2004. «Una Vita da Editore. Intervista a Raffaele Crovi». In *Tirature '04: Che Fine ha Fatto il Postmoderno?*, a cura di Vittorio Spinazzola. Il Saggiatore/Fondazione Arnoldo e Alberto Mondadori.

Gangemi, Aldo, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, e Misael Mongiovì. 2017. «Semantic Web Machine Reading with FRED». *Semantic Web* 8 (6): 873–93. https://doi.org/10.3233/SW-160240.

Gao, Yunfan, Yun Xiong, Xinyu Gao, et al. 2024. «Retrieval-Augmented Generation for Large Language Models: A Survey». https://arxiv.org/abs/2312.10997.

Garcelon, Marc. 2006. «The `Indymedia' Experiment: The Internet as Movement Facilitator Against Institutional Control». *Convergence* 12 (1): 55–82. https://doi.org/10.1177/1354856506061554.

García-González, Herminio, e Mike Bryant. 2023. «The Holocaust Archival Material Knowledge Graph». In *The Semantic Web – ISWC 2023*, a cura di Terry R. Payne, Valentina Presutti, Guilin Qi, et al. Springer Nature Switzerland.

Garofalo, Mauro. 2008. «la "tortuga" di evangelisti: pirati analogici e hacker digitali». *Generazione X 2.0*, ottobre 30. https://maurogarofalo.nova100.ilsole24ore.com/2008/10/30/la-tortuga-di-e/.

Gartner, Richard. 2016. «What Metadata Is and Why It Matters». In *Metadata: Shaping Knowledge from Antiquity to the Semantic Web*. Springer International Publishing. https://doi.org/10.1007/978-3-319-40893-4_1.

Gavin, Michael. 2022. *Literary Mathematics: Quantitative Theory for Textual Studies*. Stanford University Press.

Giagnolini, Lucia. 2018. «L'Archivio Massimo Vannucci e il dibattito sulla selezione dal cartaceo al digitale». *Clionet: per un senso del tempo e dei luoghi* 2. https://rivista.clionet.it/vol2/societa-e-cultura/archivi_vivi/giagnolini-archivio-massimo-vannucci.

Giagnolini, Lucia. 2023. «Verso una "volontà d'archivio" digitale». *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD.*, 92–98. https://cris.unibo.it/handle/11585/945614?mode=complete.

Giagnolini, Lucia. 2025a. «Nuovi archivi d'autore». In *Il futuro della memoria. Dove come dove salvare*, a cura di Giuseppe Antonelli, Paola Italia, e Giacomo Papi. Fondazione Mondadori.

Giagnolini, Lucia. 2025b. «Riflessi Digitali: Cinquanta interviste sugli archivi d'autore contemporanei». [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15063775

Giagnolini, Lucia, e Primo Baldini. 2025. «Literary Bytes: New Approaches to Born-Digital Archives at Pavia Archivi Digitali». In *The Intangible Papers. Authorial Philology and Born-Digital Texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino.

Giagnolini, Lucia, Paolo Bonora, e Francesca Tomasi. 2024. «Affinare il contesto: Estrazione di informazioni strutturate per l'arricchimento dei contesti archivistici». In *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti*. Associazione per l'Informatica Umanistica e la Cultura Digitale.

Giagnolini, Lucia, e Mariangela Giglio. 2024. «Il Futuro della Memoria: Dove, Come, Cosa Salvare / The Future of Memory: Where, How, What to Save (Milano, 5 novembre 2023)». *Ecdotica* 1: 458–62. https://doi.org/10.7385/116720.

Giagnolini, Lucia, e Inês Koch. 2024. «Semantic representation of the Registos de Baptismos da Paróquia de Aldoar (Porto, Portugal)». Institute for Systems and Computer Engineering of Porto. https://doi.org/10.25747/15YG-GD86.

Giagnolini, Lucia, Inês Koch, Francesca Tomasi, e Carla Teixeira Lopes. 2025. «Comparative insights into semantic archival modelling: evaluating RiC-O and ArchOnto representation capabilities». *Journal of Documentation* 81 (4): 1003–31. https://doi.org/10.1108/JD-12-2024-0310.

Giagnolini, Lucia, Andrea Schimmenti, Paolo Bonora, e Francesca Tomasi. 2025. «Expliciting Contexts: Semantic Knowledge Extraction from Traditional Archival Descriptions». *Umanistica Digitale* 9 (20): 115–44. https://doi.org/10.6092/issn.2532-8816/21229.

Giglio, Mariangela. 2025a. «Metodologie computazionali per l'organizzazione di archivi nati digitalmente». *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Atti del XIV Convegno Annuale AIUCD* (Verona), giugno, 208–14. https://aiucd2025.dlls.univr.it/assets/pdf/papers/50.pdf.

Giglio, Mariangela. 2025b. «Metodologie computazionali per l'organizzazione di archivi nati digitalmente». In *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Proceedings del XIV Convegno Annuale AIUCD2025*, a cura di Serena Rebora, Matteo Rospocher, e Silvia Bazzaco. AIUCD. https://doi.org/10.6092/unibo/amsacta/8380.

Gillean, Dan. 2017. «Artefactual response to RiC-CM Draft». gennaio 25. https://groups.google.com/g/ica-atom-users/c/QwSor7OQ90U.

Gilliland, Anne J. 2016. «Setting the stage». In *Introduction to Metadata*, 3rd ed., a cura di Murtha Baca. Getty Research Institute. https://www.getty.edu/publications/intrometadata/setting-the-stage/.

Girolami, Andrea. 2011. «L'intervista obliqua a Valerio Evangelisti nel 2011: Finché c'è lotta informatica c'è speranza». Wired Italia, gennaio 24. https://www.wired.it/video/watch/valerio-evangelisti-fincha-ca-lotta-informatica-ca-speranza-interviste-oblique.

Gladney, Henry. 2007. *Preserving Digital Information*. Springer Science & Business Media.

Glimm, Birte, Ian Horrocks, Boris Motik, Giorgos Stoilos, e Zhe Wang. 2014. «HermiT: An OWL 2 Reasoner». *Journal of Automated Reasoning* 53 (3): 245–69. https://doi.org/10.1007/s10817-014-9305-1.

Gooding, Paul, Jos Smith, e Justine Mann. 2019. «The forensic imagination: interdisciplinary approaches to tracing creativity in writers' born-digital archives». *Archives and Manuscripts* 47 (3): 374–90. https://doi.org/10.1080/01576895.2019.1608837.

Gorini, Adele, e Lucia Giagnolini. 2025. «Under the Surface: Interpreting Technical Metadata as Archival Narrative». *IPres Conference 2025 Proceedings, 21st International Conference on Digital Preservation* (Wellington, New Zealand (Te Whanganui-a-Tara, Aotearoa)), novembre.

