Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

CULTURE LETTERARIE E FILOLOGICHE

Ciclo 35

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

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

ONE ARCHIVE, MANY READINGS: PERSONAL ARCHIVES AS COMPLEX NETWORKS IN THE SEMANTIC WEB

Presentata da: Francesca Giovannetti

Coordinatore Dottorato

Marco Antonio Bazzocchi

Supervisore

Francesca Tomasi

Co-supervisore

Nicola Grandi

Esame finale anno 2023

One Archive, Many Readings

Personal Archives as Complex Networks in the Semantic Web

Francesca Giovannetti

PhD Dissertation
Department of Classical Philology and Italian Studies
University of Bologna
15 June 2023

Abstract

Personal archives are the archives created by individuals for their own needs and purposes. Among these are the library and documentary collections of writers and scholars. It is only recently that archival literature has begun to focus on this category of archives, emphasising how their heterogeneous nature necessitates the conciliation of different approaches to archival description, and calling for a broader understanding of the principle of provenance, recognising that multiple creators, including subsequent researchers, can contribute to shaping personal archives over time by adding new layers of contexts.

Despite these advances in the theoretical debate, current architectures for archival representation remain behind. Finding aids, whether paper-based or digital, still privilege a single point of view and do not allow subsequent users to embed their own, potentially conflicting, readings. Using semantic web technologies, particularly linked open data, this study aims to define a conceptual model for writers' archives based on existing and widely adopted models in the cultural heritage and humanities domains. The model developed can be used to represent different types of documents at various levels of analysis, such as record sets and individual records, as well as record content and components. It also enables the representation of complex relationships and the incorporation of additional layers of interpretation into the finding aid, transforming it from a static search tool into a dynamic research platform.

The personal archive and library of Giuseppe Raimondi (Bologna, 1898-1985) serves as a case study for the creation of an archival knowledge base using the proposed conceptual model. By querying the knowledge graph through SPARQL, the effectiveness of the model is evaluated. The results demonstrate that the model effectively addresses the primary representation challenges identified in archival literature, from both a technological and methodological standpoint. The ultimate goal is to bring the output *par excellence* of archival science, i.e. the finding aid, more in line with the latest developments in archival thinking.

Keywords: archival representation, digital finding aid, linked open data, semantic web, personal archives, textual scholarship.

Contents

List of Figures

List of Tables

Introduction

Personal archives are aggregations of the personal papers and other documents of individuals or families, formed over time for their own purposes and needs. A personal archive stands as evidence of its creator's life and activity. One particular category of personal archives is represented by writers' archives, which contain literary or research manuscripts in a variety of formats that relate to different writing stages, from initial notes to final versions. Such archives often hold also letters and other pieces of correspondence, either personal (family and friends) or cultural (other writers, academics, artists, etc.) – two spheres that most frequently overlap –, as well as printed material such as press clippings, books and journals.

These heterogeneous documents are interrelated to one another in complex ways, participating to multiple networks of interdependencies that are impossible to fully describe and represent within the constraints of a hierarchy oriented approach, let alone the relationships that they can bear to external documents and collections, including those from other cultural institutions such as libraries, archives, and museums.

The ability to provide meaningful context — or, better, contexts — for archival documents equates to the possibility of including complex, possibly conflicting, relationships in the representations of archives (e.g., non-hierarchical relationships, non-binary relationships, relationships that cross the boundaries of the single collection).

This thesis aims to contribute to the current debate in archival representation by exploring the challenges and opportunities of using semantic web technologies and, particularly, linked open data for creating multiperspective representations of personal archives that allow subsequent researchers to enhance the archival knowledge base with additional interpretations once the initial archivist's work of appraisal and description has been completed. The case study for the research described is

provided by the Giuseppe Raimondi Fonds, held at the Ezio Raimondi Library of the Department of Classical Philology and Italian Studies (FICLIT) of the University of Bologna, which comprises the author's personal archive and library.

The motivation for this research is explained in the first section of this introduction. The second section contains the research problem and questions. The third section anticipates the methodology and contributions of this work. The final section provides a road map for this thesis.

Research Motivation

The Ezio Raimondi Library of the Department of Classical Philology and Italian Studies (FICLIT) of the Universiry of Bologna houses over forty analogue writers' archives, mostly accumulated by former academics but also belonging to literary authors. Among these is the archive of the Bolognese writer Giuseppe Raimondi (1895-1985), which comprises the author's personal library (ca. 6,200 literature and art publications, 265 journal issues, ca. 50 exemplars of his published works) and papers (over 10,000 items including letters and postcards, drawings, photographs; 460 manuscript notebooks; 480 manuscript and typescript articles).[1]

My initial work towards completing this dissertation was motivated by the Department's goal of facilitating access to the personal archives housed in its library and supporting research on the wealth of documents contained within them. One key goal of the FICLIT Department is to enhance the existing archival descriptions of its personal archives and create high-resolution facsimiles of the documents they comprise. These facsimiles will be made publicly available online according to the International Image Interoperability Framework (IIIF) set of standards.[2]

The overarching aim of my research is twofold: on the one hand, I will use the Giuseppe Raimondi Fonds as a representative case study to rethink the representation of the Department's archival collections and, more broadly, of writers' archives in light of linked open data and semantic web technologies; the primary objective is the definition of a conceptual model and ontology for writers' archives capable of

[1] Federica Rossi and Alina Wenzlawski, 'Nello scrittoio di Giuseppe Raimondi: carte e libri di un letterato bolognese su Paul Valéry', in *Il privilegio della parola scritta: gestione, conservazione e valorizzazione di carte e libri di persona*, ed. Giovanni Di Domenico and Fiammetta Sabba (Roma: Associazione italiana biblioteche, 2020), p. 177.

[2] The International Image Interoperability Framework (IIIF), which I will discuss in more detail in Chapter 2, is "a set of open standards for delivering high-quality, attributed digital objects online" (https://iiif.io).

enabling a granular representation of heterogeneous types of documents and relations, as well as the integration of further and possibly conflicting contextualisations of the records by multiple scholars. On the other hand, I will take the opportunity to reflect on the methodological and conceptual implications of using knowledge graphs as the central and primary form of archival representation in cultural institutions focusing in particular on the profound ways in which semantic web technologies challenge some of the traditional principles of archival science. Before I move into the substance of this dissertation, I should specify that, due to the nature of my case study, my research is focused on non-digital archives. The additional personal archives which the Department will likely acquire in the near future will almost undoubtedly be in a hybrid, if not entirely digital, form: therefore, this study will require integrations.

Research Problem and Questions

Knowledge graphs have only recently been getting attention as a means for archival representation and the creation of finding aids. The data structure of knowledge graphs, compared to that of tree hierarchies, supports the creation of representations that are able to express higher orders of archival complexity. The descriptive potential of knowledge graphs is grounded on the Semantic Web architecture, which adopts the Resource Description Framework (RDF) as the basic, interoperable data model to convey information through semantic statements taking the form of subject-predicate-object expressions. Linked data as a set of publishing principles and knowledge graphs as a form of representation open up unprecedented possibilities for augmenting the descriptions of the Department's personal archives with multidimensional relationships between individual records and data, superseding the exclusively hierarchical organisation of traditional archival representation.

In light of such considerations, this thesis aims to use semantic web and linked open data technologies to improve the representation of analogue writers' archives. The goal is to transform these archives into dynamic research tools capable of expressing complex contextualizations of records and adapting over time as new relationships between records and data emerge.

The research question that this study seeks to answer is how the use of semantic web and linked open data technologies can enhance the representations of analogue writers' archives. This question is broken down into three subparts (**RQs**):

RQ1 How can semantic web technologies and linked open data be used to represent heterogeneous types of documents and relationships in a multidimensional way, particularly in complex scenarios involving multiple interconnected documents?

RQ2 How can records and their relationships be described at different levels of granularity, such as record sets, individual records, record parts, as well as content and text?

RQ3 How to integrate 'a' reading of the archive with archival description? How to represent such a reading so to make new users of the archive able to explore and evaluate such a reading? How to represent conflicting, yet coexisting, readings of the archive by ontologies?

The objective of this study is to create a conceptual model for writers' archives. The objective consists of three parts (**ROs**), each corresponding to one of the research questions mentioned above:

RO1 The conceptual model should enable a multidimensional representation of diverse types of records, allowing multi-entity descriptions to be linked to individual documents such as notebooks, letters, postcards, albums, and annotated printed books;

RO2 The conceptual model should enable the representation of records and their relationships at various levels of granularity, including record sets, individual records, and record content/components, with a particular emphasis on representing relationships among texts;

RO3 The conceptual model should enable the inclusion of multiple and possibly conflicting viewpoints in the form of supplementary interpretations by researchers in the finding aid over time.

The following section provides a brief summary of the research approach adopted for the study, outlining the methodology used to achieve the study's objectives.

Research Methodology and Contributions

The Giuseppe Raimondi Fonds was used as a case study in this research. The fonds is highly representative of the Department's collection of writers' archives because

it contains a wide range of document types. Relevant archival and digital humanities literature was reviewed in order to gain a comprehensive understanding of the characteristics and challenges that distinguish personal archives and, in particular, writers' archives, and to investigate the current archival thinking regarding their representation.

The case study was then examined using the existing finding aid as well as published research on Giuseppe Raimondi's fonds to identify the objectives and potential complexities of its representation.

For each type of document in Giuseppe Raimondi's fonds, a set of competency questions was developed.

In order to identify models suitable for reuse, a review of existing conceptual models and ontologies for the representation of archives was conducted.

Based on the competency questions generated in the previous phase, a conceptual model and ontology were created.

Using a set of ad hoc Python scripts, the ISBD finding aid for Giuseppe Raimondi's fonds was converted into an RDF dataset based on the conceptual model previously defined.

The data were enriched in various ways to show how the base archive description can be supplemented with additional reconstructions at various levels of granularity, such as record sets, individual documents, and content.

The obtained dataset was evaluated using the initial competency questions as well as additional ones.

The contributions of this study are as follows:

1. A conceptual model for writers' archives based on widely-adopted standards and ontologies that allows for multiperspective, ever-evolving representations of heterogeneous documents and data.
2. The RDF representation of Giuseppe Raimondi's archive and library, based on the conceptual model defined in the framework of this study and enriched with examples of multiple interpretations by subsequent researchers.
3. A set of Python scripts for transforming the existing ISBD descriptions of the the FICLIT Department's archival collections into RDF statements.

Thesis Outline

This work begins with introducing Giuseppe Raimondi's personal archive as a representative case study and continues with a review of the literature related to personal archives and archival representation (Chapter 1). It then provides a survey of the state of the art concerning conceptual models and ontologies for the representation of archives and cultural heritage objects in general (Chapter 2). The methodology adopted for this study is then described (Chapter 3).

The second part illustrates the conceptual model (Chapter 4) and its evaluation through creating and querying an enriched RDF finding aid of Giuseppe Raimondi's archive and library (Chapter 5). The work concludes with a discussion of the contribution and limits, both methodological and practical, of the obtained results (Chapter 6).

Chapter 1

Personal Archives: Challenges and Opportunities for Representation

This chapter provides background for this study through literature. It begins by discussing the archival literature on personal archives, concentrating on the practices and challenges of archival representation. Over the years archival researchers have highlighted limitations in current archival representation practice and have investigated the resulting gaps in finding aids. Here, I introduce some of their contributions, focusing on reinterpretations of the so-called principle of archival provenance. Archival theory has traditionally been oriented towards strictly attributing the fonds to a single named creator, namely, the individual from whom the archival tree figuratively originates. More recent debates call for archival representations, and thus for finding aids, acknowledging the contributions of multiple creators in the construction of the fonds. This new orientation opens promising opportunities for finding aids, no longer seen as passive (re)search tools but as evolving knowledge bases designed to incorporate multiple and different scholarly readings of the archives they represent.

The chapter moves onto introducing the theme of modelling in the digital humanities, aiming to frame archival representation as a modelling activity whose output is only one of the many possible. This thought prompts the following question: how can multiple perspectives be combined into a single finding aid? The chapter addresses the problem by analysing different ways of shaping archival data, from tables, to trees, to graphs. Literature from various fields, including archival theory, digital humanities, and digital textual scholarship, is discussed with the aim to support the argument that complex, or layered, graphs are highly suitable for

representing archives as multidimensional and multiperspective entities.

The goal of this dissertation is the exploration of ways to augment archival representation with additional, and possibly competing, contextualizations of the records. The types of contextualizations I am particularly interested in incorporating into the finding aid are the interpretations given by subsequent users-researchers of the archive. Such interpretations, or contextualizations, configure themselves as complex networks of relationships between heterogeneous types of archival entities including documents and document sets, persons, organisations, dates, places. Supplementing archival representation with multiple contextualizations implies a shift from a uni-dimensional, multi-level approach to a multi-dimensional, granular approach primarily centred on the single item and its intrinsic components rather than on record sets.

The model proposed in this research aims to address some of the challenges in representation highlighted in archival literature from both a technological and methodological point of view, bringing the scholarly output of archival science par excellence, the finding aid, closer to the most recent archival thinking. In particular, two main issues with current archival representation standards are outlined. First, existing archival descriptions prioritise the archival hierarchy rather than individual documents. Second, despite its importance in context reconstruction, current models rarely implement content representation.[1] Moving from the assumption that linked open data and semantic web technologies can provide an effective means of representing complex networks of relationships, this study contributes to the enhancement of the FICLIT Department archival collections through the definition of a conceptual model for the representation of personal archives which integrates different disciplinary perspectives, including literary and textual scholarship. This study also argues that reconfiguring the finding aid as a knowledge graph broadens its role in the context of literary and textual studies by allowing multiple interpretations by archivists and researchers to be incorporated into the archival representation as additional collections of interrelationships.

Experiments with CIDOC-CRM as a base ontology for representing archival information have already proved fruitful for demonstrating how classes and properties from CIDOC-CRM could also be leveraged for creating representations in the

[1] For more on the importance of content as a source of information for enriching our representations of archival collections see Francesca Tomasi and Marilena Daquino, 'The archival domain in a disciplinary-integrated ontological perspective', Number: 3, *JLIS.it* 6, no. 3 (15 September 2015): 23, accessed 1 May 2023, https://doi.org/10.4403/jlis.it-11133.

archival domain.[2] However, none of these experiments have dealt specifically with the representation of the life cycle of writers' archives, and especially of the role of subsequent users-researchers as creators of additional interconnections between texts and documents. In addition, archival description practices to date have substantially focused on the representation of record sets rather than individual documents and texts.[3] On the other hand, most existing digital scholarly editions prioritise a document-centred view of texts that uses TEI/XML markup over LOD-based representations of textual phenomena and do not address the representation of the archival dimension in which texts participate.[4]

1.1 The Personal Archives of Writers and Intellectuals

Personal archives are those created by individuals for their own purposes and needs. Among the various types of personal archives are those belonging to writers and intellectuals. The Ezio Raimondi Library of the University of Bologna holds approximately 60 of such personal archives.[5] The majority date from the twentieth century and belonged to scholars, lecturers, but also literary authors. Many of the library's archives comprise both documentary and bibliographic materials including books, periodicals, manuscripts, typescripts, notebooks, notes, letters, postcards, photographs, drawings, etc. Books, periodicals, and other bibliographic items often feature handwritten marginalia and annotations, and sometimes contain further documents.

[2] See Lina Bountouri and Manolis Gergatsoulis, 'Mapping encoded archival description to CIDOC CRM', in *First Workshop on Digital Information Management* (2011), 8–25; Marilena Daquino et al., 'Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data', Publisher: ACM New York, NY, USA, *Journal on Computing and Cultural Heritage (JOCCH)* 10, no. 4 (2017): 1–21; Inês Koch et al., 'Knowledge graph implementation of archival descriptions through CIDOC-CRM', in *International Conference on Theory and Practice of Digital Libraries* (Springer, 2019), 99–106.

[3] For an in-depth discussion of this topic, see Geoffrey Yeo, 'Contexts, Original Orders, and Item-Level Orientation: Responding Creatively to Users' Needs and Technological Change', *Journal of Archival Organization* 12, no. 3 (2015): 170–185.

[4] One exception is the digital edition of Paolo Bufalini's Notebook, which describes intratextual relationships in LOD (Marilena Daquino, Francesca Giovannetti and Francesca Tomasi, 'Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini', *Umanistica Digitale*, no. 7 (2019): 49–75, accessed 19 December 2019).