Grandjean, Martin. 2016. «Archives Distant Reading: Mapping the Activity of the League of Nations' Intellectual Cooperation». *Digital Humanities 2016* (Krakow, Poland), 531–34. https://shs.hal.science/halshs-01525565.

Grigar, Dene, e James Christopher O'Sullivan, a c. di. 2021. *Electronic Literature as Digital Humanities: Contexts, Forms, and Practices*. Bloomsbury Academic.

Grillo, Remo, Lucia Giagnolini, e Paolo Bonora. 2026. «Archival Knowledge Graphs: Balancing Semantic Richness and Accessibility of Hierarchical Structures». *Proceedings of the 22nd Conference on Information and Research Science Connecting to Digital and Library Science (IRCLD 2026)* (Modena, Italy).

Gubitosa, Carlo. 1999. *Italian crackdown : BBS amatoriali, volontari telematici, censure e sequestri nell'Italia degli anni 90 / Carlo Gubitosa . Associazione peacelink*. In *Italian crackdown BBS amatoriali, volontari telematici, censure e sequestri nell'Italia degli anni 90*. Apogeo.

Guernaccini, Fabiana, Silvia Mazzini, e Giovanni Bruno. 2019. «LOD Publication in the Archival Domain: Methods and Practices». In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH 2019)*, a cura di Antonella Poggi. Sapienza University of Rome, Faculty of Arts and Humanities.

Guo, Willis, Armin Toroghi, e Scott Sanner. 2024. «CR-LT-KGQA: A Knowledge Graph Question Answering Dataset Requiring Commonsense Reasoning and Long-Tail Knowledge». arXiv:2403.01395. Preprint, arXiv, marzo 3. https://doi.org/10.48550/arXiv.2403.01395.

Haraway, Donna. 1988. «Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective». *Feminist Studies* (College Park, Md) 14 (3): 575–99.

Harvey, Phil. 2003. «ExifTool: Platform-independent Perl library and command-line application for reading, writing and editing meta information in files». https://exiftool.org/.

Hawkins, Ashleigh. 2022. «Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web». *Archival Science* 22 (3): 319–44. https://doi.org/10.1007/s10502-021-09381-0.

Hedstrom, Margaret. 1993. «Descriptive Practices for Electronic Records: Deciding What Is Essential and Imagining What Is Possible». *Archivaria* 36 (febbraio): 53–63.

Henttonen, Pekka, e Jaana Kilkki. 2017. «"Records in Contexts" and the Finnish Conceptual Model for Archival Description». *Letonica* 36: 60–71.

Higgins, Sarah, Christopher Hilton, e Lyn Dafis. 2014. «Archives context and discovery: Rethinking arrangement and description for the digital age». *2nd Annual Conference of the International Council*

on Archives: Archives and Cultural Industries (Girona, Spain), ottobre. http://www.girona.cat/web/ica2014/eng/comunicacions.php#GJ.

Huang, Yurui, Hengwang Li, e Tianrui Liu. 2025. «The Evolution of Bulletin Board System Forum Applications in the Digital Perspective: Mechanisms, Trends, and Implications». *Proceedings of the 2024 International Conference on Digital Economy and Computer Science* (New York, NY, USA), DECS '24, 227–32. https://doi.org/10.1145/3705618.3705656.

Hudson, Graham, Alain Léger, Birger Niss, e István Sebestyén. 2017. «JPEG at 25: Still Going Strong». *IEEE MultiMedia* 24 (2): 96–103. https://doi.org/10.1109/MMUL.2017.38.

Hyvönen, Eero, Mikko Koho, e Petri Leskinen. 2021. «WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data». *Archival Science* 21 (2): 215–34.

Iadevaia, Roberta. 2022. «Letteratura&rete. La parola agli scrittori elettronici». *ENTHYMEMA*, fasc. 30: 30. https://doi.org/10.54103/2037-2426/19556.

ICA-EGAD (International Council on Archives - Expert Group on Archival Description). 2024. «Discussion on refining the definition of an attribute for a legal status». https://github.com/ICA-EGAD/RiC-O/issues/56.

IEEE/Open Group Std. 2024. «IEEE/Open Group Standard for Information Technology–Portable Operating System Interface (POSIX™) Base Specifications, Issue 8». *IEEE/Open Group Std 1003.1-2024 (Revision of IEEE Std 1003.1-2017)*, 1–4107. https://doi.org/10.1109/IEEESTD.2024.10555529.

IFLA LRMoo Working Group e CIDOC CRM SIG. 2024. *LRMoo Object-Oriented Definition and Mapping from the IFLA Library Reference Model*. https://www.cidoc-crm.org/frbroo/sites/default/files/LRMoo_V1.0.pdf.

IFLA LRMOO Working Group, CIDOC CRM Special Interest Group, Chryssoula Bekiari, Martin Doerr, Patrick Le Bœuf, e Pat Riva. 2024. *LRMOO Object-Oriented Definition and Mapping from the IFLA Library Reference Model*. Technical Specification. Versione 1.0. International Federation of Library Associations and Institutions (IFLA). https://www.ifla.org/files/assets/cataloguing/lrmoo/lrmoo_v1.0.pdf.

International Council on Archives. 2000. *ISAD(G): General International Standard Archival Description*. Second. International Council on Archives.

International Council on Archives, Expert Group on Archival Description. 2023. *Records in Contexts Conceptual Model*. Versione 1.0. International Council on Archives. https://www.ica.org/en/records-in-contexts-conceptual-model.

International Council on Archives Expert Group on Archival Description. 2023. *Records in Contexts: Foundations of Archival Description*. https://github.com/ICA-EGAD/RiC-CM/issues.

International Council on Archives, Expert Group on Archival Description. 2025a. «Next Steps». About ICA Records in Contexts Ontology (RiC-O), maggio. https://ica-egad.github.io/RiC-O/next-steps.html.

International Council on Archives, Expert Group on Archival Description. 2025b. «Why Use RiC-O?» About ICA Records in Contexts Ontology (RiC-O), maggio 22. https://ica-egad.github.io/RiC-O/why-use-RiC-O.html.

International Council on Archives, Expert Group on Archival Description. s.d. «Records in Contexts (RiC)». *International Council on Archives*. https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/.

International Organization for Standardization. 1989. *Information processing systems – Open Systems Interconnection – Basic Reference Model – Part 2: Security Architecture*. International Standard ISO 7498-2:1989. ISO. https://www.iso.org/standard/14256.html.

InterPARES 1 Authenticity Task Force. 2002. *Requirements for Assessing and Maintaining the Authenticity of Electronic Records*. InterPARES Project. https://www.interpares.org/book/interpares_book_k_app02.pdf.

InterPARES Trust. 2016. «InterPARES Trust response to EGAD-RiC». https://interparestrust.org/trust/article/interpares-trust-response-to-egadric.