[5] The Ezio Raimondi Library is the Humanities library of the Department of Classical Philology and Italian Studies of the University of Bologna. As of November 4, 2022, the total number of personal archives held by the library was 59. An up-to-date list of collections can be found at https://bur.sba.unibo.it/chi-siamo/fondi-bibliografici-e-documentali-della-sezione-di-filologia-classica-e-italianistica.

These archives are unique sources for research and study, not only for the diverse individual items they contain, but also because of the heterogeneous interconnections that aggregate such items into unitary, meaningful systems. The study of these systems support the development of a deeper understanding and interpretation of the archive creator's work and life. Such interconnections exist not only between an author's documentary records, that is her or his papers, but also between the papers and the library.[6] Within the framework of each distinct archive and between related archives, the different items forming part of the documentary and bibliographic collections are inextricably linked and dialogue with one another. These interconnections extend beyond the canonical boundaries of archival representation to reach other cultural heritage resources in the library and museum domains. As Tomasi puts it:

> Gli archivi di persona e, in particolare, di personalità, conservati presso istituzioni archivistiche, o donati alle biblioteche, rappresentano un'interessante prospettiva di studio perché sono portatori di un sapere che travalica i principi della descrizione archivistica per abbracciare i sistemi di descrizione dell'intero patrimonio culturale, rappresentato da documenti, ma anche da libri, artefatti o ancora risorse e oggetti, anche nativi digitali, in un sistema di profonde interconnessioni spesso latenti.[7]

Although it is not fully reflected in the current descriptions of the archives, such an interdependency clearly represents an invaluable resource for gaining insight into the documents – including autographs. Some scholars even argue that the books and publications in a writer's personal library bear significance not only because they may feature annotations, but also for the sole reason of being included in the collection.[8] Among them, Caproni has emphasised on several occasions the

[6] The documentary and the bibliographic collections that make up the personal archives of the Ezio Raimondi Library are conceptually preserved as unitary fonds. The descriptions of the collections, however, are logically separated, so that the label 'archive' is often used to refer solely to the papers, whereas 'library' refers to the bibliographic collections. The latter are not explicitly attributed to the personal archive they belong to, but their provenance can be deduced from their shelfmark, which indicates their physical location within the Ezio Raimondi Library.

[7] Francesca Tomasi, 'Archivi di persona in linked open data: il modello concettuale come strumento di integrazione nei GLAMs', *AIB studi* 57, no. 2 (20 November 2017): 284, accessed 13 September 2019, https://doi.org/10.2426/aibstudi-11647.

[8] Sabba, along with other archival scholars, claims that the inclusion of marginal notes, underlines, bookmarks, and dog-eared pages gives books an archival value, transforming them from mass-produced items into unique documents on par with autograph manuscripts (Fiammetta Sabba, 'Biblioteche e carte d'autore: tra questioni cruciali e modelli di studio e gestione', *AIB studi* 56, no. 3 (2016)). Before Sabba, Crocetti had called for the personal books of writers to be considered

value of exposing the interrelationships between an author's papers and books. He contends that personal archive descriptions should avoid treating papers and books as logically distinct entities for, to fully understand authors and their literary or scholarly production, three different dimensions of their writing activity should be considered. Instead of studying authors solely as writers, which focuses scholars' attention primarily on the autographs that have survived, they should also be studied as persons who publish and, most importantly, read.[9] Archival descriptions, in particular, should ideally convey the reconstruction and the explication of the connections that exist between autographs (the author as a writer), the published versions of such autographs (the author as a publisher), and the readings that influenced the writing process (the author as a reader), with the latter dimension being crucial for scholars' and intellectuals' archives.

Aside from the fact that they contain deeply interconnected papers and bibliographic resources, Crocetti, a library scholar who extensively researched twentieth-century writers' archives, identifies a number of other distinguishing features of these particular category of personal archives. For example, he notices that they are most commonly found in public or university libraries rather than state archives. Libraries are seen as the most appropriate and cohesive settings for facilitating access to and contextualising the materials, as well as enhancing their dual nature as documentary and bibliographic. State archives, on the other hand, primarily store administrative and financial documents and serve a broader, less specialised audience. As a result, they are generally thought to be unsuitable for preserving personal archives.[10] Crocetti, like Caproni, recognises the immense value of bibliographic records in the archives of writers and intellectuals. In light of this, he strongly advises archivists and scholars to treat these records with the same respect as other archival documents such as manuscripts and autographs. He suggests that one of the primary responsibilities of archivists and archival researchers should be to regard both papers and books as equally important, and to create maps that allow for easy navigation between these two types of materials.[11]

archival materials solely on the basis of their inclusion in the writer's archive, regardless of whether or not they bear signs of reading or use (Luigi Crocetti, 'La tradizione culturale italiana del Novecento', *AIAA Journal* 17, no. 9 (1999): 25).

[9] Attilio Caproni, 'Biblioteca privata: ipotesi di definizione', *Bibliotheca* 5, no. 1 (2006): 22–28; by the same author see also Attilio Caproni, 'Le biblioteche d'autore: definizione, caratteristiche e specificità: alcuni appunti', *Bibliotheca* 3, no. 1 (2004): 15–22.

[10] Andrea De Pasquale, 'Private archives in the library. Types, acquisition, treatment and description', 10, no. 3 (2019): 34–46.

[11] "Libri e carte sono da porre sullo stesso piano, ciò che dobbiamo fare è disegnare una mappa"

Another shared characteristic of twentieth-century writers' archives, according to Crocetti, is their size and the diverse range of document types they encompass. In comparison to personal archives of earlier centuries, twentieth-century writers' archives are extraordinarily large and feature an incredibly wide range of document types. The variety and multiplicity of documents found in these personal archives far exceed those found in collections from previous centuries.[12] For example, Giuseppe Raimondi's archive – the case study adopted for this research – features autograph manuscripts as well as typescripts but also personal collections of books (often annotated), notebooks, cut-out articles from newspapers and other types of publications, photographs (some gathered in thematic albums), private correspondence (mostly received letters, postcards, and invitations). Correspondence, in particular, contributes to broadening the scope of the archive beyond the original creator's work to include the works of other writers, scholars, and artists, including unpublished texts. Incoming letters and other forms of correspondence enable archive users to examine the original creator's work within the larger intellectual community to which they belonged, allowing them to cross the archival boundary and access other personal archives. Given this, Crocetti proposed classifying these collections using the broader label 'cultural archives' rather than 'literary archives'.[13]

Crocetti also underlines the heterogeneous nature of twentieth-century writers' and intellectuals' archives. Autographs are no longer the only meaningful records that a writer's archive may contain; they interact with other types of documents forming an inseparable and meaningful unit. Papers, books, letters, notebooks, newspaper clippings, and photographs are all essential elements for interpreting and explicating the fonds and its creator. For example, a particular passage of a manuscript may be better understood considering a particular book owned by the author and possibly annotated; or data from a letter may cast light on the genesis of a work or its writing process. Literary and textual research in writers' archives should not only focus on autographs but should also leverage any other types of

(Crocetti, 'La tradizione culturale italiana del Novecento', 141).

[12] Laura Desideri, 'Le raccomandazioni di Luigi Crocetti', in *Conservare il Novecento: gli archivi culturali. Ferrara, Salone internazionale dell'arte del restauro, 27 marzo 2009: atti del convegno*, ed. Laura Desideri and Giuliana Zagra (Roma: Associazione italiana biblioteche, 2010), 9, https://doi.org/10.1400/285768.

[13] The Italian expression suggested by Crocetti for large twentieth-century writers' archives that comprise highly heterogeneous types of document is 'archivi culturali'.See Luigi Crocetti, 'Parole introduttive a Conservare il Novecento', in *Conservare il Novecento: convegno nazionale, Ferrara, Salone internazionale dell'arte del restauro e della conservazione dei beni culturali e ambientali, 25-26 marzo 2000: atti*, ed. Maurizio Messina and Giuliana Zagra (Roma: Associazione italiana biblioteche, 2001), 24.

available documents (e.g., books, letters, photographs, clippings, etc.) within the archive and in related archives so to develop a complete understanding of the original creator's production. Continuing the map metaphor, Crocetti claims that archivists and researchers working with archival materials should act as cartographers sketching maps of the paths that connect the different documents in the collections.[14] Therefore, an ideal finding aid should support users in navigating the archive as a network of meaningfully interrelated documents.

In the case of the Giuseppe Raimondi archive, and of the FICLIT Department's collection of archives in general, researchers have given multiple interpretations providing interconnections between the documents and other archival data. This, according to Thomassen, is precisely the nature and goal of archival research, which he defines as "research on relations".[15] Examples of research on the relations within Giuseppe Raimondi's personal archive includes the reconstruction of the reception of the work of Paul Valéry in Raimondi's literary production;[16] the transcription and analysis of the private correspondence between Raimondi and Giuseppe Ungaretti;[17] the connection between the draft manuscripts contained in Raimondi's archive and the corresponding published works. Giuseppe Raimondi himself processed selected documents of his archive to assemble thematic albums dedicated to specific correspondents.[18] Regardless of their relevance, such scholarly reconstructions are neither described in the existing finding aids of the collections nor used to improve the user experience, despite finding aids having a crucial function in the research cycle as they are fundamental tools frequently serving as the starting point and primary point of reference for researchers.

1.2 Representing Personal Archives

The term 'representation' is commonly used to refer to both the process or activity of representing as well as the final product or object that results from that activity.

[14] Crocetti, 'Parole introduttive a Conservare il Novecento'; see also Crocetti, 'La tradizione culturale italiana del Novecento'.

[15] Theo Thomassen, 'A first introduction to archival science', Publisher: Springer, *Archival science* 1, no. 4 (2001): 373.

[16] See Rossi and Wenzlawski, 'Nello scrittoio di Giuseppe Raimondi'.

[17] See Eleonora Conti, ed., *Giuseppe Ungaretti. Lettere a Giuseppe Raimondi, 1918-1966*, vol. 53 (Pàtron, 2004).

[18] One example is the album dedicated to Filippo De Pisis, which contains 15 photographs, 41 original drawings and 27 reproductions, 5 manuscripts, a book, 9 invitation cards, and a poster in addition to 86 letters and postcards. See Giuseppe Raimondi, 'Album De Pisis' (Album, BFICLIT FR.A ALBUM DEPISI, RDm 835, 1916).

According to Yakel, representation in the archival context includes both the process of arrangement, which is the process of (re)constructing order, and the process of description, which is the process of developing tools and systems for collection access and management, as well as surrogates or models of the archival collection.[19] Finding aids, realised in different forms, analogue or digital, convey archival representations while serving as both access tools for users and collection management tools for archivists.

The archival work of representation is guided by two interconnected principles: the principle of *respect pour les fonds* and the principle of respect for original order. The principle of *respect pour les fonds* states that records created by an individual or organisation should be kept together and never mixed with others, whereas the principle of respect for original order states that records should be kept in the same sequence and arrangement as established by the original creator. Both principles, according to Horsman, can be thought of as subprinciples of the broader principle of provenance.[20] Adherence to such principles ensures the preservation of the archival bond, defined as "the network of relationships that each record has with the records belonging in the same records aggregation".[21] In order to ensure adherence to these principles, archivists have traditionally used a hierarchical representation system in which the fonds sits at the top of a reversed tree, from which the description narrows as one progresses down the tree to the series, to the file, and finally to the item level. Hierarchical relationships between records and their creators form the architecture of the finding aid and, naturally, the most immediately visible paths for users of the archive to follow in their navigation.

When personal archives are examined, however, these principles – or, more specifically, the way they have traditionally been understood and applied to record description – are called into question. Cook, for example, emphasises how the arrangement or order given to an archive after processing should be viewed as a representation of one of many possible orders, given that what is customarily referred to as 'original' order, like the archetype in textual criticism, is nearly impossible to reconstruct or even never existed in the first place. MacNeil, in line with Cook, puts into question the concept of original order and the way it is understood among

[19] Elizabeth Yakel, 'Archival representation', *Archival Science* 3, no. 1 (2003): 1–25.

[20] Peter Horsman, 'Taming the Elephant: An Orthodox Approach to the Principle of Provenance', in *The Principle of Provenance: Report from the First Stockholm Conference on the Archival Principle of Provenance, 2-3 September 1993* (Stockholm: Swedish National Archives, 1994), 51–63.

[21] InterPARES Trust, *The InterPARES 2 Project Glossary: archival bond* (2024), 5, accessed 17 February 2023, http://interparestrust.org/terminology/term/archival%20bond.

archival practitioners. She maintains that original order is "simply one of many possible orders a body of records will have over time".[22]

The archival literature also challenges the notion of provenance, which – as explained earlier – is perceived as closely tied to *respect pour les fonds*. Hurley, for example, emphasises how important it is acknowledging that multiple and different creators can influence and shape a personal archive over time.[23] Similarly, Douglas argues for including "in the description of archives a fuller representation of different types of creators", such as researchers who prove new insights into the collections as they engage with them after the initial processing by the archivist.[24] Light and Hyry also noted that finding aids do not include additional interpretations by researchers, instead presenting a single point of view on the material they describe.[25] Despite advancements in digital archival representation since Light and Hyry's article, the situation remains largely unchanged.

Yakel points out that finding aids serve as both access tools for users and collection management tools for archivists. While collection management is primarily based on the hierarchical grouping and organisation of records, access to records by users-researchers begins with the hierarchy but gradually moves away from it to embrace graph-like paths of discovery and meaning construction.[26] The reflective literature on archival description has long argued against an exclusively hierarchical approach to record contextualization. For instance, the same Yakel noted that "the very act of archival representation, designed to order ad provide access to collections through finding aids, can also create barriers to use".[27] This had also been emphasised by Boles in his article titled 'Disrespecting Original Order', in which he claims that "the ordering imposed upon documents by their creator has evidential value. The ordering makes a statement about the creator. Despite this, original order can be changed. [...] Original order is to be respected when it is usable".[28]

[22] Heather MacNeil, 'Archivalterity: Rethinking Original Order', *Archivaria*, 2008, 24.

[23] Chris Hurley, 'Parallel provenance [Series of parts]: Part 1: What, if anything, is archival description?', An earlier version of this article was presented at the Archives and Collective Memory: Challenges and Issues in a Pluralised Archival Role seminar (2004: Melbourne). *Archives and Manuscripts* 33, no. 1 (May 2005): 110.

[24] Jennifer Lynn Douglas, 'Archiving Authors: Rethinking the Analysis and Representation of Personal Archives' (PhD diss., University of Toronto, 2013), 196; See also Jennifer Douglas, 'A call to rethink archival creation: exploring types of creation in personal archives', Publisher: Springer, *Archival Science* 18, no. 1 (2018): 29–49.

[25] Michelle Light and Tom Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', Publisher: Society of American Archivists, *The American Archivist* 65, no. 2 (2002): 216.

[26] Yakel, 'Archival representation'.

[27] Yakel, 2.

[28] Frank Boles, 'Disrespecting Original Order', Publisher: Society of American Archivists, *The*

According to Boles, one important guiding principle of archival representation, particularly arrangement, should be usability. This entails prioritising subsequent the ability of the users to access documents over a less efficient or nonexistent filing system imposed on the archive by its creator.[29] Similarly, in 1940, Posner proposed that the principle of usability be followed not so much in arranging physical documents as in arranging surrogates – or models – of such documents, in order to allow the coexistence of multiple orders.[30] This would preserve the 'original' order while providing users with more effective, possibly non-hierarchical, orders for accessing and interpreting the documents and the relationships that link them one to the other.

These observations all suggest that the archival hierarchy is only one of many possible contexts for a record and that additional contexts can exist beyond it. Considered in this light, the underlying architecture of the archive becomes that of a hypertext, with multiple access points and paths to follow in addition to the primary one, which usually corresponds to the name of a single creator at the top of the archival hierarchy.[31]

1.3 The Archival Fonds as a Conceptual Model

The introduction of knowledge graphs into archival description practices prompts a need for a broader understanding of the concept of archival fonds and opens up new possibilities for archival representation. Knowledge graphs not only provide archival practitioners with more advanced technological means for description, but they also challenge traditional approaches to archival representation both methodologically and conceptually. From a methodological perspective, the use of knowledge graphs in archival representation leads to a shift in the structure and organisation of the archive from a fixed hierarchy of records to a fluid network of logical interdependencies that can be arranged and rearranged into new representations. The semantic range of these interdependencies is virtually infinite, as terms from multiple different ontologies can be combined within one graph, and new relation-

American Archivist 45, no. 1 (1982): 29–32, accessed 26 March 2021, http://www.jstor.org/stable/40292457.