Istituto Centrale per gli Archivi - ICAR. 2025. «Arco4Archives». giugno 27. https://icar.cultura.gov.it/standard/standard-san/arco4archives.com.

Italia, Paola, a c. di. 2021. *A carte scoperte : come lavorano le scrittrici e gli scrittori contemporanei : dieci domande a Andrea Bajani... [et al.]*. Bononia University Press.

Italia, Paola, e Monica Zanardo. 2023a. *Volontà d'archivio: l'autore, le carte, l'opera*. I libri di Viella 469. Viella.

Italia, Paola, e Monica Zanardo, a c. di. 2023b. *Volontà d'archivio: l'autore, le carte, l'opera*. I libri di Viella 469. Viella.

Jaillant, Lise. 2019. «After the digital revolution: working with emails and born-digital records in literary and publishers' archives». *Archives and Manuscripts* 47 (3): 285–304. https://doi.org/10.1080/01576895.2019.1640555.

Jaillant, Lise. 2022a. «How Can We Make Born-Digital and Digitised Archives More Accessible? Identifying Obstacles and Solutions». *Archival Science* 22 (3): 417–36. https://doi.org/10.1007/s10502-022-09390-7.

Jaillant, Lise. 2022b. «More Data, Less Process: A User-Centered Approach to Email and Born-Digital Archives». *The American Archivist* 85 (2): 533–55. https://doi.org/10.17723/2327-9702-85.2.533.

Jaillant, Lise. 2024. «Introduction to the Special Issue: Using Visual AI Applied to Digital Archives». *Digital Humanities Quarterly* 018 (3). https://www.digitalhumanities.org/dhq/vol/18/2/000752/000752.html.

Jaillant, Lise, Katie Aske, Eirini Goudarouli, e Natasha Kitcher. 2022. «Introduction: Challenges and Prospects of Born-Digital and Digitized Archives in the Digital Humanities». *Archival Science* 22 (3): 285–91. https://doi.org/10.1007/s10502-022-09396-1.

Jaillant, Lise, e Annalina Caputo. 2022. «Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data». *AI & SOCIETY* 37 (3): 823–35. https://doi.org/10.1007/s00146-021-01367-x.

Jaillant, Lise, e Lingjia Zhao. 2025. «Introduction: When data turns into archives: making digital records more accessible with AI». *AI & SOCIETY*, pubblicazione online ad accesso anticipato, aprile 30. https://doi.org/10.1007/s00146-025-02374-y.

James, Ryan, e Leon De Kock. 2013. «The Digital David and the Gutenberg Goliath: The Rise of the "Enhanced" E-Book». *English Academy Review: A Journal of English Studies* 30 (1): 107–23. 202017112809. https://doi.org/10.1080/10131752.2013.783394.

Jockers, Matthew. 2023. *Introduction to the Syuzhet Package*. CRAN / R Project. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html.

«John Updike Papers, 1940–2009 (MS Am 1793)». s.d.

Jordan, Tim, e Paul Taylor. 2004. *Hacktivism and Cyberwars: Rebels with a Cause?* 1ª ed. Routledge. https://doi.org/10.4324/9780203490037.

Karagiannopoulos, Vasileios. 2021. «A Short History of Hacktivism: Its Past and Present and What Can We Learn from It». In *Rethinking Cybercrime: Critical Debates*, a cura di Tim Owen e Jessica Marshall. Springer International Publishing. https://doi.org/10.1007/978-3-030-55841-3_4.

Kartography CIC. 2023. «The ResearchSpace System - Underlying Principles». https://kartography.org/researchspace.html.

Kerrisk, Michael. 2010. *The Linux Programming Interface: A Linux and UNIX System Programming Handbook*. No Starch Press.

Ketelaar, Eric. 2023. «Archival contexts». *Archeion* CXXIV: 35–56. https://doi.org/10.4467/26581264ARC.23.003.17863.

Kilbride, William. 2024. «DP and Artificial Intelligence - A Four Point Plan - Digital Preservation Coalition». *Digital Preservation Coalition*, giugno 24. https://www.dpconline.org/blog/dp-and-artificial-intelligence-a-4-point-plan.

Kim, Haklae. 2023. «A knowledge graph of interlinking digital records: the case of the 1997 Korean financial crisis». *The Electronic Library* 42 (1): 60–77. https://doi.org/10.1108/EL-05-2023-0131.

King, Owen C. 2024. «Archival Meta-Metadata: Revision History and Positionality of Finding Aids». *Archival Science* 24 (3): 509–29. https://doi.org/10.1007/s10502-024-09443-z.

Kirschenbaum, Matthew G. 2007. *Mechanisms: New Media and the Forensic Imagination*. The MIT Press. https://doi.org/10.7551/mitpress/7393.001.0001.

Kirschenbaum, Matthew G. 2013. «The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary». *Digital Humanities Quarterly* 7. https://api.semanticscholar.org/CorpusID:31174925.

Koch, Inês Dias. 2025. «Integration of models for linked data in cultural heritage and contributions to the FAIR principles». PhD thesis, Faculdade de Engenharia da Universidade do Porto. https://hdl.handle.net/10216/167821.

Koch, Inês, Nuno Freitas, Cristina Ribeiro, Carla Teixeira Lopes, e João Rocha Da Silva. 2019. «Knowledge Graph Implementation of Archival Descriptions Through CIDOC-CRM». In *Digital Libraries for Open Knowledge*, a cura di Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond

Aalberg, e Adam Jatowt, vol. 11799. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-30760-8_8.

Koch, Inês, Cristina Ribeiro, e Carla Teixeira Lopes. 2020. «ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives». In *Digital Libraries for Open Knowledge*, a cura di Mark Hall, Tanja Merčun, Thomas Risse, e Fabien Duchateau, vol. 12246. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-54956-5_10.

Koch, Inês, Carla Teixeira Lopes, e Cristina Ribeiro. 2023. «Moving from ISAD(G) to a CIDOC CRM-Based Linked Data Model in the Portuguese Archives». *Journal on Computing and Cultural Heritage* 16 (4): 1–21. https://doi.org/10.1145/3605910.

Krautter, Benjamin. 2024. «The Scales of (Computational) Literary Studies: Martin Mueller's Concept of Scalable Reading in Theory and Practice». In *Exploring Scale in Digital History and Humanities*, a cura di Florentina Armaselu e Andreas Fickers. De Gruyter Oldenbourg. https://doi.org/doi:10.1515/9783111317779-011.

Krtalić, Maja, e Jesse David Dinneen. 2024. «Information in the Personal Collections of Writers and Artists: Practices, Challenges and Preservation». *Journal of Information Science* 50 (1): 189–203. https://doi.org/10.1177/01655515221084613.