[29] Boles, 'Disrespecting Original Order', 29–32.

[30] Ernst Posner, 'Some Aspects of Archival Development since the French Revolution', Society of American Archivists, *The American Archivist* 3, no. 3 (1940): 169.

[31] See Diana Toccafondi, 'Gli archivi letterari del Novecento: un laboratorio per la collaborazione tra professionisti', in *Conservare il Novecento: gli archivi culturali. Ferrara, Salone internazionale dell'arte del restauro, 27 marzo 2009: atti del convegno* (Associazione italiana biblioteche, 2010), 44, accessed 27 January 2022, https://www.torrossa.com/it/resources/an/5088757.

ships can be defined when necessary. From a conceptual perspective, rethinking the archival fonds as a graph model points towards new directions for finding aids.

In his article 'The Conceptual Fonds and the Physical Collection', Yeo explores the relationship between two fundamental concepts in the world of archives, namely fonds and collection.[32] In traditional archival theory, fonds are generally understood to be naturally accumulated, while collections are intentionally or artificially assembled. The focus of archival science is on the former, rather than the latter. However, Yeo proposes a different perspective on the traditional opposition between fonds and collection. He proposes that collections be viewed as physical aggregations, whereas fonds be viewed as conceptual: the body of records that materially reaches the cultural institution designated for its preservation is a physical collection of documents that can be objectively observed and circumscribed. The fonds, on the other hand, is a conceptual representation of the collection and of any other records to which such a collection may be related, independently of their physical location. Collections are what archivists manage, preserve, deliver, and examine in the real world, whether these comprise analogue or digital items. The reformulation proposed by Yeo acknowledges Terry Cook's observation that linking the notion of fonds to a logical concept and a physical object at the same time is problematic.[33] This is true especially when it comes to personal archives. Distinguishing between natural versus artificial accumulation in personal archives can be difficult as creators often make conscious choices about which documents to keep and which to discard. The traditional concept of fonds as an organically accumulated body of records is thus difficult to apply consistently to this category of archives.

The fonds can in fact involve records that are outside of the collection, for example, those that are part of other archives, library collections, or even museums. A simple but significant example is given by the correspondence often kept in author archives, including that of Giuseppe Raimondi. In many cases, an author's archive only contains the letters, postcards, and other documents received, and only in rare cases are copies of the author's own responses present, whereas replies are often found in the archives of other personalities. Therefore, the fonds of one author and another could overlap in this perspective, including the complete correspondence even though it physically belongs to another collection. Another example of the non-correspondence between the physical collection and the fonds as a model is

[32] Geoffrey Yeo, 'The Conceptual Fonds and the Physical Collection', *Archivaria* 73 (2012): 43–80.
[33] Terry Cook, 'The Concept of the Archival Fonds in the Post-Custodial Era: Theory, Problems and Solutions', *Archivaria* 35 (1992): 25.

also offered by the case of Giuseppe Raimondi. If one analyses the finding aid, which conveys the reconstruction of the fonds, a gap in the shelfmark number of the manuscripts of articles can be noticed, which is not reflected in the inventory number. This jump is due to the fact that one of the documents was probably retained by the heirs at the time of the transfer of the physical collection from the author's home to the university library. Though physically missing from the collection, conceptually the document is still part of the fonds. it is still part of the fonds, even though it is not included in the physical collection. Another significant example is represented by the notebook of Paolo Bufalini also preserved in the Department's Library. The notebook is physically separated from its author's archive, as it was donated to the Department by his heirs due to the presence of unpublished translations from Latin to Italian within it, which make this document particularly significant for classical philology studies. The notebook is composed of fragments of quotes and comments. It can be imagined how each of these fragments is connected to Bufalini's readings, and therefore to the volumes present in his personal library, still in possession of the family, which are likely to be rich in underlines and margin notes that would certainly contribute to further expand and contextualise the contents of the notebook.[34]

From this perspective, the concept of personal fonds can be given a new definition as the full set of interrelated documents produced and received over a person's lifetime. These records may actually exist in separate collections and might have never been physically united,[35] but the use of digital technology in virtual environments could make it possible. Each collection housed in the Ezio Raimondi Library of the Department may not correspond to a distinct personal fonds. Rather, when we consider the context of the archival heritage of the library of the FICLIT Department, we can see that the personal fonds of the different creators are inextricably linked and intertwined. The relationships between these creators, whether they are direct or indirect, give rise to a complex and interconnected system of records that reflect the activities of their creators. An interesting example of this reality in the Giuseppe Raimondi Fonds is provided by an autographed copy of *I tetti sulla città*, a work by Giuseppe Raimondi published in the late 1970s. The copy, which contains

[34] An in-progress digital edition of Paolo Bufalini's notebook can be found at https://projects.dharc.unibo.it/bufalini-notebook. See Daquino, Giovannetti and Tomasi, 'Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini'.

[35] Geoffrey Yeo, 'Bringing Things Together: Aggregate Records in a Digital Age', *Archivaria*, 2012, 49.

underlines and annotations by Ezio Raimondi, was gifted by the author to Ezio Raimondi.

Following these considerations, one might conclude that reconstructing the fonds is nearly impossible because not all of its elements are part of the same collection at the same time, and the collection can expand in multiple directions. "A collection is rarely 'the whole of the records', but is usually a fragment of that whole" at a given time.[36] Moreover, several scholars argue that there is no unique, original fonds, just as there is no single context for a document.

Because the resulting surrogates constitute different representations or models of the archive, archival description can be thought of as a modelling activity. According to different scholars, archival arrangement too is an activity that generates a representation, because it necessitates an unavoidable recontextualisation of the material and, as a result, the development of a new model of the archival bond. Cook, in particular, emphasises how the order given to an archive after processing should be viewed as a model of one of many possible orders, given that what is customarily referred to as 'original' order, like the archetype in textual criticism, is nearly impossible to reconstruct or even never existed in the first place.[37]

Adopting this perspective, the fonds is to be understood as a representation or model that represents the records of the collection in context, reconstructing the relationships that these records have with each other. As such, distinguishing between a collection as a physical entity and a fonds as a conceptual entity, allows us to imagine the possibility of different fonds or representations or models coexisting for the same collection. Yeo also states that "one fonds may not be wholly discrete from another".[38] On the other hand, "identification of collections is less subjective. As material entities, they may change over time, but at any given moment their boundaries are usually easy to recognise".[39] In this perspective, where the fonds is independent of the collection, the idea that the fonds and the collection must necessarily coincide is abandoned.[40]

From this perspective, the fonds is seen as a completely subjective construct. As Duff and Harris put it:

> No approach to archival description, no descriptive system or architec-

[36] Yeo, 'The Conceptual Fonds and the Physical Collection', 74.
[37] Cook, 'The Concept of the Archival Fonds in the Post-Custodial Era'.
[38] Yeo, 'Bringing Things Together', 49.
[39] Yeo, 49.
[40] Yeo, 'The Conceptual Fonds and the Physical Collection'.

> ture, can escape the reality that it is a way of constructing knowledge through processes of inscription, mediation, and narration. No architecture can escape the bias of its developers. Archivists cannot describe records in an unbiased, neutral, or objective way. [...] The narrative construction attempts to give [records] a shape, a pattern, a closure – to end their inevitable openness.[41]

Similarly, Yakel states that "archival representation processes are neither objective nor transparent".[42] Therefore, it is essential to ensure that this inherent subjectivity of representation, and thus of its expression, the finding aid, is made visible and apparent to the end-user.[43] The use of linked open data enables metadata to become data in an increasing layer of complexity, allowing for *describing the description* and thus explicating information about the archival description process by formalising it.[44]

Yakel's emphasis on "open attribution" is an essential aspect of acknowledging and addressing the subjectivity inherent in creating finding aids.[45] This is because finding aids, like any form of archival representation, are not objective or transparent. Instead, they are shaped by the biases and perspectives of those who create them. Therefore, it is crucial to make these biases and perspectives explicit, particularly given that no two scholars would approach the same object in the same way or make the same connections. This issue is particularly relevant to the description of logical relationships between objects within an archive. As Yeo notes, different observers may have different interpretations of the nature and extent of these relationships: "the nature and extent of logical relationships can be open to dispute; a relationship [...] may be apparent to one observer but not to another".[46]

A personal fonds is thus a representation or model, among many possible ones, that captures an individual's personal records and the meaningful relationships that define them. The finding aid serves as a means of conveying this model. Digital Humanities scholarship is largely concerned with the activity of modelling humanistic

[41] Wendy M. Duff and Verne Harris, 'Stories and names: archival description as narrating records and constructing meanings', *Archival Science* 2, no. 3 (2002): 275.
[42] Yakel, 'Archival representation', 25.
[43] Light and Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', 217.
[44] A clear boundary between data and metadata in the linked open data realm is hardly traceable as further layers of metadata can always be added on top of existing ones to describe other metadata, increasing the complexity of the representation.Seth van Hooland and Ruben Verborgh, *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata* (London: Facet Publishing, 2014)
[45] Yakel, 'Archival representation', 25.
[46] Yeo, 'Bringing Things Together', 88.

knowledge to make it machine-readable. Scholars agree on the idea that modelling is more than straightforward data structuring. Ciula and Eide, for example, describe modelling as a "creative process of thinking and reasoning where meaning is made and negotiated through the creation and manipulation of external representations".[47] Modelling is an intellectual process which leads to the creation of a conceptual model on the basis of close analysis and interpretation of the domain under study. Being independent of any formalisation, conceptual models take often the form of entity-relationship diagrams that precede and inform the implementation of knowledge graphs.

As the product of a subjective process of meaning-making, a conceptual model can only convey its modeller's personal interpretation of the modelled object; there is no such thing as an objective model. Problematic as it may seem, this nature aligns models with the very nature of humanistic inquiry, whose richness is found also in the availability of multiple and potentially conflicting perspectives on the same topic.

These premises suggest a parallel between the conceptual model and the archival fonds. Archival theory has traditionally given primacy not to the collection but to the fonds. Nonetheless, attempts made by some scholars to draw a clear distinction between the concept of collection and that of fonds have led to a promising line of research, with significant implications for archival representation. Yeo views collections as physical aggregations of records brought together either in terms of ownership, custody, or co-location, while arguing for fonds to be understood as conceptual constructions.[48] According to Jennifer Douglas, archival science should not only admit but embrace "the 'constructedness' of the fonds".[49] Those constructing the fonds include not only the original creators of the records contained in the fonds, but also subsequent custodians, archivists, researchers, and users in general. They shape the archive by creating new meaning as they interact with it at different times. To borrow Bearman's words, "the fact of processing, exhibiting, citing, publishing and otherwise managing records become significant to their meaning as records".[50]

I argue that these layers of creatorship should be seen as interacting conceptual

[47] Arianna Ciula and Øyvind Eide, 'Modelling in digital humanities: Signs in context', Publisher: Oxford University Press, *Digital Scholarship in the Humanities* 32 (2017): 34.
[48] Yeo, 'The Conceptual Fonds and the Physical Collection'.
[49] Douglas, 'Archiving Authors: Rethinking the Analysis and Representation of Personal Archives', 237.
[50] David A. Bearman, 'Documenting Documentation', *Archivaria* 34 (1992): 41.

models that we can formalise as knowledge graphs. Indeed, if we look at the conceptual fonds as the product of an intellectual process of archival interpretation, then we can see the conceptual fonds as a conceptual model and the creator of that fonds as a modeller.[51]

1.4 Augmenting Descriptions with Additional Interpretations

If we consider the fonds to be a model and archival representation to be a modelling activity, we can accept multiple fonds for the same collection, just as competing models for the same observed domain can coexist. In the context of the personal archives preserved at the Ezio Raimondi Library, these considerations translate into the possibility of enhancing archival descriptions by incorporating additional layers of scholarly interpretation, possibly even conflicting, into the finding aids as research on the material progresses. In this respect, Light and Hyry maintain that "over time, as reference archivists and researchers use a processed collection, they come to know more about the records or understand them in new ways".[52] They also point out that "the finding aid privileges the first reading of a collection, arresting its evolution at a particular moment in time. Other valid or accumulative readings, which may be useful for discovery and analysis, are lost or recorded in disparate places removed from the collection".[53]

The possibility of incorporating further layers of scholarly interpretation leaves the archival representation open to new knowledge that would integrate the initial description carried out by the archivist—which is frequently partial and incomplete due to limited time and resources on the one hand and the large number of records that most collections contain on the other.

Our current perception of Giuseppe Raimondi's cultural archive has been shaped by different creators: Giuseppe Raimondi himself, who selected and arranged his

[51] This view is also applicable to the original creators, i.e. the individuals who performed the activities that resulted in record aggregation. The original creators too convey their individual perspective on the collection as they operate a more or less conscious selection and organisation of their documents. In this respect, Cook speaks of the "abstract notion of the creator's activity" as something different from "the concrete or physical realities of the resulting records": on the one hand, the domain that the archive is supposed to represent; on the other, just one of the possible models of such a domain (Cook, 'The Concept of the Archival Fonds in the Post-Custodial Era', 28).

[52] Light and Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', 226.

[53] Light and Hyry, 226.

own records; his heirs, who performed a partial reordering of the archive (a process which likely included arbitrary additions to the collection); at least two archivists, who appraised the fonds and represented it according to a traditional archival hierarchy both on paper and electronically in different times; several researchers working with and interpreting the fonds, who revealed numerous meaningful connections within and beyond Giuseppe Raimondi's records.

The researchers' creatorship is the primary focus of this study. The idea is that we can represent each view of the archive or, as we might say in light of the argument pursued by this chapter, each conceptual fonds as a distinct but interconnected graph. The objective is to provide a multiperspective representation of the archive that admits the possibility of multiple and even conflicting conceptual fonds realised in the form of knowledge graphs. The idea of viewing a fonds as one out of many possible representations is not entirely new to archival debate. In an seminal article, Murray and Tillett claimed that descriptions of cultural heritage resources should be seen as "observations organised into graphs".[54]

Embracing this approach opens up promising opportunities. In the case of Raimondi, for example, we foresee the formalisation of philological reconstructions of the author's writing process by connecting together the books (whose highlight and marginalia convey ideas and suggest how other authors influenced Raimondi's work), the notebooks (which contain annotations and drafts), the different versions of his articles, the newspaper clippings of articles he wrote or which inspired him, the letters (both sent and received, with inevitable inter-archival connections), the records of published articles held at different newspaper archives. Such connections are clearly the product of the researcher's subjective interpretation of the archive, which can be viewed as the researcher's conceptual fonds. By bringing together different readings of the archive as graphs, we offer a multiperspective representation that can grow over time and are able to formally tie each assertion to its creator.

Yakel maintains that "archivists should begin to think less in terms of a single, definitive, static arrangement and description process, but rather in terms of continuous, relative, fluid arrangements and descriptions as on-going representational processes".[55] In support of this claim, she argues that representational tools such as finding aids serve more than just access; they also serve as collection management tools. The transition from paper-based to digital representations has resulted

[54] Ronald J. Murray and Barbara B. Tillett, 'Cataloging theory in search of graph theory and other ivory towers', *Information Technology and Libraries* 30, no. 4 (2011): 182.
[55] Yakel, 'Archival representation', 4.

in many representations of archives being revisited and updated in response to new knowledge, possibly after decades of fixity.