Kulmukhametov, Artur, Andreas Rauber, e Christoph Becker. 2021. «Improving data quality in large-scale repositories through conflict resolution». *International Journal on Digital Libraries* 22 (4): 365–83. https://doi.org/10.1007/s00799-021-00311-0.

Kuny, Terry. 1997. «A Digital Dark Ages? Challenges in the Preservation of Electronic Information». *63rd IFLA Council and General Conference, Workshop: Audiovisual and Multimedia joint with Preservation and Conservation, Information Technology, Library Buildings and Equipment, and the PAC Core Programme* (Copenhagen, Denmark), settembre. mailto:terry.kuny@xist.com.

Langdon, John. 2016. «Describing the Digital: The Archival Cataloguing of Born-Digital Personal Papers». *Archives and Records* 37 (1): 37–52. https://doi.org/10.1080/23257962.2016.1139494.

Larry, Daniel, e Daniel Lars. 2011. *Digital Forensics for Legal Professionals: Understanding Digital Evidence From the Warrant to the Courtroom*. Syngress, Burlington. https://www.ebsco.com/ebooks.

Lee, Christopher. 2012. «Archival application of digital forensics methods for authenticity, description and access provision». *Comma* 2012 (2): 133–40.

Lee, Christopher A., Kam Woods, Matthew Kirschenbaum, e Alexandra Chassanoff. 2013. *From Bitstreams to Heritage: Putting Digital Forensics into Practice*.

Leighton John, Jeremy. 2009. *Digital Lives: An Initial Synthesis*. British Library. https://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf.

Levy, Josh. 2022. «An Introduction to Born Digital Collections at the Manuscript Division, or How to Cross the Equator». Library of Congress. https://blogs.loc.gov/manuscripts/2022/01/an-introduction-to-born-digital-collections-at-the-manuscript-division-or-how-to-cross-the-equator/.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. 2020. «Retrieval-augmented generation for knowledge-intensive NLP tasks». *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA), NIPS '20.

Libraries, University of California Systemwide. 2017. *UC Guidelines for Born-Digital Archival Description * ottobre 26. https://escholarship.org/uc/item/9cg222jc.

Library of Congress. 2023. «Exif Exchangeable Image File Format, Version 2.2». Web page. novembre 3. https://www.loc.gov/preservation/digital/formats/fdd/fdd000146.shtml.

Library of Congress. 2025. *Recommended Formats Statement 2025-2026*.

Light, M. 2014. «Managing Risk with a Virtual Reading Room: Two Born-Digital Projects». In *Reference and Access: Innovative Practices for Archives and Special Collections*, a cura di K. Theimer. Rowman & Littlefield Publishers.

Lindlar, Micky. 2022. «PREMIS For All, For Good, Forever!» *Open Preservation Foundation*, settembre 21. https://openpreservation.org/blogs/premis-for-all-for-good-forever/.

Lodolini, Elio. 1981. «L'ordinamento dell'archivio: nuove discussioni». *Rassegna degli Archivi di Stato* 41 (1–3): 38–56.

Lowood, Henry E. 2023. *Essential Writings on Software Preservation and Game Histories*. Johns Hopkins University Press.

MacNeil, Heather. 2005. «Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor». *The American Archivist* 68 (2): 264–78. JSTOR.

Mameli Andrea. 2012. «Nicolas Eymerich, il primo videogame in latino (completamente accessibile) a Bologna il 16 Novembre». *Nicolas Eymerich, il primo videogame in latino (completamente*

accessibile) a Bologna il 16 Novembre, ottobre 27. https://linguaggio-macchina.blogspot.com/2012/10/nicolas-eymerich-il-primo-videogame-in.html.

Marilù Oliva. 2013. «VALERIO EVANGELISTI». *libroguerriero*, dicembre 2. https://libroguerriero.wordpress.com/2013/12/02/valerio-evangelisti/.

Matt, Luigi. 2009. «Chi è stregato dallo Strega? Rilievi di stile sugli ultimi romanzi vincitori (2002-2009)». *Lingua Italiana d'oggi* (Roma), 251–86.

Maugeri, Massimo. 2011. *L'e-book e (è?) il futuro del libro: opinioni emotive sul libro elettronico da parte degli addetti ai lavori del mondo della scrittura, dell'editoria e della critica letteraria*. Tascabili. Historica.

McKean, Callum. 2025. «Personal Digital Archives at the British Library: Capture, Description, Access». In *The Intangible Papers. Authorial Philology and Born-Digital Texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino. https://doi.org/10.1401/9788815414328/c5.

Meyerson, Jessica, Zac Vowell, Wendy Hagenmaier, et al. 2017. «The Software Preservation Network (SPN): A Community Effort to Ensure Long Term Access to Digital Cultural Heritage». *D-Lib Magazine* 23 (5/6). https://doi.org/10.1045/may2017-meyerson.

Michela Trigari. 2012. «Ad Handimatica il primo videogioco per non vedenti - Corriere.it». *Corriere della sera*, novembre 20. https://web.archive.org/web/20130123074223/http://www.corriere.it/salute/disabilita/12_novembre_20/handimatica-videogioco-ciechi_e1c8cda8-3255-11e2-942f-a1cc3910a89d.shtml.

Michetti, Giovanni. 2009. «Ma è poi tanto pacifico che l'albero rispecchi l'archivio?» *Archivi & Computer* 1: 85–95.

Microsoft. 2024. «How to customize folders with Desktop.ini». Microsoft Learn. https://learn.microsoft.com/en-us/windows/win32/shell/how-to-customize-folders-with-desktop-ini.

Microsoft Corporation. 2021. «File Times». https://learn.microsoft.com/en-us/windows/win32/sysinfo/file-times.

Microsoft Developer Documentation. 2024a. «File Locking». https://docs.microsoft.com/en-us/windows/win32/fileio/locking-and-unlocking-byte-ranges-in-files.

Microsoft Developer Documentation. 2024b. «Globally Unique Identifiers». https://docs.microsoft.com/en-us/windows/win32/api/guiddef/ns-guiddef-guid.

Microsoft Italia. 2014. «Microsoft lancia OneDrive a livello mondiale il servizio di archiviazione». gennaio. https://news.microsoft.com/it-it/2014/01/19/microsoft-lancia-onedrive-a-livello-mondiale-il-servizio-di-archiviazione/.

Micunovic, Milijana, Hana Marčetić, e Maja Krtalić. 2016. «Literature and Writers in the Digital Age: A Small-Scale Survey of Contemporary Croatian Writers' Organization and Preservation Practices». Preservation, Digital Technology & Culture (PDT&C). *Preservation, Digital Technology & Culture* 45 (1): 2–16. https://doi.org/10.1515/pdtc-2015-0028.

Mikhaylova, Daria, e Daniele Metilli. 2023. «Extending RiC-O to Model Historical Architectural Archives: The ITDT Ontology». *J. Comput. Cult. Herit.* (New York, NY, USA) 16 (4). https://doi.org/10.1145/3606706.

Milone, Federico. 2025. «Philological Investigation in the Digital Era: The Case of Laura Pugno's Sirene». In *The Intangible Papers. Authorial Philology and Born-Digital Texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino.

Minarso, Christinia, Tamara Adriani Salim, Rahmi, e Mad Khir Johari Abdullah Sani. 2023. «Strategies and Challenges of Personal Digital Archiving (PDA) in the Digital Era». *Proceedings of the fourth Asia-Pacific Research in Social Sciences and Humanities, Arts and Humanities Stream (AHS-APRISH 2019)*, 457–71. https://doi.org/10.2991/978-2-38476-058-9_36.

Modern Language Association. 1995. *Statement on the Significance of Primary Records*. Statement. Modern Language Association. https://www.mla.org/Resources/Guidelines-and-Data/Reports-and-Professional-Guidelines/Significance-of-Primary-Records/Read-the-Report-Online/Statement-on-the-Significance-of-Primary-Records-Modern-Language-Association.

Mohammed, Farah. 2017. «The Rise and Fall of the Blog». *JSTOR Daily*, dicembre 27. https://daily.jstor.org/the-rise-and-fall-of-the-blog/.

Monforte, Alessandro d'Arminio. 2025. «The Digital Legacy in Italy». In *The Intangible Papers: Authorial Philology and Born-Digital Texts*, a cura di Giuseppe Antonelli, Lucia Giagnolini, e Federico Milone. Il Mulino. https://doi.org/10.1401/9788815414328/c2.

Moraitou, Efthymia, John Aliprantis, Yannis Christodoulou, Alexandros Teneketzis, e George Caridakis. 2019. «Semantic Bridging of Cultural Heritage Disciplines and Tasks». *Heritage* 2 (1): 611–30. https://doi.org/10.3390/heritage2010040.

Moretti, Franco. 2000. «Conjectures on World Literature». *New Left Review* 1 (febbraio): 54–68.

Moretti, Franco. 2013. *Distant Reading*. Verso.

Nesmith, Tom. 2002. «Seeing archives: postmodernism and the changing intellectual place of archives». *The American Archivist*, 24–41.

Nesmith, Tom. 2005. «Reopening Archives: Bringing New Contextualities into Archival Theory and Practice». *Archivaria*, 259–74.

Nesmith, Tom. 2006. «The concept of societal provenance and records of nineteenth-century Aboriginal–European relations in Western Canada: implications for archival theory and practice». *Archival Science* 6 (3): 351–60. https://doi.org/10.1007/s10502-007-9043-9.

Niu, Jinfang. 2015. «Original order in the digital world». *Archives and Manuscripts* 43 (1): 61–72. https://doi.org/10.1080/01576895.2014.958863.

Note, Margot. 2025. «The Ethical Use of Born-Digital Materials in Archives». Lucidea. Think Clearly Blog, settembre 8. https://lucidea.com/blog/ethical-born-digital-archives/.

Oh, Junghoon, Sangjin Lee, e Hyunuk Hwang. 2022. «Forensic Recovery of File System Metadata for Digital Forensic Investigation». *IEEE Access* 10: 111591–606.

Oldman, Dominic, e Diana Tanase. 2018. «Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace». In *The Semantic Web – ISWC 2018*, a cura di Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, et al. Springer International Publishing.

Oliveira, Caliel Cardoso de, Marieta Marks Löw, e Thiago Henrique Bragato Barros. 2024. «Knowledge Organization Possibilities for Archives: Comparative Semantic Analysis Between CIDOC-CRM and RiC-CM». *KO KNOWLEDGE ORGANIZATION* 51 (5): 362–70. https://doi.org/10.5771/0943-7444-2024-5-362.

Open Knowledge. s.d. «Glossary, Triple store». Open Data Handbook. https://opendatahandbook.org/glossary/it/terms/triple-store/.

Pacheco, André, Carlos Guardado Da Silva, e Maria Cristina Vieira De Freitas. 2023. «A metadata model for authenticity in digital archival descriptions». *Archival Science* 23 (4): 629–73. https://doi.org/10.1007/s10502-023-09422-w.

Palmero Aprosio, Giovanni, e Giovanni Moretti. 2019. «Tint 2.0: An All-Inclusive Suite for NLP in Italian». In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It 2018*, a cura di Elena Cabrio, Alessandro Mazzei, e Fabio Tamburini. Accademia University Press.

Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, e Xindong Wu. 2024. «Unifying Large Language Models and Knowledge Graphs: A Roadmap». *IEEE Transactions on Knowledge and Data Engineering* 36 (7): 3580–99. https://doi.org/10.1109/tkde.2024.3352100.

Park, EunKyung, e Sanghee Oh. 2018. «A comparative study of personal digital archiving practices between American and Korean researchers». *Journal of Information Science* 44 (6): 767–81. https://doi.org/10.1177/0165551517741393.

Parmar, Sharika. 2025. «Forging an Interdisciplinary Lens for Understanding Community Digital Archives of South Asian Diaspora». *Frontiers in Sociology* 10 (giugno). https://doi.org/10.3389/fsoc.2025.1450641.

Pashkeeva, Natalia, Chowra Makaremi, e Johanna Saadoune. 2024. «De l'ethnographie critique des archives à la modélisation d'une base de données pour l'étude de l'Iran postrévolutionnaire». *Revue des mondes musulmans et de la Méditerranée* 156 (2). https://doi.org/10.4000/remmm.21709.

Paul Gabriele, Weston, Emmanuela Carbé, e Primo Baldini. 2017. «If Bits Are Not Enough: Preservation Practices of the Original Contest for Born Digital Literary Archives». *Bibliothecae.It* Vol 6 (giugno): 154-177 Pages. 154-177 Pages. https://doi.org/10.6092/ISSN.2283-9364/7027.

Pendergrass, Keith L., Walker Sampson, Tim Walsh, e Laura Alagna. 2019. «Toward Environmentally Sustainable Digital Preservation». *The American Archivist* 82 (1): 165–206. https://doi.org/10.17723/0360-9081-82.1.165.

Penzo Doria, Gianni. 2022. «A new archives definition». *JLIS.it* 13 (2). https://doi.org/10.36253/jlis.it-465.

Piazza, Isotta. 2020. «Dal web al volume: il caso Carbé». *Prassi Ecdotiche della Modernità Letteraria*, fasc. 5/I: 5/I. https://doi.org/10.13130/2499-6637/13164.

Pires, Catarina, Inês Koch, e Sergio Nunes. 2023. «ArchOnto Ontology Representation of Portuguese Archival Description Units (Baptism Records and Passports)». https://doi.org/10.25747/X78E-1A27.