In this respect, Light and Hyry describe finding aids as "centres for the accumulation of knowledge about a collection".[56] This concept of the finding aid as a centre of accumulation implies that the finding aid is not static, but rather grows and evolves. They also observe that most finding aids provide a single perspective on the collection they represent and are not designed to support subsequent modifications by archivists and additional interpretations by researchers. The authors emphasise the requirement to make the inherent subjectivity of the finding aid visible to final users and, accordingly, incorporate multiple perspectives into the archival representation. Regardless of the new opportunities provided by digital technologies, their observations from around two decades ago remain relevant to archival representation today. They suggest that archivists and researchers using a collection over time (and space, now that virtual access to high-quality images of archival documents has become a reality) go through "many of the same processes of discovery, interpretation, explanation, valuation, and understanding of those archivists who initially undertook arrangement and description" gaining "new insight into the context and content of the collection, the significance of certain records, and the relationships to other collections".[57] Scholarly users should be able to contribute to the existing finding aid as their use of the archive reveals new information about the records and their interconnections. The idea of opening finding aids to continuous enrichment is also advanced by Yakel. In her article on archival representation, Yakel urges archivists to "think less in terms of a single, definitive, static arrangement and description process, but rather in terms of continuous, relative, fluid arrangements and descriptions as on-going representational processes".[58] She notes that "the creation of each inventory or guide negates, reinforces, extends, or transcends previous artefacts. Thus each new representational artefact contributes to the knowledge base of the repository at the same time it changes it".[59]

Incorporating additional readings of records and collections would also allow the same archival knowledge base to convey multiple, potentially differing perspectives. This is especially important for collections of heterogeneous record types such as the Giuseppe Raimondi Archive and Library, whose description requires

[56] Light and Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', 228.
[57] Light and Hyry, 226.
[58] Yakel, 'Archival representation', 4.
[59] Yakel, 5.

the collaboration of scholars from various disciplines in order to move beyond the record sets to the level of the single items. While urging archival scholars to transform representations into ever-evolving systems that respond to subsequent uses of the archive, Light and Hyry acknowledge that allowing successive modifications in archival representations may jeopardise the authoritativeness of the archival description. They recognise that, in a scenario in which archival descriptions are constantly allowed to grow or be updated, archivists should act as gatekeepers, despite the fact that this would add yet another layer of mediation to the process of archival representation.[60]

In their analysis, which was grounded in a strongly document-oriented archival practice of description (i.e. where the finding aid, whether paper-based or digital, is organised as a unitary, cohesive document), Light and Hyry discuss the beneficial possibility of adding further annotations to existing finding aids. Developing this idea further, the present research views a reading or interpretation of a record within an archive not as a simple annotation – as imagined by Light and Hyry – but rather as a (re)construction of a web of cross-references to other records, collections, or pieces of scholarship, as well as contextual relationships among relevant individuals, families, organisations, places, and events that becomes and integral part of and expands the archival knowledge base.

However, scholars have also recognised that allowing successive modifications in archival representations may jeopardise the authoritativeness of the archival description. Despite this, Light and Hyry argue that archival representations should evolve into ever-evolving systems that respond to subsequent uses of the archive. In this scenario, archivists would act as gatekeepers to ensure that the evolving archival descriptions maintain their authority, despite the addition of subsequent readings.[61]

The envisioned scenario is one in which the finding aid, supplemented with subsequent and diverse readings, evolves into a growing archival knowledge base. Users would be clearly informed about the origins of each reading, and each reading could be viewed separately or at least distinguished from the others. This approach would allow scholars to gain a more nuanced understanding of the archival materials while also ensuring the archival description's ongoing authority.

Light and Hyry discuss the practice of adding "annotations" to existing finding

[60] Light and Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', 229.
[61] Light and Hyry.

aids in their document-oriented archival discourse. They believe that annotations can enhance the value of documents for a community of users, and that annotations should be kept separate from the document itself, giving users a choice in which annotations to view.[62] In contrast, my study views the interpretation of records within an archive as a (re)construction of a web of cross-references to other records, collections, or pieces of scholarship, as well as contextual relationships among relevant individuals, families, organisations, places, and events. This expands the archival knowledge base and becomes an integral part of the interpretation, rather than just an annotation, allowing the finding aid to evolve over time as new intra- and inter-relationships emerge.

1.5 The Finding Aid as an Evolving Representation

The finding aid plays a critical role in the research cycle as it facilitates the organisation and navigation of collections. However, the role of the finding aid can broaden significantly if researchers using, studying, and interacting in general with the archive could supplement the archival representation with their additional interpretations, as Light and Hyry suggest:

> Why not capture and accumulate the experiences of re-readings as well? With Web-based annotations to online finding aids, archivists can exploit their roles as mediators and producers of knowledge to create a powerful tool for description, revision, reference, and research.[63]

According to the authors, the finding aid could serve as "the center for the accumulation of knowledge about a collection", especially when enriched with additional layers of interpretation that reconstruct relationships among records.[64] The addition of these interpretive layers would transform the finding aid into "a powerful tool for description, revision, reference, and research".[65] In this enriched form, the finding aid becomes a map supporting users in the exploration of records through their interconnections.

[62] "Annotations allow documents to grow, respond, and increase in value for a community of users. The annotations are kept separate from the document itself, giving other users choice in which annotation to view" (Light and Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', 226–227).
[63] Light and Hyry, 226.
[64] Light and Hyry, 228.
[65] Light and Hyry, 226.

Users could opt for viewing a traditional representation that presents the fixed arrangement created by the original creator and interpreted by the archivist. Or, users could choose to view alternative arrangements and contextualisations, such as those by a specific researchers or those added in a specific time period. Digital technologies allow the preservation of past organisational structures while enabling the creation of new logical connections. By turning finding aids into an evolving system, each new addition enriching the existing base with new readings can be recorded so as to maintain provenance information. As Yeo puts it, what appears to be needed is a richer framework that does not require us to impose a single set of boundaries, allows us to represent collections as well as fonds, acknowledges that fonds may overlap and may not correspond precisely to any existing collection, and recognises that items or sets of items may move from one collection to another as collections are formed and re-formed over time".[66]

1.6 The Finding Aid and the Edition: A Reciprocal Enhancement?

Archival and textual research are both concerned with the creation of representations. On the one hand, archival research generates representations of archival collections that are realised as finding aids. Textual research, on the other hand, generates representations of texts that are realised as editions. Editing a text equals to mediating between the text and the reader.[67] The same can be claimed about archival representation, where the archivist, like an editor, acts as a mediator between the archive and the user. Textual and archival representations, like their outputs, i.e. the edition and the finding aid, are never neutral; they always reflect the subjective viewpoint of the scholars who created them. Both the editor and the archivist creating a representation of a text or archive acts also as a maker of connections that provide new context and ground for interpretation.[68]

In a provocative paper reflecting on the relationship between digital humanities and textual scholarship, Robinson emphasises that editions convey arguments

[66] Yeo, 'The Conceptual Fonds and the Physical Collection', 71.

[67] Paul Eggert, 'The Archival Impulse and the Editorial Impulse', *Variants*, no. 14 (20 March 2019): 3–22, accessed 14 June 2021, https://doi.org/10.4000/variants.570, http://journals.openedition.org/variants/570.

[68] Dirk Van Hulle used this expression in the context of a lecture about the Samuel Beckett Digital Manuscript Project given at the University of Bologna on December 12, 2019.

about texts. Without editions, digital surrogates of archives are useless for research.[69] This chapter, however, has discussed the finding aid as a research output that conveys an argument about an archival collection, just as an edition conveys an argument about a text. This final section explores the boundary between edition and archive to suggest ways they could dialogue with one another to reciprocally enhance their arguments and representations.

Digital surrogates of archives are distinguished from electronic editions because of their breadth of scope and because they do not usually offer critical introductions, commentaries, or apparatuses.[70] Van Hulle and Nixon, however, observe that "the strict boundary between digital archives and electronic editions is becoming increasingly permeable, resulting in a continuum rather than a dichotomy".[71] They claim that digital tools, such as CollateX, allow archives to be transformed into editions as textual apparatuses can be generated on the fly, without any editorial intervention.[72] In the same wake, Robinson encourages experiments in the direction of a connection between digital libraries and digital editions claiming that "there should be an unbroken continuum between the multiple resources created by libraries as they produce and put online more and more images of the manuscripts and books [...] through to the scholarly editions we might build within and on top of all these materials".[73]

This idea of a direct connection between digital archives and editions is also advanced by Eggert in two different articles.[74] Editions, according to Eggert, could become enriching layers for digital libraries and archives, while libraries and archives could contribute to the development of digital scholarly editions by providing a solid foundation. In the articles, he underlines how in the digital dimension the definition of 'archive' gets increasingly blurred as the term is used to indicate di-

[69] "An edition is an argument about a text. We need arguments; without arguments, our archives are inert bags of words and images" (Peter Robinson, 'Why digital humanists should get out of textual scholarship', Social, Digital, Scholarly Editing Conference (University of Saskatchewan, 11 July 2013), 1).

[70] See Dirk Van Hulle and Mark Nixon, *Samuel Beckett Digital Manuscript Project: Editorial Principles and Practice* (December 2021), accessed 24 June 2021, https://www.beckettarchive.org/editorial.jsp; Kenneth M. Price, 'Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?', *Digital Humanities Quarterly* 3, no. 3 (2009): 11.

[71] Van Hulle and Nixon, *Samuel Beckett Digital Manuscript Project: Editorial Principles and Practice*.

[72] Ronald H. Dekker and Gregor Middell, *CollateX, Version 1.7.1*, in collab. with The Interedition Development Group, 2015, https://collatex.net/.

[73] Robinson, 'Why digital humanists should get out of textual scholarship'.

[74] Paul Eggert, 'The reader-oriented scholarly edition', *Digital Scholarship in the Humanities* 31, no. 4 (December 2016): 797–810, accessed 6 May 2021, https://doi.org/10.1093/llc/fqw043; Eggert, 'The Archival Impulse and the Editorial Impulse'.

gital surrogates of analogue archives, digital-born archives, or thematic collections in general. In these resources, what Eggert calls the 'archival impulse' starts bending towards the 'editorial impulse', meaning that digital libraries and archives, like digital scholarly editions, provide readers with some level of analysis and interpretation of the materials they display.[75] He claims that a digital archive or library that already differentiates between a work and variant versions of that work begins to acquire the characteristics and intent of an edition by providing readers with an initial argument about the resources it contains.[76]

Following this line of thought, the archive and library, which provide facsimiles and transcriptions of books, manuscripts, and other documents, including visual items, can serve as a foundation for the edition, which can be built on top of and within the digital archive. In this vision, the archive and the edition are no longer separate data silos, but rather interconnected resources that form an interdependent system. In a similar system, the edition, according to Eggert, serves as "the editorial layer of the complete project" and "the normal point of entry [to the archive] for the reader".[77] Among the editorial tasks is to establish the connection between the 'editorial layer' and the 'archival evidence'. Each digital scholarly edition built on top of a digital archive, according to Eggert, would be 'an argument directed at the reader about the archive" that "adds value to what is independently available in the archive".[78]

This concept of reciprocal enhancement between digital archives and editions can also be applied to metadata in finding aids. Layers of editorial metadata can indeed be added to archival metadata. The following chapter will discuss how metadata in the form of linked data based on RDF can be easily supplemented with new statements without having to alter the already existing dataset, and how the provenance of each new set of statements can be openly attributed and described with additional metadata. In this vision, any scholar can contribute to the

[75] The distinction between different scholarly resources such as archives, libraries, and editions that was once clear in the analogue dimension becomes more blurred in the digital dimension as archives begin to offer arguments on the materials they encompass, including editorial features. As a result, Eggert proposes visualising "the relationship of archive and edition as a horizontal slider or scroll bar running from archive on the left to edition on the right" (Eggert, 'The Archival Impulse and the Editorial Impulse', 8).

[76] This distinction is consistent with the WEMI model classification of Work, Expression, Manifestation, and Item proposed by IFLA's FRBR conceptualisation, which uses the notions of Work and Expression to identify the abstract intellectual content of a creation and the actual text — seen as a combination — of signs realising such a creation. See Chapter 2 for a thorough discussion of FRBR in the context of archival representation.

[77] Eggert, 'The Archival Impulse and the Editorial Impulse', 16.

[78] Eggert, 'The reader-oriented scholarly edition', 805.

digital archival representation, which, like digital scholarly editions, becomes "always open to correction, revision and extension".[79]

1.7 Concluding Remarks

In light if the issues discussed so far, which technical approach to archival representation should we adopt to take into account, combine, and mutually complement multiple perspectives? It is important to stress that the goal is not just to create a set of knowledge graphs corresponding to distinct representations of the archive. The goal must rather be to support the interlinking of different knowledge graphs so that users can explore different points of view on the archive. The final result should be that of a layered archival representation where each new layer complements, critiques, and extends the layers below.

In this thesis I suggest that knowledge graphs challenge our understanding of archival representation both methodologically and conceptually. Understanding conceptual fonds as conceptual models, and archival creators as modellers, opens up unprecedented opportunities for archival practice. Through the case study of Giuseppe Raimondi's fonds, our project sets out to explore such opportunities. I propose to employ semantic web technologies and linked data to represent the archive as a multi-layered entity, with each layer providing a perspective tied to a specific provenance. The expectation is for this project to be able to show how archival description can evolve not only from the point of view of technology, but also as a theory.

[79] Eggert, 'The Archival Impulse and the Editorial Impulse', 11.

Chapter 2

Organising and Representing Cultural Heritage Knowledge on the Web

The previous chapter examined the challenges of representing personal archives, particularly writers' archives. Drawing upon archival and digital humanities literature, the chapter identified several complexities when attempting to represent these materials, highlighting three key issues with current archival representation standards.

First, archival description standards are not oriented towards the item level but tend to prioritise context, understood as the archival hierarchy into which individual documents nest. Second, current archival representation practices do not focus on content, especially text, even though it is a critical source of information for the (re)construction of context.[1] Third, most archival descriptions privilege a single fixed point of view on collections without allowing the incorporation of additional scholarly readings into the finding aid.

The chapter concluded that focusing on individual documents – or, in Yeo's words, orienting the archival description towards the individual item level –[2] is required in order to chart new paths for the representation of writers' archives, in which text in the broadest sense of the term is the primary research focus.

In keeping with these ideas, this chapter provides an overview of the state of the art concerning the organisation and representation of knowledge in the cultural heritage domain, focusing on a selection of existing models and practices. The chapter discusses the Semantic Web and Linked Open Data as data-driven methodologies –

[1] See Tomasi and Daquino, 'The archival domain in a disciplinary-integrated ontological perspective', 23.

[2] See Yeo, 'Contexts, Original Orders, and Item-Level Orientation'.

not just technologies – for representing archives and cultural heritage resources in general. The chapter then examines existing ontologies and vocabularies in archival, museum, library, and textual studies, emphasising their suitability for representing personal archives. The chapter concludes with an overview of existing digital projects that use linked open data to represent and disseminate personal archives and scholarly knowledge about them.

2.1 The Semantic Web and Linked Open Data

Semantic web technologies and linked open data (LOD) have been investigated for over a decade as a means of publishing and increasing access to cultural heritage collections. The digital humanities and the archival, museum, and library information science have explored using LOD to represent cultural heritage data, such as texts and editions, as interconnected networks.

Compared to the first *Companion to Digital Humanities*, published in 2004, the *New Companion to Digital Humanities*, published in 2015, includes a series of new contributions in this direction, demonstrating how the semantic web had entered the digital humanities discourse and practice during the years that separate the distribution of the second companion from that of the first one.[3] The new companion offers a definition of the concepts of semantic web and linked data that clearly explains how these are related to one another:

> The Semantic Web [...] is the web of meaningful data that can be processed by computers and employs 'Linked Data' as the mechanism for publishing structured data to the World Wide Web where that data can be linked and integrated.[4]

The reader should focus on the attributes 'meaningful' and 'integrated', which describe data in the semantic web context. Such linked data are meaningful because the interconnected networks they form carry semantics that both machines and humans can interpret. Linked data can also be integrated as datasets produced by

[3] See Susan Schreibman, Ray Siemens and John Unsworth, eds., *A companion to digital humanities* (John Wiley & Sons, 2008), https://companions.digitalhumanities.org/DH; Susan Schreibman, Ray Siemens and John Unsworth, eds., *A new companion to digital humanities* (John Wiley & Sons, 2015).

[4] Dominic Oldman, Martin Doerr and Stefan Gradmann, 'Zen and the art of Linked Data: new strategies for a Semantic Web of humanist knowledge', in *A new companion to digital humanities*, ed. Susan Schreibman, Ray Siemens and John Unsworth, Publisher: Wiley Online Library (Wiley, 2015), 251.

different institutions – for example, a museum, a library, and an archive – can be used and queried as one unitary source of information.