Pledge, Jonathan, e Eleanor Dickens. 2018. «Process and Progress: Working with Born-Digital Material in the Wendy Cope Archive at the British Library». *Archives and Manuscripts* 46 (1): 59–69. https://doi.org/10.1080/01576895.2017.1408024.

Polley, Katherine Louise, Vivian Teresa Tompkins, Brendan John Honick, e Jian Qin. 2021. «Named Entity Disambiguation for Archival Collections: Metadata, Wikidata, and Linked Data». *Proceedings*

of the Association for Information Science and Technology 58 (1): 520–24. https://doi.org/10.1002/pra2.490.

Poveda-Villalón, María, Asunción Gómez-Pérez, e Mari Carmen Suárez-Figueroa. 2014. «OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation». *International Journal on Semantic Web and Information Systems (IJSWIS)* 10 (2): 7–34.

Pozzoli, C. 1986. *Scrivere con il computer. Introduzione alla videoscrittura con personal computer e word processor*. Mondadori.

Pratesi, Alessandro. 2018. *Genesi e forme del documento medievale*. 1ª ed. Vol. 28. Historica. Jouvence.

Prayudi, Yudi, e Azhari Sn. 2015. «Digital chain of custody: State of the art». *International Journal of Computer Applications* 114 (5).

PREMIS Editorial Committee. 2015a. «PREMIS Data Dictionary for Preservation Metadata, Version 3.0». https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

PREMIS Editorial Committee. 2015b. *PREMIS Data Dictionary for Preservation Metadata, Version 3.0*. Library of Congress. http://www.loc.gov/standards/premis/v3/index.html.

Prom, Christopher J., e Thomas J. Frusciano, a c. di. 2013. *Archival Arrangement and Description*. Trends in Archives Practice. Society of American Archivists.

Ragusa, Carmen. 2021. «Il computer per scrivere: per una storia della videoscrittura in Italia tra il 1983 e i primi anni Novanta». Tesi di laurea magistrale, Università di Pisa.

Rajh, Arian. 2024. «Archival Description Turns Truly Collaborative: An Exercise in Records in Contexts Standard». *Moderna Arhivistika* 2024 (7) (1): 63–82. https://doi.org/10.54356/MA/2024/JVGE4567.

Rana, Anshul. 2024. *Understanding the DRAM: How Does Computer Memory Work?* Articles. settembre 16. https://storedbits.com/dram-working/.

Regesta.exe, redazione. 2013. *» Linked open data archivistici*. ottobre 14. https://www.regesta.com/2013/10/14/linked-open-data-archivistici/.

Ries, Thorsten. 2018. «The rationale of the born-digital dossier génétique: Digital forensics and the writing process: With examples from the Thomas Kling Archive». *Digital Scholarship in the Humanities* 33 (2): 391–424. https://doi.org/10.1093/llc/fqx049.

Ries, Thorsten. 2022. «Digital history and born-digital archives: the importance of forensic methods». *Journal of the British Academy* 10: 157–85. https://doi.org/10.5871/jba/010.157.

Ries, Thorsten. 2025. *Textgenese und digitale Forensik: Exemplarische Studien zu Thomas Kling und Michael Speier*. Vol. 120. Beihefte zum Euphorion. Zeitschrift für Literaturgeschichte. Winter.

Ries, Thorsten, e Gábor Palkó. 2019. «Born-Digital Archives». *International Journal of Digital Humanities* 1 (1): 1–11. https://doi.org/10.1007/s42803-019-00011-x.

Riva, Pat, M. Žumer, e T. Aalberg. 2023. «LRMoo, Navigating Standards Development Processes in Two Communities». *World Library and Information Congress, 88th IFLA General Conference and Assembly* (Rotterdam, The Netherlands). https://repository.ifla.org/handle/123456789/2668.

Rogers, Corinne. 2019. «From time theft to time stamps: mapping the development of digital forensics from law enforcement to archival authority». *International Journal of Digital Humanities* 1 (1): 13–28. https://doi.org/10.1007/s42803-019-00002-y.

Romagna, Marco. 2020. «Hacktivism: Conceptualization, Techniques, and Historical View». In *The Palgrave Handbook of International Cybercrime and Cyberdeviance*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-78440-3_34.

Rupp, Florian, Benjamin Schnabel, e Kai Eckert. 2022. «Easy and Complex: New Perspectives for Metadata Modeling Using RDF-Star and Named Graphs». In *Knowledge Graphs and Semantic Web*, a cura di Boris Villazón-Terrazas, Fernando Ortiz-Rodriguez, Sanju Tiwari, Miguel-Angel Sicilia, e David Martín-Moncunill. Springer International Publishing. https://doi.org/10.1007/978-3-031-21422-6_18.

San Emeterio de la Parte, Mario, José-Fernán Martínez-Ortega, Néstor Lucas Martínez, e Vicente Hernández Díaz. 2025. «SISS: Semantic Interoperability Support System for the Internet of Things». *IEEE Internet of Things Journal* 12 (16): 33769–91. https://doi.org/10.1109/JIOT.2025.3577776.

Santos, Catarina, e Jorge Revez. 2023. «Applying Records in Contexts in Portugal: the case of the scientific correspondence from António de Barros Machado and Dora Lustig archive». *Archival Science* 23 (2): 137–58. https://doi.org/10.1007/s10502-022-09401-7.

Sanzogni, Carlotta, a c. di. 2023. *Dove si scrive, come si scrive*. Argentovivo. Rizzoli.

Savoy, Jacques. 2020. «Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling». *Machine Learning Methods for Stylometry*. https://api.semanticscholar.org/CorpusID:221954648.

Schaefer, Sibyl, e Janet M. Bunde. 2013. «Standards for Archival Description». In *Archival Arrangement and Description*, a cura di Christopher J. Prom e Thomas J. Frusciano. Society of American Archivists.

Scott, Paul J. 1966. «The record group concept: a case for abandonment». *American Archivist* 29 (4): 493–504.

Sebastiani, Alberto. 2018. *Nicolas Eymerich : il lettore e l'immaginario in Valerio Evangelisti*. Odoya library 332. Odoya.

Secondulfo, Giovanni. 2000. «Eymerich Mailing List». In *Il mondo dei Fan Club*, a cura di Fulvio Paloscia e Luca Scarlini. Prima scelta 21. Adnkronos libri.

Sgherza, Alessio. 2016. «Giorgio Rutigliano, papà della prima Bbs italiana: "I nostri gruppi antenati di Facebook"». la Repubblica, aprile 29. http://www.repubblica.it/tecnologia/2016/04/29/news/giorgio_rutigliano_papa_della_prima_bbs_italiana_i_nostri_gruppi_antenati_di_facebook_-138691361/.