Compared to web 2.0, which primarily focuses on linking documents, the semantic web relies on conceptual models and ontologies to formally express relationships between atomic data so that machines can interpret their meaning and reconstruct context.[5] In summary, web 3.0 extends web 2.0 with additional standards that enable the formalisation of how different types of data, such as documents and texts, are related. The possibility of expressing semantic relationships between resources (i.e. relationships that use terms from published ontologies) allows scholars to meaningfully enrich their data with metadata, provenance, and context.

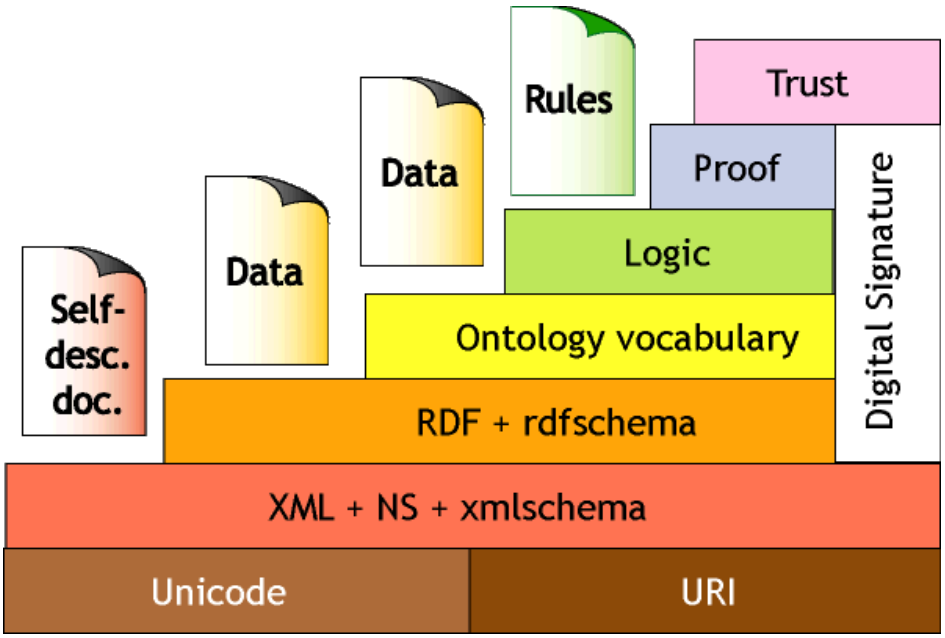

Figure 2.1: The Semantic Web stack of technologies (from https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html).

Figure 2.1 illustrates the architecture of the semantic web, also known as the semantic web stack, a hierarchy of technologies in which each layer relies on and uses the capabilities of the layers below.[6] The bottom layers contain well-known hypertext web technologies that serve as the foundation for the semantic web.

The standard method for encoding HTML documents on the web is UTF-8, a

[5] To link web pages, web 2.0 (also known as the web of documents) uses the 'href' attribute that indicates a generic hyper-reference, which produces an implicit connection between the documents. In web 3.0 (also known as the web of data), the meaning of links is made explicit by leveraging terms from ontologies, enabling individual entities described in documents to be interconnected via typed links. For example, instead of defining a generic hyperlink from one document to another, web 3.0 allows specifying that some person described in one of the documents is the son of a person described in another document.

[6] Tim Berners-Lee, 'Semantic Web on XML' (XML 2000 Washington DC, 6 December 2000), accessed 24 April 2023, http://www.w3.org/2000/Talks/1206-xml2k-tbl.

variant of Unicode,[7] the universal character encoding standard that establishes codes for the characters that make up all of the most widely used modern writing systems. Adopting a global standard for character encoding ensures that characters are correctly exchanged, processed, and displayed across different languages and countries.

Uniform Resource Identifiers (URI) are unique strings of characters that identify resources unambiguously (every piece of data on the semantic web, including relationships, is represented by means of a URI).

XML allows the creation of semi-structured documents organised into data hierarchies.[8]. Because XML is a metalanguage (i.e. a language to create markup languages), different XML applications may employ different elements and attributes. As such, when using multiple XML vocabularies together in a single document, XML namespaces must be used to prevent name conflicts. XML schemas define the structure, elements, attributes, and any other rules that must be followed by XML documents using such schemas.[9]

These lower layers serve as the foundation for the middle layers of the semantic web stack.[10] The Resource Description Framework (RDF), a data model for describing information about resources as graphs of statements formed by a subject, a predicate, and an object, constitutes one of these layers. RDF can be serialised in XML, but other commonly used formats are available.[11]

RDF Schema (RDFS), an extension of the basic RDF vocabulary, provides the basic components for developing vocabularies used in RDF triple sets, while the Web Ontology Language (OWL) is the main language for defining ontologies. Because it is based on description logic, OWL extends RDFS allowing for more expressive representations and more powerful reasoning over knowledge bases. The Simple Protocol and RDF Query Language (SPARQL) is the query language for RDF data. The higher layers of the semantic web stack shown in Figure 2.1 are yet to be implemented.

[7] See Richard Ishida, *Who uses Unicode?* (17 March 2023), https://www.w3.org/International/questions/qa-who-uses-unicode.

[8] Texts encoded according to the TEI standard are an example of semi-structured XML documents.

[9] The TEI, for example, is implemented as a collection of XML schemas, either predefined or user-created.

[10] Tim Berners-Lee titled one of his first presentations of the semantic web architecture "Semantic Web *on* XML" (the italics are mine) to emphasise that the semantic web is built on XML, extending the existing web architecture rather than replacing it.

[11] Other popular serialisation formats for RDF are Turtle, N3, and JSON-LD. See W3C, 'RDF Syntax', 31 January 2023, accessed 24 April 2023, https://www.w3.org/wiki/RdfSyntax.

To help frame this research, the following section delves deeper into three critical semantic web concepts: the RDF data model, the role of ontologies, and the principles of linked open data.

2.1.1 The RDF Data Model

RDF provides a framework for structuring information on the web as graphs of statements composed of a subject, a predicate, and an object (also called RDF triples). A statement's subject or object can also be another statement's subject or object. When this happens, separate statements connect to form a graph.[12] Unique URIs represent each of the three components of a statement. Listing 2.1 below shows an RDF statement about one of Giuseppe Raimondi's manuscript notebooks. The statement asserts that the notebook (S) is currently located at (P) the Ezio Raimondi Library (O):

```
S: <https://w3id.org/giuseppe-raimondi-lod/notebook/rdq1/object> # Giuseppe Raimondi's notebook RDq1
P: <http://erlangen-crm.org/current/P55_has_current_location> # has current location
O: <https://w3id.org/ficlitdl/place/biblioteca-ezio-raimondi> # Ezio Raimondi Library, Bologna
```

Listing 2.1: Example of an RDF statement formed by a subject, a predicate, and an object.

The same statement could be made for all items currently located at the Ezio Raimondi Library. If that were the case, a graph containing as many statements as the total number of items in the library would emerge. By navigating such a graph, users could easily find other objects whose current location is the Ezio Raimondi Library.

As previously stated, RDF is a data model rather than a language, and it can be expressed using a variety of syntaxes such as XML (the standard format), Turtle, N3, JSON-LD, or N-triples. How XML is used to represent RDF data differs significantly from how the same language is used for annotating texts. XML can markup data according to different degrees of structuredness, from semi-structured to highly structured. Because RDF enforces a data-centric way of describing information, an RDF dataset serialised in XML contains highly structured data. XML, however, can also be used in a narrative- or document-centred way, as in TEI documents. In this case,

[12] See, as an example, 'RDF 1.1 Primer', Figure 1: Informal graph of the sample triples, accessed 27 April 2023, https://dvcs.w3.org/hg/rdf/raw-file/tip/rdf-primer/Overview.html.

XML elements are interspersed within free text documents, making manipulation and interchange more challenging.

An XML schema is not required when creating RDF datasets in the form of XML documents because RDF defines the data structure. However, vocabularies and ontologies are needed to determine the semantics of the data. The original purposes of XML schemas and ontologies are very different. XML schemas provide integrity constraints for semi-structured documents, while ontologies use descriptive logic to specify more complex domain theories.[13] An XML schema defines the syntactic structure of an XML document but is unaware of its meaning.[14]

In some cases, a triple is insufficient to describe information. The strength of RDF is that it allows anything to become the subject or object of a statement. As a result, it is always possible to supplement the existing description with additional metadata.[15] If the subject or object of a statement is another statement rather than an individual entity, the statement as a whole can also be given a unique identifier so that it can later be referred to. RDF quadruples can be used to accomplish this. An RDF quadruple is constituted by a subject, a predicate, an object, and the name of the graph to which the triple belongs. N-Quads and TriG are the syntaxes for writing RDF quadruples.[16]

2.1.2 Vocabularies and Ontologies

If the RDF data model determines the structure of RDF data, vocabularies and ontologies define their semantics. Vocabularies are controlled lists of terms, which are organised hierarchically and by association.[17] Ontology and vocabulary are terms that are frequently used interchangeably. However, because ontologies allow for the formal expression of interrelationships, axioms, and classes using logical constructs, they enable the creation of more complex representations of the domains they reflect.

[13] Michel Klein et al., *The relation between ontologies and XML schemas*, 2001.

[14] Silvio Peroni, Aldo Gangemi and Fabio Vitali, 'Dealing with markup semantics', in *Proceedings of the 7th International Conference on Semantic Systems* (2011), 111–118.

[15] In the context of linked data, distinguishing between data and metadata is hardly possible because metadata can be further described by additional metadata in a scenario of increasing complexity.

[16] Because the quadruple construct allows for attributing a name to one or more triples, RDF quadruples enable the creation of so-called named graphs.

[17] A data model for representing various types of controlled vocabularies, such as thesauri, classification schemes, subject heading lists, and taxonomies, on the web is offered by the Simple Knowledge Organisation System (SKOS).

An ontology is an explicit formalisation of a conceptualisation capable of capturing "the objects, concepts, and other entities that are assumed to exist in some area of interest, and the relationships that hold among them".[18] Ontologies, according to Gruber, are closer to first-order logic than other existing data models because the languages in which they are expressed allow for a high level of abstraction. For this reason, ontologies are frequently referred to as conceptual models[19].

2.1.3 The Principles of Linked Open Data

Linked Open Data (LOD) refers to a set of evolving best practices for publishing structured data on the web rather than a specific standard or technology.[20] These data are primarily intended for machine consumption and are a key feature of the Semantic Web, which Bizer et al. describe as "a web of things in the world, described by data on the web".[21] Berners-Lee identifies four rules for creating linked data on the web:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs. so that they can discover more things.[22]

Adherence to these principles allows the creation of interconnected datasets that users can exploit in new, unforeseen ways. As Berners-Lee indeed underlines, "it is the unexpected re-use of information which is the value added by the web".[23] In the context of cultural heritage, when datasets produced and published independently by different institutions cease to be siloed and become interconnected to generate new shared knowledge, the value of such datasets increases disproportionately. The

[18] Thomas R. Gruber, 'Toward principles for the design of ontologies used for knowledge sharing?', *International Journal of Human-Computer Studies* 43, no. 5 (1 November 1995): 907, accessed 29 March 2023, https://doi.org/10.1006/ijhc.1995.1081, https://www.sciencedirect.com/science/article/pii/S1071581985710816.

[19] See Julia Flanders and Fotis Jannidis, *Knowledge organization and data modeling in the humanities* (2015), https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-111270.

[20] For a critical introduction to linked open data and their use in the library, archival, and museum domains see Hooland and Verborgh, *Linked Data for Libraries, Archives and Museums*.

[21] 207 Christian Bizer, Tom Heath and Tim Berners-Lee, 'Linked data: The story so far', in *Semantic services, interoperability and web applications: emerging concepts* (IGI global, 2011), 205–227.

[22] Tim Berners-Lee, 'Design issues: Linked data', *http://www.w3.org/DesignIssues/LinkedData*, 2006,

[23] Berners-Lee.

section that follows will discuss the benefits of using linked open data to describe cultural heritage collections before reviewing various datasets and ontologies for knowledge organisation and representation in the library, archive, and museum domains, with the goal of identifying gaps and opportunities in existing models for the formal representation of personal archives.

2.2 Some Models For Cultural Heritage Description Across the GLAM Domain

GLAM is an abbreviation for galleries, libraries, archives, and museums. These organisations are in charge of gathering, preserving, and disseminating cultural heritage materials such as artworks, books, manuscripts, photographs, and other types of artefacts, as well as knowledge about these materials. GLAM institutions are beginning to explore the added value of sharing resources as linked open data.[24] As Marden et al. explain:

> Linked open data offers a new way for cultural heritage institutions to share their holdings with a wider audience, and to change the traditional relationship between the holder of knowledge, the interpreter of knowledge, and the consumer of knowledge. With a strong user interface built upon a linked open dataset, users with all levels of expertise can access and analyze information once siloed in many different LAMs. This new way to interpret and access cultural heritage information might allow us to update how we define cultural heritage itself.[25]

Interconnecting data with other datasets from different institutions is fundamental for improving user experience in exploring cultural heritage information. Adopting an integrated approach is especially critical when dealing with personal archives. Personal archives comprise a wide range of items, including papers, books, photographs, drawings, artwork, and various other objects, such as writing tools, in line with the archival principle that any item possessing evidential value of the life and activities of its original owner or creator should be treated as an archival document. More than just focusing on the archival domain is required when representing personal archives. Instead, attention must be directed towards the broader

[24] Examples of GLAM linked open datasets are the Library of Congress subject headings (https://id.loc.gov/authorities/subjects.html); the Europeana Portal (https://www.europeana.eu/en); the Sistema Archivistico Nazionale Linked Open Data (http://dati.san.beniculturali.it/dataset).

[25] Jeff Edelstein et al., 'Linked Open Data for Cultural Heritage', 2013, 6.

cultural heritage representation landscape, including museums, libraries, and text representation models, to fully and effectively represent the variety and complexities characterising personal archives.

The upcoming section delves into a selection of various models used to represent cultural heritage resources and knowledge. It starts by exploring archives and gradually moves on to museums and libraries. Additionally, it will introduce other models that are useful for representing or formalising specific aspects of personal archives.

2.2.1 The Records-in-Context Conceptual Model and Ontology (RiC-CM and RiC-O)

According to Hooland and Verborgh, a proper understanding of the advantages offered by RDF, the data model underlying the linked data approach, can only be achieved by considering the historical progression of previous data models.[26] This section traces the evolution of the International Council on Archives (ICA)'s archival description standards, from ISAD(G) and related standards to RiC-CM and RiC-O, a conceptual model and ontology created by the ICA that aims to be a standard for producing interconnected archival descriptions based on the linked data paradigm.

The General International Standard Archival Description (ISAD(G)), the International Standard Archival Authority Records — Corporate Bodies, Persons, and Families (ISAAR(CPF)), the International Standard for Describing Functions (ISDF), and the International Standard for Describing Institutions with Archival Holdings (ISDIAH) were the four international standards for archival description developed by the ICA "to be used in conjunction with existing national standards or as the basis for the development of national standards".[27].

In the late 1990s, the Society of American Archivists was pivotal in developing two XML standards: the Encoding Archival Description (EAD) and Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF), both mapped to ICA's standards. EAD is mainly used for encoding information related to archival materials, while EAC-CPF is used for encoding information about the creators of the materials.

[26] Hooland and Verborgh, *Linked Data for Libraries, Archives and Museums*, 14.

[27] Internationaler Archivrat and International Council of Archives, eds., *ISAD(G): General international standard archival description; adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 Sept. 1999*, 2. Ed, ICA Standards (Ottawa: International Council of Archives, 2000), 7.

The development of EAD and EAC-CPF represented a crucial first step towards enabling dialogue and integration between archival descriptions. These standards are defined by XML schemas and vocabularies that allow for syntactic interchangeability of finding aids produced as XML documents. However, as noted by Yakel, the technological shift from paper-based to XML finding aids does not also necessarily signify a shift in paradigm. XML/EAD archival descriptions, being based on previously existing paper-based finding aids, are document-centred. In these finding aids, XML is leveraged in a semi-structured way which makes reuse of and interoperability between archival descriptions a complex task.[28]

In more recent years, since 2012, the Experts Group on Archival Description committee (EGAD), established by the ICA Programme Commission, has been actively working on developing a new standard realised as a conceptual model and formalised as an ontology that could overcome the shortcomings of XML.

RiC, a standard for archival description, comprises two distinct components: a conceptual model (RiC-CM) and an ontology (RiC-O). The primary objective of RiC is to consolidate the four existing ICA standards — ISAD(G), ISAAR(CPF), ISDF, and ISDIAH — into a unified model. In addition, RiC-O intends to provide an ontological framework for publishing archival descriptions as linked open data on the web and aims to facilitate integration between information systems in GLAM institutions.

In 2017, the ICA decided to develop an ontology based on RiC-CM, named RiC-O, first released in December 2019. RiC-O implements the entities and relationships defined in RiC-CM into a formal ontology, enabling the semantic representation of archival information.

Despite the effort put forth in developing a standard ontology for the archival domain, the decision to create RiC-O as an isolated model without first considering other systems of description used in libraries, museums, and other cultural domains is a critical issue if ICA intends for RiC-O to support dialogue and integration between archival representations across different cultural institutions around the world.

Long before the ICA, beginning in 2000, the International Council of Museums (ICOM) defined the CIDOC Conceptual Reference Model (CIDOC-CRM), as an ontology for cultural heritage information. The International Federation of Library Associations (IFLA) expanded CIDOC-CRM in 2006 by reformulating FRBR, an entity-

[28] See Yakel, 'Archival representation', 5.

relationship model intended to represent bibliographic information, as a conceptual model. The decision to extend CIDOC-CRM rather than create a new ontology was made to facilitate the integration, mediation, and interchange of bibliographic and museum information.

In the future, the RiC-O documentation mentions mapping their model to existing ontologies in the cultural heritage domain (although there is yet to be a detailed plan to do so). Even so, the fact that it was created without leveraging existing models widely used by numerous cultural institutions as a foundation makes completing a full mapping problematic.

Although at the very beginning of the RiC-CM documentation, it is claimed that the model "may be used to support scholarly descriptions of individual records or sets of records that are based on other perspectives and requirements"[29], the model does not concern itself with the multiperspective representation of archives.[30] The most recent version of RiC-O mentions the option of using named graphs to describe provenance information. Still, it does not make the issue of describing provenance central to the scope and objectives of the model despite provenance being one of the cornerstones of archival science and practice. As the Interpares Trust points out:

> RiC-CM appears totally unaware of the international debate – raised by postmodernist voices and then discussed by the broader community – on the role of archivists as mediators, and on the value of their professional action, supporting and guaranteeing the authenticity of records on the one side, shaping the cultural memory hence the identity of communities on the other side. In other words, it seems impossible to define a model of descriptive elements if a model of the archivist's role is not defined. Instead, EGAD considers this as an action that can be postponed.
>
> It bears stating also that description is not always the work of archivists. Increasingly, in interactive online environments users are asked to contribute and descriptions need to accommodate multiple perspectives. When two or more users (or indeed two or more archivists) examine the same entity, they are likely to view it in different ways and see different relationships between it and other entities. But there appears to be nothing in the standard that would support representation of different

[29] *Records in Contexts Conceptual Model*, Consultation Draft v0.2 (International Council on Archives Expert Group on Archival Description, 2021), 4.
[30] *Comments on "Records in Context"* (InterPARES Trust, 10 December 2016), 7.

> and possibly conflicting viewpoints; indeed, EGAD does not reflect the movement toward user participation.[31]

The reason for the isolation of RiC-O from existing models and ontologies and for not taking into account the most recent archival thinking, as pinpointed by Feliciati and Duranti, appears groundless.[32] RiC-O is set to remain isolated in the global effort to interconnect heterogeneous cultural heritage data until it is aligned with CIDOC-CRM and other widely used ontologies.

In light of what has been said above, RiC-O is insufficient for representing the personal archives of writers. More effective methods for describing parallel provenance in archival descriptions are required to allow multiple scholars to contribute to enriching finding aids over time, as well as to bring archival description standards in line with the most recent archival thinking, which acknowledges the role of the archivist (and any other agent contributing to the archival description) as a mediating role and the nature of the finding aid as a subjective model of a collection.

Besides, as Tomasi points out, the highly heterogeneous nature of the documents contained in personal archives necessitates a harmonised and integrated use of multiple conceptual models and the adoption of different disciplinary perspectives, in contrast with the self-referential approach which EGAD opted for in developing RiC-O. Such isolation of the model is particularly problematic when creating item-level representations of personal archives.[33]

The primary focus of ICA's standards is on the representation of collections at the level of the (sub)fonds, the (sub)series, and the (sub)file. Representing individual items below the file level requires adopting other modelling approaches due to the variety of record types (photos, notes, letters, newspaper clippings, excerpts, books, periodicals, artefacts, etc.) that a personal archive may contain. This heterogeneity can make describing and representing a single personal archive as complex and challenging as describing cultural heritage materials preserved and maintained across different GLAM institutions.

Attempting to reduce the descriptions of personal archives to a single model would inevitably result in a loss of the expressive capacity of each document.[34] Sim-

[31] *Comments on "Records in Context"*, 5.

[32] See Pierluigi Feliciati, 'Archives in a Graph : the Records in Contexts Ontology within the framework of standards and practices of Archival Description', *JLIS*, no. 1 (2021): 100, accessed 2 May 2023, https://doi.org/10.4403/jlis.it-12675, https://doi.org/10.4403/jlis.it-12675.

[33] Tomasi, 'Archivi di persona in linked open data', 285–286.

[34] Tomasi, 285–286.

ilarly, Bowker and Star argue that each standard naturally and unavoidably privileges a particular perspective while marginalising others.[35] Considering this, the ability to combine multiple models for describing complex systems such as personal archives is extremely valuable.

Examples of existing models created and used in the context of different disciplines and domains that provide relevant solutions for the representation of personal archives are:

- CIDOC-CRM, which was born in the context of museums but has expanded to become a cross-domain ontology for the harmonisation of cultural heritage datasets;
- FRBRoo, which is used in the bibliographic domain but can also handle text and manuscript representation;
- the Open Annotation (OA) data model, which allows the description of annotations to documents on the web but can also be leveraged for representing annotations on documents in general (e.g. manuscript annotations added to a physical book by its owner);
- PROV-O, a general model that can be employed to represent provenance information related to archival descriptions;
- PRO, an ontology developed in the context of the publishing domain to describe the different agents and roles involved with the life-cycle of a cultural resource;
- the TEI which is the de facto standard in the humanities for the markup of texts and the production of digital scholarly editions;
- the Nanopublication data model, from the biology domain, which allows the creation of representations composed of multiple named graphs, one for each agent or time when new assertions are added to the archival knowledge base.

[35] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge, MA: MIT press, 1999), 5.

2.2.2 The CIDOC Conceptual Reference Model (CIDOC-CRM)

The CIDOC Conceptual Reference Model (CIDOC-CRM) provides a framework for describing the concept and relationships typically found in cultural heritage documentation. Doerr and Crofts point out that "the primary role of the CRM is to serve as a basis for mediation of cultural heritage information and thereby provide the semantic 'glue' needed to transform today's disparate, localised information sources into a coherent and valuable global resource".[36]

CIDOC-CRM is an event-based model. CIDOC-CRM events are time-bound temporal entities in which a group of persistent physical or conceptual entities participate. The interaction between persistent and temporal entities constitutes the core of the CIDOC-CRM model. For example, using CIDOC-CRM, it is possible to represent the creation of a text as an event or, more specifically, an activity involving and interconnecting four persistent entities (a text, an author, a place, a time) and one temporal entity (the text creation).[37]

CIDOC-CRM entities all originate from the highest level class, E1 Entity, which acts as the root of the CIDOC-CRM poly-hierarchy. This poly-hierarchy structure is advantageous in that an entity can have multiple parent classes, allowing for the representation of complex phenomena, such as textual variation or the conceptual belonging of an individual item to multiple record sets. These phenomena often generate a problem of overlap, where an entity can be classified in different ways simultaneously depending on the context. Multi-parentage, which enables entities to have multiple parents, addresses this problem. This technique allows entities to be classified in various ways based on the requirements of the specific application or use case.[38]

In CIDOC-CRM, conceptual objects (belonging to the E28 Conceptual Object class) are further specialised into propositional objects. Propositional objects are "immaterial items, [...] sets of propositions about real or mental things and that are

[36] Martin Doerr and Nicholas Crofts, 'Electronic Communication on Diverse Data - The Role of the oo CIDOC Reference Model', 1998, 2.

[37] This construct is based on the n-ary relation pattern, which allows expressing non-binary relationships that involve multiple entities. The pattern suggests that the relationship is 'reified', i.e. represented as a class rather than a property, to which the multiple entities involved are linked.see Natasha Noy and Alan Rector, *Defining N-ary Relations on the Semantic Web*, W3C Working Group Note 12 April 2006 (2006), accessed 3 May 2023, https://www.w3.org/TR/swbp-n-aryRelations/.

[38] See C. Sperberg-McQueen et al., 'Drawing inferences on the basis of markup', 1 January 2002,

documented as single units or serve as topic of discourse. This class also comprises items that are 'about' something in the sense of a subject".[39] For example, a FRBRoo work, as the next section will introduce, is a type of propositional objects (i.e. a subclass of the E89 Propositional Object class in the CIDOC-CRM). Here, the concepts defined in FRBRoo intersect with those belonging to CIDOC-CRM, providing an integrated conceptual model for describing cultural heritage.

As the TEI is the standard for text encoding in the humanities, the TEI Ontologies Special Interest Group's proposal to align the TEI model with CIDOC-CRM is a logical next step for the two models. A mutual exchange of data and information could greatly benefit the fields of textual, archival, museum, and cultural heritage research in general.[40]

2.2.3 The Functional Requirements for Bibliographic Records object-oriented (FRBRoo)

FRBRoo, soon to be renamed LRMoo (the Library Reference Model), is an object-oriented model designed to represent the concepts and relationships that characterise the bibliographic domain. The model uses CIDOC-CRM as a foundation, promoting dialogue and integration between library and museum resources.

The central concepts of FRBRoo are Works, Expressions, Manifestations, and Items. Works are abstract concepts belonging to the FRBRoo class F1 Work and are the direct outcome of an intellectual process of creation performed by one or more agents, such as an author or a cultural institution. Expressions are closely related to works and represent the realisation of works in texts, poems, music scores, etc. Any change to an expression results in a new expression being created. Evidence of the existence of a work can only be provided through the existence of its expression. Two distinct classes in FRBRoo describe manifestations: F3 Manifestation Product Type and F4 Manifestation Singleton. F3 represents abstract entities that group all instances of a particular publication, whereas F4 represents human-made physical objects, such as manuscripts, that carry an instance of F2 Expression. The F5 Item class defines items as physical objects that carry a publication expression and are

[39] See the scope note of the E89 Propositional Object class in Nick Crofts et al., eds., 'CIDOC CRM: CIDOC Conceptual Reference Model', November 2011, http://www.cidoc-crm.org/cidoc-crm/.

[40] On the topic of TEI and CIDOC-CRM integration, see Christian-Emil Ore and Øyvind Eide, 'TEI and cultural heritage ontologies: Exchange of information?', *Literary and Linguistic Computing* 24, no. 2 (1 June 2009): 161–172, accessed 28 March 2023, https://doi.org/10.1093/llc/fqp010.

produced by industrial means.

OWL implementations of both CIDOC-CRM and FRBRoo are provided in the framework of the Erlange project.[41]

CIDOC-CRM and FRBRoo have already been adopted in archival representation projects as base ontologies. For example, Bountouri and Gergatsoulis show how the Encoded Archival Description, which is based on ISAD(G) like is RIC-CM, can be mapped to CIDOC-CRM;[42] Daquino and Tomasi demonstrate how CIDOC-CRM and FRBRoo can be leveraged as a basis for representing the Federico Zeri's photo archive;[43] Koch et al. explain how they have adopted and built upon CIDOC-CRM to produce linked open descriptions of the Portuguese National Archives.[44]

2.2.4 The Web Annotation Ontology (OA)

The Open Annotation model, whose name changed in 2013 to the Web Annotation model, aims to provide a standard framework for the description of annotations.

An OA annotation is a "rooted, directed graph that represents a relationship between resources".[45] Such resources are the body of the annotation (related to an oa:Annotation through the property *oa:hasBody*) and the resource being annotated (related to the oa:Annotation through the property *oa:hasTarget*). The body and the target are the primary resources participating in the relationship. Typically, the information conveyed by the body is information about the target. A straightforward example of annotation is a comment about a particular paragraph of an academic article.

Each annotation is associated with a creator (as a rule, this is achieved via the property *dcterms:creator*).[46] The motivation for creating the annotation can also be included in the graph by linking an oa:Motivation to an oa:Annotation via the property *oa:motivatedBy*. In addition, it is also possible to declare the date and the time of

[41] The Erlangen CIDOC-CRM and FRBRoo OWL implementations are available at http://erlangen-crm.org. The official creators of the models, namely ICA and IFLA, do not provide OWL versions of their conceptual models.

[42] Bountouri and Gergatsoulis, 'Mapping encoded archival description to CIDOC CRM'.

[43] Marilena Daquino and Francesca Tomasi, 'Linked Cultural Objects: dagli standard di catalogazione ai modelli per il web of data. Spunti di riflessione dalla Fototeca Zeri', Number: 1, *Umanistica Digitale*, no. 1 (1 October 2017), accessed 10 February 2021, https://doi.org/10.6092/issn.2532-8816/7195.

[44] Koch et al., 'Knowledge graph implementation of archival descriptions through CIDOC-CRM'.

[45] Robert Sanderson, Paolo Ciccarese and Benjamin Young, *Web annotation data model. W3C Recommendation*, 2017-02-23, http://web.archive.org/save/https://www.w3.org/TR/annotation-model/ (2017).

[46] DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms/.

the creation or modification of an oa:Annotation via the properties dcterms:created and dcterms:modified.

The body of an annotation can take several forms, including plain text, HTML, XML, RDF. The format of the body can be declared using *dcterms:format*, with values taken from a controlled list of Media Types.[47]

2.2.5 The Provenance Ontology (PROV-O)

Provenance is broadly defined as information about the physical or conceptual origins of an entity. A provenance record describes the agents and activities that were involved in the creation of such an entity. Provenance is a crucial instrument in both archival representation and academic research in general. In a digital environment, where resources form a network of often contradictory or even questionable information, identifying the entities responsible for specific statements becomes even more critical.

The PROV model provides a convenient framework for representing provenance in archival description. The three main PROV-O classes for describing provenance information are prov:Entity, prov:Activity, and prov:Agent. A prov:Entity is "is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary".[48] An example of prov:Entity is a scholarly statement about a collection. A prov:Activity is "something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities".[49] Archival description is an example of prov:Activity. An Agent is "something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity".[50] The archivist generating a record description on the basis of a process of analysis is an example of prov:Agent. An entity is related to the agent responsible for its production through the property *prov:wasAttributedTo*, while an activity is related to the agent who carried it out through the property *prov:wasAssociatedWith*.

In PROV-O, the same activity can generate multiple entities. The connection between an entity and the activity that produced it is expressed through the prop-

[47] IANA Media Types, https://www.iana.org/assignments/media-types/media-types.xhtml.
[48] Luc Moreau and Paolo Missier, *PROV-DM: The PROV Data Model*, W3C Recommendation 30 April 2013 (2013), accessed 3 May 2023, https://www.w3.org/TR/2013/REC-prov-dm-20130430/.
[49] Moreau and Missier.
[50] Moreau and Missier.

erty *prov:wasGeneratedBy*. An entity can be a derivative of (prov:wasDerivedFrom) another entity. Derivation is described in PROV-O as "a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity".[51] By employing the property *prov:wasDerivedFrom* we can build provenance chains, i.e. sequences of entities deriving from one another.