Shiri, Fatemeh, Van Nguyen, Farhad Moghimifar, John Yoo, Gholamreza Haffari, e Yuan-Fang Li. 2024. «Decompose, Enrich, and Extract! Schema-Aware Event Extraction Using LLMs». https://arxiv.org/abs/2406.01045.

Shu, Junliang, Yuanyuan Zhang, Juanru Li, Bodong Li, e Dawu Gu. 2017. «Why Data Deletion Fails? A Study on Deletion Flaws and Data Remanence in Android Systems». *ACM Trans. Embed. Comput. Syst.* (New York, NY, USA) 16 (2). https://doi.org/10.1145/3007211.

Sikos, Leslie F., e Dean Philp. 2020. «Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs». *Data Science and Engineering* 5 (3): 293–316. https://doi.org/10.1007/s41019-020-00118-0.

Simon, Bart. 2007. «Geek Chic: Machine Aesthetics, Digital Gaming, and the Cultural Politics of the Case Mod». *Games and Culture* 2 (3): 175–93. https://doi.org/10.1177/1555412007304423.

Simonetti, Gianluigi. 2023. *Caccia allo Strega : anatomia di un premio letterario*. Extrema ratio. Nottetempo <casa editrice>.

Sinn, Donghee, Sujin Kim, e Sue Yeon Syn. 2017. «Personal digital archiving: influencing factors and challenges to practices». *Library Hi Tech* 35 (2): 222–39. https://doi.org/10.1108/LHT-09-2016-0103.

Sistema Archivistico Nazionale. 2014. «Pubblicate le versioni 1.0 dell'Ontologia SAN LOD e del Thesaurus in formato SKOS del SAN». luglio 16. http://www.san.beniculturali.it/web/san/dettaglio-notizia-san?p_p_id=56_INSTANCE_X7Qi&articleId=2925424&p_p_lifecycle=1&p_p_state=normal&groupId=10704&viewMode=normal.

Skorobogatov, Sergei. 2005. «Data Remanence in Flash Memory Devices». In *Cryptographic Hardware and Embedded Systems – CHES 2005*, a cura di Josyula R. Rao e Berk Sunar. Springer Berlin Heidelberg.

Smith, Lisa, Jenny Wood, Greg Oakes, e Madalyn Grant. 2021. *Exploring Ethical Considerations for Providing Access to Digital Heritage Collections*. Digital Preservation Coalition. https://doi.org/10.7207/twgn21-18.

Sosio, Silvio. 1996. «Il creatore di Nicholas Eymerich: intervista con Valerio Evangelisti ∂ Fantascienza.com». Fantascienza.com, luglio 15. https://www.fantascienza.com/175/il-creatore-di-nicholas-eymerich-intervista-con-valerio-evangelisti.

Sosio, Silvio. 2004. «Dieci anni di fantascienza ∂ Fantascienza.com». Fantascienza.com, novembre 20. https://www.fantascienza.com/6788/dieci-anni-di-fantascienza.

Spadini, Elena. 2025. «Grant Search: Bit Philology». https://data.snf.ch/grants/grant/226460.

Stollar Peters, C. 2006. «When Not All Papers are Paper: A Case Study in Digital Archivy». *Provenance. Journal of the Society of Georgia Archivists* 24 (1): 22–34.

Sundqvist, Anneli. 2021. «Things That Work: Meditations on Materiality in Archival Discourse». *Journal of Contemporary Archival Studies* 8 (1): Article 7.

The National Archives. 2016. *Redaction Toolkit: Editing Exempt Information from Paper and Electronic Documents Prior to Release*. The National Archives. https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

The National Archives. 2021. «The National Archives Implement the ResearchSpace Platform». *ResearchSpace*, novembre 8. https://researchspace.org/blog/national_archives_researchspace/.

The Sedona Conference. 2013. «Commentary on Ethics & Metadata». *Sedona Conference Journal* 14: 169.

Thibodeau, Kenneth. 2002. «Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years». *The State of Digital Preservation: An International Perspective* (Washington, DC), CLIR Publication, fasc. 107.

Timms, Kat. 2013. «The Devil is in the Details: Describing Born-digital Records Using the Rules for Archival Description». *Society of American Archivists Annual Meeting* (New Orleans), agosto. http://files.archivists.org/conference/nola2013/materials/701-TimmsA.pdf.

Tomasi, Francesca. 2017. «La preservazione del contenuto degli oggetti culturali: formalizzare la provenance». *Bibliothecae.it* 6 (2): 2. https://doi.org/10.6092/issn.2283-9364/7531.

Tomasi, Francesca. 2022. *Organizzare la conoscenza: Digital Humanities e Web semantico*. Editrice Bibliografica. https://doi.org/10.53134/9788893573573.

Tomasi, Francesca. 2023. «Archival Finding Aids in Linked Open Data between Description and Interpretation». *JLIS.It* 14 (3): 3. https://doi.org/10.36253/jlis.it-557.

Tomasi, Francesca, e Marilena Daquino. 2015. «Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare». *JLIS.it* 6: 13–38.

Toroghi, Armin, Willis Guo, Mohammad Mahdi Abdollah Pour, e Scott Sanner. 2025. «Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering». https://arxiv.org/abs/2403.01390.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. 2023. «LLaMA: Open and Efficient Foundation Language Models». *ArXiv* abs/2302.13971. https://api.semanticscholar.org/CorpusID:257219404.

Trace, Ciaran B. 2011. «Beyond the Magic to the Mechanism: Computers, Materiality, and What It Means for Records to Be "Born Digital"». *Archivaria*, dicembre 2, 5–27.

Treccani Vocabolario Online. s.d. «Netiquette - Significato ed etimologia». Treccani. Consultato 2 settembre 2025. https://www.treccani.it/vocabolario/netiquette/.

Turkle, Sherry. 1995. *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster Trade.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. The University of Chicago Press.

Università di Bologna. 1976. «Facoltà di Scienze Politiche, fascicolo studente Valerio Evangelisti». N. 2901. Archivio Storico dell'Università di Bologna.

University at Buffalo Libraries. s.d. «Research Guides: Archival Processing Documentation: Born Digital». Consultato 11 settembre 2025. https://research.lib.buffalo.edu/processing/digital-materials.

University of Washington, Paul G. Allen School of Computer Science & Engineering. 2019. «Bulletin Board Systems».

US National ArchivesArchives. 2016. «Metadata in Electronic Records Management». *Records Express*, novembre 21. https://records-express.blogs.archives.gov/2016/11/21/metadata-in-electronic-records-management/.

Valacchi, Federico. 2022. «The Parts and the Whole. Integrate Knowledge». *JLIS.It* 13 (3): 3. https://doi.org/10.36253/jlis.it-477.

Valacchi, Federico. 2023. «Not the Institutions but the Subjects Matter. Beyond the Necessary Approximation of Finding Aids?» *JLIS.It* 14 (3): 3. https://doi.org/10.36253/jlis.it-539.