The PROV-O relationship *prov:wasDerivedFrom* is a subproperty of a more generic relationship *prov:wasInfluencedBy*. Influence is defined within PROV-O as "the capacity of an entity, activity, or agent to have an effect on the character, development, or behavior of another by means of usage, start, end, generation, invalidation, communication, derivation, attribution, association, or delegation".[52]

2.2.6 The Publishing Roles Ontology (PRO)

The Publishing Roles Ontology (PRO) is an ontology for representing the roles of agents in the publication process, including people, corporate bodies, and computational agents. Examples of roles are authors, editors, reviewers, publishers, and librarians.[53] The expressive power of this ontology is that it leverages the n-ary relation pattern to represent roles as time-bound concepts. PRO introduces the concept of Role In Time to represent roles in a specific context. Each instance of Role In Time can be linked to a specific document, agent, role, and time that characterise it. Using PRO, for example, a person has the role of author only in relation to a specific authored object and/or in a specific time-frame, allowing for an accurate representation of roles and agency. Adopting models like PRO becomes even more important in the context of archival representation, where provenance is a fundamental principle and multiple provenances are possible. By contextualising roles, PRO can help preserve the integrity of archival descriptions and ensure that the provenance of both records and descriptions is accurately represented.

[51] Moreau and Missier, *PROV-DM: The PROV Data Model*.
[52] Moreau and Missier.
[53] PRO – The Publishing Roles Ontology, http://purl.org/spar/pro. See also Silvio Peroni, David Shotton and Fabio Vitali, 'Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents', in *Proceedings of the 8th International Conference on Semantic Systems* (ACM, 2012), 9–16.

2.2.7 The Text Encoding Initiative (TEI)

TEI is the only model presented among the relevant models for archival representation in this chapter that is not based on the linked open data paradigm but is realised as a set of XML schemas that, as already mentioned in the previous sections, allows encoding texts in a semi-structured way.[54] TEI schemas do not provide XML/TEI documents with formal semantics as an ontology would. Still, it is possible to leverage TEI markup to automatically extract information about the text to represent such information as RDF triples. Using this approach, the information contained in texts (e.g. information about text revision and variation or named entities occurring in the text, such as persons, places, and events) could contribute to the archival knowledge base as foreseen by Tomasi and Daquino.[55] Linked data from TEI (LIFT) is a tool which I developed precisely for this scope (see Chapter 6 for a more detailed discussion of the application in the context of the representation of writers' archives).[56]

2.2.8 The Nanopublication Data Model (NP)

Nanopublications are a linked data format for publishing scholarly data that has received considerable uptake in the last few years. Nanopublications provide a method to attach provenance and metadata to information at an atomic level (for example, a single RDF triple). As illustrated in Figure 2.2, a nanopublication is formed of four distinct graphs: the graph of statements that contains, in the archival representation context, the assertion(s) made by the archivist or scholar; the graph of provenance that contains the provenance of the statements; the graph of publishing information, which describes the provenance of the nanopublication itself; and a graph that has the function of interlinking these three components of nanopublications listed above.[57] Nanopublications, as Chapter 4 will discuss, offers an effective way of creating archival representations that incorporate multiple voices and perspectives into a single knowledge base about the collection.

[54] TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 4.5.0, https://web.archive.org/save/https://tei-c.org/guidelines/ (TEI Consortium, 2019), http://www.tei-c.org/Guidelines/P5/.

[55] Tomasi and Daquino, 'The archival domain in a disciplinary-integrated ontological perspective'.

[56] See also Francesca Giovannetti and Francesca Tomasi, 'Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation', *Digital Humanities Quarterly* 16, no. 2 (2022).

[57] See Paul Groth, Andrew Gibson and Jan Velterop, 'The anatomy of a nanopublication', *Information Services & Use* 30, no. 1 (2010): 51–56.

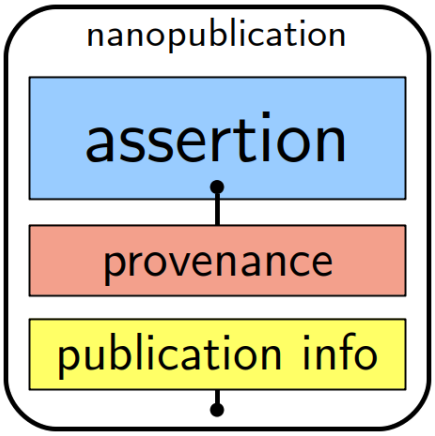

Figure 2.2: The anatomy of a nanopublication (from https://nanopub.net/).

2.3 Concluding Summary

The review presented in this chapter has shown that it is necessary to look beyond archival representation models and towards representation models in the general GLAM domains in order to define a model for representing heterogeneous writers' archives at a granular level, i.e. that of the individual document.

By adopting a model that enables a data-centric, item-oriented archival description, researchers other than the archivist could contribute to extending and deepening the archival description over time. The sole archivist's effort and knowledge is insufficient given the variety of materials found in such archives and the complex relationships between them that need to be formalised in order to provide final users with finding aids that are not only search but research tools that allow them to explore collections according to multiple perspectives and entry points.

Finding aids should implement multidimensional and multiperspective representations of individual records and record parts, including the representation of content and text. In order to accomplish that, archival description standards must be integrated with further models capable of filling the representation gaps identified both at the individual document and at the content level. Moreover, a methodological and, more importantly, conceptual shift from document-centred to data-centred finding aids – that is a shift from creating finding aids as fixed documents rather that as fluid collections of atomic archival data – is required.

By combining and extending existing ontologies drawn from various domains, we can define a new model for personal archives that can help us rethink the finding aid as a complex network of multiple and interacting scholarly contributions

providing different perspectives on the materials, rather than as a close-fixed document.

Experiments with CIDOC-CRM as a base ontology for representing archival information have already proven fruitful for demonstrating how CIDOC-CRM classes and properties could also be leveraged for the archival domain. Given these premises, the model for representing personal archives proposed in this study could be based on CIDOC-CRM and FRBRoo, which provide harmonised classes and properties suitable for representing museum, library, and textual entities as one integrated knowledge base, while also allowing for a high degree of flexibility in organising, manipulating, and describing archival resources.

Chapter 3

Research Methodology

First, it restates the research problem and questions this study seeks to address. It then describes the research hypotheses, which were derived from a review of the archival literature and the state of the art in conceptual modelling for archival representation, both of which have already been discussed in Chapters 1 and 2. The chapter continues by outlining the objectives and contributions of this research before delving into the description of the research design used to conduct the study.

3.1 Framing the Research

Research Problem and Questions

This research employs Giuseppe Raimondi's personal archive as a case study to investigate how linked open data and semantic web technologies can prompt a reconsideration of how writers' archives, and specifically those held and curated by the FICLIT Department, are represented. Through this study, I aim to reflect on the implications and opportunities for cultural institutions of using knowledge graphs as the primary means of archival representation. I specifically explore the possibility of supplementing the finding aid with additional interpretations by subsequent researchers after the archivist has completed the initial process of arrangement and description. The most recent archival theory advocates for a broader understanding of the principle of provenance as it is currently applied to archival representation. According to this post-modernist theory, archival representation ought to be open to multiple perspectives and interpretations instead of describing collections according

to a single fixed viewpoint.[1]

The main purpose of this thesis is to address the research problem of enhancing the representation of paper-based, twentieth-century writers' archives using semantic web technologies and particularly linked open data. The ultimate goal is to transform these archives into dynamic research tools capable of conveying complex relationships and evolving over time as new connections between records and data are discovered.

Because the goal of this project is to improve representations of writers' archives within a university department specialising in philology and literature, I will concentrate on supplementing the existing archival description with textual interpretations, which Greetham defines as "all activities associated with the discovery, description, transcription, editing, glossing, annotating, and commenting on texts".[2] Using practical examples, I will show how the finding aid for Giuseppe Raimondi's archive, once transformed into a linked open dataset, can be enhanced with literary and textual interpretations from different scholars. The goal is to demonstrate how finding aids may evolve from static search tools to open research platforms, and how their role may consequently change as the line between archival description and editorial interpretation starts blurring.

The research questions (**RQs**) that guided this study were as follows:

RQ1 How can semantic web technologies be used to represent heterogeneous types of documents and relationships in a multidimensional way, particularly in complex scenarios involving multiple interconnected documents?

RQ2 How can records and their relationships be described at different levels of granularity, such as record sets, individual records, record parts, as well as content and text?

RQ3 How to integrate 'a' reading of the archive with archival description? How to represent such a reading so to make new users of the archive able to explore and evaluate such a reading? How to represent conflicting, yet coexisting, readings of the archive by ontologies?

[1] See Douglas, 'Archiving Authors: Rethinking the Analysis and Representation of Personal Archives', 260.

[2] David C. Greetham, *Textual Scholarship: An Introduction*, Google-Books-ID: RDJ4nTA_3Z8C (Routledge, 1994), 2.

Research Hypotheses

For each of the three subquestions presented above, this study is based on one hypothesis as follows:

Hypothesis for RQ1: It is possible to extend existing ontologies that are commonly used in the LAM domain with a minimal number of new terms to represent various types of writers' documents that have different characteristics.

Hypothesis for RQ2: Using data-centric representations based on RDF, instead of document-centric representations such as EAD/XML, can enable the representation of archival entities at different levels of granularity, from sets of documents to individual document components and content. Moreover, RDF-based data models allow entities to be attributed to multiple classes concurrently.

Hypothesis for RQ3: Leveraging established models like nanopublications can allow for the seamless integration of multiple interpretations into a single archival knowledge base. This approach also permits the inclusion of provenance information to RDF assertions, thereby improving the dependability and transparency of the data.

Research Objectives and Contributions

The ultimate goal of this study is to develop a conceptual model and ontology for representing the archives of writers in the FICLIT Department. The model should be capable of expressing complex interrelationships between heterogeneous types of documents and data, and it should be flexible enough to accommodate different levels of granularity, ranging from record sets to individual items to specific item components. Moreover, the model should be designed to enable multiple agents to contribute to the archival representation, thereby enriching it with new contextualisations expressed in the form of networks of meaningful relationships. The primary objective of the model is to facilitate knowledge discovery within the writers' archives of the department. However, it should also be applicable to other paper-based personal archives from the twentieth century.

In order to satisfy the specified requirements, the conceptual model must achieve three objectives (**ROs**), each corresponding to one of the research questions listed earlier.

RO1 Firstly, the model should permit the multidimensional representation of various record types, enabling a multi-entity description of individual documents such as notebooks, books with annotations, letters, postcards, and albums.

RO2 Secondly, it should facilitate the expression of records and relationships at varying levels of granularity, ranging from sets of records to individual records to record components.

RO3 Lastly, the model should allow for the incorporation of multiple — possibly competing — perspectives, especially those of textual researchers, into the finding aid.

This study makes the following contributions (**RCs**) to the field of archival representation:

RC1 The development of a conceptual model for writers' archives based on widely-adopted standards and ontologies. This model allows for multiperspective, ever-evolving representations of heterogeneous documents and data.

RC2 The creation of an RDF representation of Giuseppe Raimondi's archive and library, which is based on the conceptual model developed in this study.

RC3 The design of a set of Python scripts that can transform the existing ISBD descriptions of the FICLIT Department's archival collections into RDF statements.

3.2 Research Design

The Giuseppe Raimondi Fonds was chosen as a representative case study of archives belonging to twentieth-century writers from the collection of personal archives held at the Ezio Raimondi Library of the FICLIT Department. The existing finding aid for the Giuseppe Raimondi Fonds, as well as published literature on the material it contains, were examined with the goal of identifying the characteristics and complexities for each of the different types of records in the collection, as well as defining the modelling requirements for transforming the existing finding aid into an archival knowledge base.

The state of the art in archival representation, particularly in the use of linked open data, was reviewed in order to assess the suitability of existing conceptual models and ontologies for the representation of twentieth-century writers' archives,

with an eye towards integration and interoperability with data curated by cultural institutions other than archives, particularly libraries and museums. While the state of the art presented in Chapter 2 revealed a number of gaps in RiC (the ICA standard for archival representation), CIDOC-CRM and its bibliographic extension FRBRoo were identified as particularly well-suited base models for developing an ontology for writers' archives.

A conceptual model and ontology for representing writers' archives was developed using competency questions as a means of formalising the modelling requirements and determining the scope of the representation. Competency questions are user-oriented queries that the developed ontology should be able to answer.[3] The same questions were also used later to evaluate the expressivity and completeness of the representation.

An RDF dataset based on the FICLITDL ontology was created from the existing finding aid, originally expressed according to the ISBD standard. RDFLib, a Python library for working with RDF data, was used to create a series of Python scripts that performed the transformation. The data extraction from the ISBD finding aid to the RDF dataset heavily relies on the use of regular expressions in order to extract atomic pieces of data from the text-based descriptions. The ISBD input, the transformation scripts, and the RDF output can be found at https://github.com/fgiovannetti/giuseppe-raimondi-lod, a public Github repository containing all code and data produced for this project.

The following sections describe the main phases of the research in more detail.

3.2.1 Analysis of the Giuseppe Raimondi Fonds

A case study approach was used to gain a better understanding of the heterogeneous entities and relationships that characterise the FICLIT Department's personal archives, with the goal of identifying the main objectives and requirements for the definition of a conceptual model for representing writers' archives.

The initial step of the process involved selecting and analysing a representative case study for the twentieth-century writers' and scholars' archives of the FICLIT Department. The Giuseppe Raimondi Fonds, which is located at the library of the Department of Italian Studies at the University of Bologna (which, together with the

[3] Michael Grüninger and Mark S. Fox, 'Methodology for the design and evaluation of ontologies', in *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*, Montreal (1995).

Department of Classical Philology forms the FICLIT Department), was chosen as a case study to develop the conceptual model. There were three primary reasons for selecting this archive. First, it contains a diverse range of document types that are highly representative of the materials found in twentieth-century writers' archives. Second, a digital inventory of all the records it contains was already available, with records described at the item level or, at least, at the file level. Finally, published research on the archive is available to test the possibility of enriching the base archival description with multiple additional interpretations.

Giuseppe Raimondi, born in Bologna in 1898, played a critical role in the cultural landscape of both Bologna and Italy throughout the twentieth century. In 1918, he established *La Raccolta*, an arts and literary journal that featured works by European authors, including Vincenzo Cardarelli, Giuseppe Ungaretti, Guillaume Apollinaire, and Blaise Cendrars. Many of these writers Raimondi encountered during World War One and the immediate aftermath.[4] The archive exemplifies what Tomasi refers to as 'multiple sedimentation', as it contains diverse materials, including notebooks, loose papers, albums, newspaper clippings, printed volumes, letters, drafts, notes, newspaper and journal articles, illustrated postcards, drawings, and photographs.[5]

In 1988, the Soprintendenza per i beni librari e documentari acquired the collection and proceeded to reorganise and catalogue all of its materials.[6] The Raimondi family, along with the Soprintendenza and the Department of Italian Studies at the University of Bologna, had a common intention that the collection would remain intact and not be dispersed. Instead, they wanted it to be preserved with its original composition and integrity, serving as both a source and an object of study. In 1996, the collection was transferred permanently to the Department of Italian Studies Library to facilitate its accessibility and use.[7]

The analysis of the case study comprised two main phases, both aimed at identifying complexities and requirements for the archival description of heterogeneous writers' archives that feature both personal papers and libraries: an analysis of the

[4] See Giuseppe Raimondi, 'Così nacque La Raccolta', *Il Resto del Carlino*, 20 August 1970.

[5] Francesca Tomasi, 'L'edizione delle carte degli scrittori: l'archivio letterario incontra il (nativo) digitale', *Biblioteche oggi Trends* 7, no. 1 (2021): 81.

[6] The Soprintendenza per i beni librari e documentari dell'Emilia-Romagna was a division of the Istituto per i beni artistici, culturali e naturali della Regione Emilia-Romagna (IBC). Recently, both entities have undergone a rebranding process, with the Soprintendenza being renamed as Soprintendenza archivistica e bibliografica dell'Emilia-Romagna and the IBC now referred to as Settore Patrimonio Culturale Emilia-Romagna.

[7] https://patrimonioculturale.regione.emilia-romagna.it/biblioteche/retrospettiva/graphe-1/schede-graphe/il-fondo-raimondi.

existing finding aid for the collection, and a survey of the published research on the collection.

The original finding aid for Giuseppe Raimondi's personal archive and library, dating from the early 1990s, was compiled by the IBC following the International Standard Bibliographic Description (ISBD), a set of rules produced by the International Federation of Library Associations and Institutions (IFLA) to create bibliographic descriptions in a standard, human-readable form, especially for use in bibliographies or library catalogues. ISBD was adopted because it is the standard in use in the cultural institution designated for conservation — the Ezio Raimondi Library — and because a large portion of the material is of a bibliographic nature.