Valacchi, Federico. 2024. *L'archivio aumentato. Tempi e modi di una digitalizzazione critica*.

Vallverdú i Segura, Jordi. 2009. «Computational Epistemology and e-Science: A New Way of Thinking». *Minds and Machines* 19 (4): 557–67. https://doi.org/10.1007/s11023-009-9168-0.

Van Hooland, Seth, e Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata*. Facet Publishing.

Varagnolo, Davide, Guilherme Antas, Mariana Ramos, Sara Amaral, Dora Melo, e Irene Rodrigues. 2022. «Evaluating and Exploring Text Fields Information Extraction into CIDOC-CRM»: *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 177–84. https://doi.org/10.5220/0011550700003335.

Varagnolo, Davide, Dora Melo, Irene Pimenta Rodrigues, Rui Rodrigues, e Paula Couto. 2023. «Archives Metadata Text Information Extraction into CIDOC-CRM». In *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, a cura di Frans Coenen, Ana Fred, David Aveiro, et al. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43471-6_9.

Verona, Giulia. 2019. «Campiello: il premio dai lettori ai lettori. La giuria popolare termometro di cultura tra storia e lingua». In *Visto si premi: i retroscena dei premi letterari. Quaderni del Master di editoria 12*. Edizioni Santa Caterina.

Vilar, Polona, e Alma Šauperl. 2017. «Archivists about students as archives users». *Information Research* 22 (1): Colis paper 1617.

Vitali, Stefano. 2014. «La descrizione degli archivi nell'epoca degli standard e dei sistemi informatici». In *Archivistica. Teorie, metodi, pratiche*, di Linda Giuva e Maria Guercio. Carocci.

Vitali, Stefano. 2017. «Introduzione». *Quaderni del Mondo degli archivi* 2 (Records in Contexts: A conceptual model for archival description: Il contributo italiano): 3–7.

W3C Semantic Web Best Practices and Deployment Working Group. 2006. *Defining N-ary Relations on the Semantic Web*. Working Group Note. World Wide Web Consortium (W3C). https://www.w3.org/TR/swbp-n-aryRelations/.

Wallace, David A. 1995. «Managing the Present: Metadata as Archival Description». *Archivaria* 39 (1): 11–21.

Waltl, Bernhard, Gerald Bonczek, e Florian Matthes. 2018. «Rule-Based Information Extraction: Advantages, Limitations, and Perspectives». *Jusletter IT*, febbraio, 4.

Wei, Jason, Yi Tay, Rishi Bommasani, et al. 2022. «Emergent Abilities of Large Language Models». *Transactions on Machine Learning Research*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2022. «Chain-of-Thought Prompting Elicits Reasoning in Large Language Models». In *Advances in Neural Information Processing Systems*, a cura di S. Koyejo e others, vol. 35. Curran Associates.

Weston, Paul G., Primo Baldini, Emmanuela Carbé, e Laura Pusterla. 2019. «Archivi digitali di persona PAD – Pavia Archivi Digitali e gli archivi degli scrittori». *DigItalia* 14 (1).

Weston, Paul G., Emmanuela Carbé, e Primo Baldini. 2020. «Conservare e rendere accessibile un archivio letterario digitale: il caso di PAD – Pavia Archivi Digitali». In *Storie d'autore, storie di persone. Fondi speciali tra conservazione e valorizzazione*, a cura di Francesca Ghersetti, Alessia Martorano, e Elisabetta Zonca. Associazione Italiana Biblioteche.

White, Matt, Ibrahim Haddad, Cailean Osborne, et al. 2024. «The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence». https://arxiv.org/abs/2403.13784.

Wilson, Simon. 2012. «Arrangement and Description: Case Study: The Papers of Stephen Gallagher». In *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*. AIMS Work Group. https://escholarship.org/uc/item/1031p8xq.

Windows Team. 2012. «A New Way to Backup File History in Windows 8». dicembre. https://blogs.windows.com/windowsexperience/2012/12/20/a-new-way-to-backup-file-history-in-windows-8/.

Yakel, Elizabeth. 2011. «Balancing Archival Authority with Encouraging Authentic Voices to Engage with Records». In *A Different Kind of Web: New Connections between Archives and Users*, a cura di Karen Theimer. Society of American Archivists (SAA).

Yasmeen, Shazia, Irfan Ali, e Nosheen Fatima Warraich. 2019. «Personal digital information management practices of engineering faculty: finding, organizing, and re-finding information». *Pakistan Journal of Information Management & Libraries* 21: 88–103.

Zhang, Jane. 2012. «Original Order in Digital Archives». *Archivaria* 74 (novembre): 167–93.

Zhao, Yezhou, Xiaolin Duan, e Hua Yang. 2019. «Postgraduates' personal digital archiving practices in China: Problems and strategies». *The Journal of Academic Librarianship* 45 (5): 102044. https://doi.org/10.1016/j.acalib.2019.06.002.

Zou, Qing, e Eun G. Park. 2024. «Archival Context, Provenance, and a Tool to Capture Archival Context*». *Archival Science* 24 (4): 801–24. https://doi.org/10.1007/s10502-024-09457-7.

Żuchowska-Skiba, Dorota. 2024. «Hacktivism». In *Encyclopedia of Diversity, Equity, Inclusion and Spirituality*. Springer, Cham. https://doi.org/10.1007/978-3-031-32257-0_48-1.

Dichiarazione sull'uso di Intelligenza Artificiale Generativa

In conformità con le linee guida dell'Università di Bologna sull'utilizzo dell'Intelligenza Artificiale Generativa (GenAI)[281], si dichiara che nella redazione della presente tesi sono stati impiegati strumenti di GenAI in modo limitato alle seguenti attività:

- Revisione linguistica: alcune parti del testo della tesi sono linguisticamente riformulate con l'assistenza di *ChatGPT* (OpenAI GPT-5, 2025) e *Claude Sonnet 4.5*, al fine di migliorarne la chiarezza, la coerenza e il tono accademico, sotto la piena supervisione e revisione critica dell'autrice.
- Assistenza alla programmazione: gli script in linguaggio *Python* sono stati redatti con il supporto di *Claude Sonnet 4.5*. Tutto il codice è stato verificato, eseguito e validato dall'autrice.

Tutti i contenuti concettuali, strutturali e interpretativi, comprese l'analisi teorica, la progettazione metodologica e l'interpretazione dei dati, sono interamente frutto del lavoro originale dell'autrice. L'utilizzo degli strumenti di IA generativa è avvenuto nel pieno rispetto dei requisiti di trasparenza, supervisione critica e responsabilità stabiliti dall'Università di Bologna.

[281] https://www.unibo.it/it/ateneo/statuto-norme-strategie-bilanci/intelligenza-artificiale/intelligenza-artificiale