The finding aid created for the Giuseppe Raimondi Fonds was meticulously crafted to preserve the original arrangement of materials as found in the writer's house. Raimondi himself had meticulously organised the materials, including manuscripts, correspondence, and albums. In particular, the albums were created in 1977 for an exhibition that Raimondi co-organised with Clemente Mazzotta, and they represent a later reorganisation of selected pieces of correspondence with various authors and intellectuals. While notebooks and manuscript articles are described at the item level, correspondence is described at the file level. Each albums is described separately, but the documents they gather are not.

The Giuseppe Raimondi Fonds consists of two distinct subfonds: one containing the author's papers and the other containing the author's library. The former is divided into four series: manuscripts of articles, correspondence, albums, and notebooks. The albums technically form part of the correspondence series as they contain pieces taken from that series.

The first series comprises Raimondi's manuscripts of articles (479 in total) and is further divided into two subseries: a first subseries of 17 folders, sorted by year by Raimondi himself, containing fair copies of articles and short stories written on half composition sheets that were published in different newspapers, including *Il Resto del Carlino*, *La Nazione*, and the *Corriere della Sera*. The second subseries contains 27 manuscripts for prefaces, conferences, and homages, dating from 1947 to 1971.

The second series, the correspondence, is also organized by Raimondi and is structured into folders containing letters, postcards, invitations, and other materials from each of the 823 folders dedicated to one particular correspondent. Sometimes the folders contain copies of Raimondi's replies.

The third series consists of 14 albums dedicated to key correspondents of Rai-

mondi, including Emilio Cecchi, Paul Valéry, Giuseppe Ungaretti, Carlo Carrà, Leo Longanesi, Filippo De Pisis (2 albums), the Dada art movement, Guillaume Apollinaire, Clemente Rebora, Ottone Rosai, Riccardo Bacchelli, and Vincenzo Cardarelli (2 albums).

The fourth series is comprised of notebooks (about 460 in total) and is further divided into three parts. The first subseries of notebooks consists of 308 manuscript notebooks written between 1954 and 1976, containing notes, short stories, reviews, and articles. The notebooks are organised into folders by year by Raimondi. The second subseries of notebooks consists of 140 notebooks, organised into folders by year in line with the first subseries. The third group, which does not constitute a unitary subseries, includes a set of 16 smaller notebooks of notes and 11 loose notebooks.

The personal library of the author features over 6,000 publications, including the author's personal copies of some of his published works.

Table 3.1 below describes the structure of the Giuseppe Raimondi Fonds.

3.2.2 Definition of a Conceptual Model for the Representation of Writer's Archives

The analysis of the case study of Giuseppe Raimondi's personal archive and library was followed by a review of existing conceptual models and ontologies in the cultural heritage domain that could be leveraged to define a conceptual model for the representation of writers' archives.[8]

The developed model, based on CIDOC-CRM, reused and merged existing ontologies, extending them with new classes and predicates only when strictly necessary. The development process, as already said, was based on competency questions that were formulated as a result of the analysis of the different types of documents that are comprised in Giuseppe Raimondi's archive. Such questions, divided by document type, were used as a guidance as well as a litmus test for evaluating the expressivity of the obtained linked archival dataset.

[8] See Chapter 2 for the review of existing models and Chapter 4 for the detailed presentation of the developed conceptual model.

Biblioteca	Monografie			Volumi
	Riviste			Fascicoli
Archivio	Articoli	[Manoscritti]	Carpette per anno	Singolo articolo
		[Manoscritti di Prefazioni, Conferenze, Omaggi]		Singolo articolo
	Corrispondenza		Carpette ordinate alfabeticamente secondo il cognome del mittente	Singola lettera, cartolina, o altro documento
	Album		Album dedicati a importanti corrispondenti di Raimondi	Singola lettera, cartolina, o altro documento
	Quaderni	[Quaderni manoscritti]. 1954-[1976]	Carpette per anno	Singolo quaderno
		[Quaderni manoscritti]	Carpette per anno	Singolo quaderno
		Gruppo di 27 quaderni sciolti		Singolo quaderno

Table 3.1: The structure of the Giuseppe Raimondi Fonds. The columns (left-to-right) lists the subfonds, the nested series and subseries, information about series and subseries are organised, and the analytical level at which the objects are described.

The adoption of CIDOC CRM as a base model was motivated by the need to pursue the integration of different descriptive models (not only in the domain of archives, but also in the domains of libraries, museums, and textual studies) in order to represent heterogeneous collections such as writers' archives. The bibliographic resource description model FRBRoo is already CIDOC-CRM compatible and versatile. It can, in fact, be used to describe not only mass-produced bibliographic resources, but also manuscripts and thus unique documents. FRBRoo can also represent the text and its fragments, as well as their relationships, and can be linked to a TEI-encoded transcription of the text.

Because neither CIDOC-CRM nor FRBRoo have an official OWL implementation, I have chosen to use the Erlangen versions of these conceptual models. The Erlangen CRM/OWL and FRBRoo/OWL, maintained by the University of Erlangen-Nuremberg, are widely used interpretations of the CIDOC CRM and FRBRoo in a logical framework that attempts to be as close as possible to the texts of the original specifications.

3.2.3 Creation of the Knowledge Base of the Giuseppe Raimondi Fonds

On the basis of the model developed, the original ISBD finding aid of the Giuseppe Raimondi Fonds was transformed into an RDF dataset composed of distinct named graphs, one for each different contributor to the archival knowledge base, with the open possibility of adding further interpretations, and thus further named graphs, as new research on the material is carried out. The next sections describe the main steps which were undertaken in order to perform the transformation.

The transformation process was preceded by the digital scanning of the manuscript articles and notebooks, while the scanning of the other documents (albums, correspondence, annotated books) is still ongoing. The scanning was done by the ADLab of the University of Bologna using two different flatbed scanners (a Plustek A3 OpticPro A320L and a Plustek OpticBook 4800).[9] The obtained facsimiles are at a resolution of 300 DPI and are available on the FICLIT Department Digital Library at https://dl.ficlit.unibo.it/s/lib/faceted-browse/3, along with basic metadata. It is possible to download the metadata in JSON-LD as well as the IIIF manifest of the

[9] The ADLab (Laboratorio Analogico-Digitale, https://site.unibo.it/adlab/it) is dedicated to the digital acquisition and publication of the documentary heritage of the University of Bologna and other cultural institutions.

facsimiles for each notebook in the digital library, allowing users and applications, including digital scholarly editions, to access and interact with the images.

The FICLIT Digital Library is a work in progress. Omeka-S is the current platform on which the digital library is built. The platform allows for the online publication of metadata in the form of linked open data.[10] However, experiments with Omeka-S revealed several limitations when working with RDF datasets. One of these is that its underlying data model is relational rather than graph-based, which makes it difficult to support seamless exploration of large RDF datasets, especially if they use named graphs. As a result, it was decided for the time being to publish only some of the statements that constitute Giuseppe Raimondi's archival knowledge base and to continue experimenting with other platforms that are becoming available, such as ResearchSpace, as discussed in greater detail in Chapter 6.

Definition of a URI Scheme

Any piece of information on the Semantic Web is represented by a unique URI, as explained in Chapter 2. The structure used to create the URIs of the resources populating the dataset of Giuseppe Raimondi's archive and library complies with the W3C's best practices guidelines.[11] One of the principles predicated by the W3C is to distinguish between different representations of a resource by assigning a distinct URI to each dimension being represented.

For example, a real-world person, a digital image of that person, and an HTML page about that person will all have three distinct but related URIs. Assigning URIs to the resources in a dataset necessitates the creation of a consistent URI scheme. This is especially important when dealing with composite objects like albums, as multiple unique identifiers are required to identify the different analytical levels of description: the album as a whole, a specific album page, particular items attached to the pages (e.g. a photograph or a letter), the content of such items, and so on. On the contrary, one accession number for the entire album was sufficient within analogue archival descriptions. Data-centred digitisation enables representations to capture complexity better.

The URI scheme was created with the FICLIT Digital Library in mind, which will feature the representation of multiple interconnected personal archives. The scheme

[10] Omeka-S is available at https://omeka.org/s/.

[11] See Leo Sauermann and Richard Cyganiak, *Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008* (3 December 2008), accessed 9 January 2020, https://www.w3.org/TR/cooluris/.

is built on a hierarchical pattern that represents concatenations of concepts and identifiers starting at the top level (the entire digital library, the personal archives it contains, the series within such archives, and so on) and working down to the individual item and its content.

The following is the schema used to generate URIs for the entities in the archival knowledge base:

Base URI: https://w3id.org/ficlitdl/

Dataset URI: [base uri]giuseppe-raimondi-lod/

Archival documents

Document: [dataset uri][document type]/[document id]

Physical document: [dataset uri][document type]/[document id]/object

Document content (expression): [dataset uri][document type]/[document id]/text

Archival record sets

Archival record set: [dataset uri]/record-set/[record set id]

Archival record set (physical): [dataset uri]/record-set/[record set id]/object

Other entities

Person: [base uri]person/[person id]

Organisation: [base uri]org/[organisation id]

Place: [base uri]place/[place id]

Date: [base uri]time-span/[date]

Dimension: [base uri][dimension type]/[dimension]

The unique identifiers used to identify the individual documents correspond to the alphanumeric inventory numbers assigned by the archivist who performed the initial appraisal and description process.

Development of ISBD-to-RDF Transformation Scripts

Using a set of Python transformation scripts, the existing ISBD finding aid of Giuseppe Raimondi's personal archive and library was converted to RDF. The scripts used RDFLib, a Python library designed specifically for working with RDF.[12] Regular expressions were used to extract specific information from textual descriptions.

Different sets of scripts were created to address the specific characteristics and requirements of each document type (articles, notebooks, correspondence, albums, books).[13] The semantics of the RDF dataset is informed by the conceptual model for writers' archives, the FICLITDL ontology (FICLITDL-O), which is extensively described in Chapter 4.

An iterative cleaning process was performed on the input description in parallel with the development of the transformation scripts. When the scripts were applied to the original finding aid, typos and inconsistencies in the ISBD description of the archival collection were indeed identified and corrected.

Named Entities Recognition and Reconciliation

Personal names were automatically recognised and extracted using NER techniques. Each personal name was then reconciled to VIAF, the Virtual International Authority File, and Wikidata on order to provide users with an additional entry point to the archival knowledge base. [14] The recognition, extraction, and classification of the named entities was performed using spaCy, an open-source Python library for natural language processing which provides a model for Italian.[15] Since the documents of the Giuseppe Raimondi Fonds have not been yet transcribed systematically, the extraction was performed on the descriptions of each individual item provided by the archivist as part of the finding aid (the same descriptions from which the majority of the data were extracted using regular expressions to populate the knowledge base). Once transcriptions of the documents become available, it will be possible to extend the operation using the same spaCy script.

The output produced by the NER script was a CSV file containing all extracted entities associated with the inventory number of the individual document or record

[12] The documentation of RDFLib is available at https://rdflib.readthedocs.io/en/stable/index.html.

[13] All scripts are available at https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/scripts.

[14] Place names, organisation names, and bibliographic references will be reconciled and matched to external authority files in the future.

[15] SpaCy is described at https://spacy.io/.

set they refer to. The CSV file was then uploaded to OpenRefine, a browser-based, open-source tool for working with large datasets that allows for external data to be added.[16] The personal names mentioned in the record descriptions for Giuseppe Raimondi's archive were reconciled with VIAF and Wikidata using OpenRefine, with the goal of disambiguating and normalising them.

3.2.4 Enrichment of the Knowledge Base with Additional Contextualisations

The RDF dataset created by converting the existing finding aid was compiled into a named graph that forms the base description or archivist's graph for the archival collection.[17] The provenance of the named graph was described using the Nanopublication data model, which allows to represent not only the agents responsible for the assertions, namely the archivists at IBC, but also the agent responsible for the creation of the dataset.

To test the feasibility of enriching the archival knowledge base with multiple additional interpretations, additional named graphs (the researchers' graphs) were created and assigned different progressive names, one for each different contribution to the knowledge base.[18]

One additional named graph containing connections between Giuseppe Raimondi's manuscript notebooks and the published editions of the texts as featured in the OPAC catalogue was incorporated into the dataset and attributed to Maria Chiara Tortora, a PhD student working on the Fonds.

Furthermore, four additional interpretations were reconstructed and added to the knowledge base. Such interpretations are based on either published literature on Giuseppe Raimondi's archive or direct analysis of the archive itself and the interrelationships that characterise it. The first reconstruction represents the interconnections between heterogeneous documents across the archive bearing different expressions of the same work; the second reconstruction represents the process of authorial revision on the title of a short story; the third reconstruction represents the influence of Paul Valéry on Giuseppe Raimondi's production; a fourth reconstruction describes the exchange of a letter and a manuscript between Giuseppe Raimondi and Giuseppe Ungaretti.

[16] OpenRefine is available at https://openrefine.org/.
[17] The name attributed to the graph is [dataset uri]nanopub/nanopub-base/assertion.
[18] The first graph, for example, is named [dataset uri]nanopub/nanopub1/assertion.

Each additional layer is represented by a distinct named graph that includes provenance information. It is worth noting that nanopublications can represent not only the provenance of a set of assertions, but also the provenance of the nanopublication itself. This is especially important because any transformation from one medium to another (such as the transformation from ISBD to RDF used in this study) adds another interpretative layer to the data as the person (or software) who creates the RDF triples differs from the person who created the assertions.

3.2.5 Evaluation of the Conceptual Model

The evaluation of the conceptual model was done on the enriched archival knowledge base. Specifically, the set of competency questions developed for each enrichment scenario during the modelling phase were used. The process is described in detail in Chapter 5, where each additional reconstruction is illustrated and queried. All competency questions, their SPARQL formulation, as well as the obtained results are available from https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries.

3.3 Methodological Limitations

The study only examined twentieth-century, paper-based writers' archives. It is important to note, however, that as the FICLIT Department acquires additional writers' and scholars' archives in the near future, they will almost certainly be in a hybrid or entirely digital format, necessitating significant adjustments to the conceptual model developed for this research.

Additionally, the complexity and variability of archives should be taken into consideration when attempting to generalise the results of this study. Archives, like texts, are intricate systems, each with its unique characteristics. No two writers' archives are identical, which means that the transformation scripts developed in the framework of this study may need to be modified to accommodate document types other than those considered for this research, such as notebooks, albums, loose manuscripts, postcards, letters, and annotated printed volumes.

Chapter 4

Modelling Writers' Archives: the FICLITDL Conceptual Model

The principal objective of this study is the definition of a conceptual model and ontology for representing writers' archives in the FICLIT Digital Library and other similar frameworks. The literature on archival representation discussed in Chapter 1 suggests a need for a conceptual and methodological change in current practices. It recommends that the hierarchical structure of archives should be expanded to incorporate graph-based representations, allowing for more intricate connections among individual documents and data both within and across collections. The literature also highlights the importance of reorienting archival descriptions towards the level of the individual document. This shift would facilitate the analytical representation of relationships among individual documents and broaden the representation of archival context as a network rather than solely a hierarchical structure. Furthermore, the literature emphasises how archival descriptions are the result of a subjective interpretation process that is never neutral. As a result, it is critical that this subjectivity be acknowledged and explicitly represented in finding aids. Finally, finding aids, according to the literature, should allow for the incorporation of additional contextualisations other than the archivist's, each accompanied by information about its provenance, thereby providing multiperspective and ever-evolving representations rather than monolithic ones.

This chapter presents and critically discusses the proposed conceptual model and ontology. The model addresses the representation of different document types and interrelationships, including notebooks, printed volumes bearing annotations, letters, postcards, and albums. This selection of document types, which necessar-

ily reflects the composition of the fonds used as a case study, does not claim to be exhaustive, but only suggests some of the many possible document types that can appear in a twentieth-century writer's archive. Similarly, the FICLIT Digital Library (FICLITDL) conceptual model does not cover all possible scenarios of critical enrichment of the base archival description by subsequent scholars. It can, however, be expanded to meet new representational needs as they arise because it is based on widely used models in the cultural heritage domain.

After reviewing the existing conceptual models for personal archives, it became apparent that there is a lack of a comprehensive model that addresses the representation requirements identified during the analysis of Giuseppe Raimondi's archive and library as a case study. However, the analysis of existing ontologies for cultural heritage suggests that a combination of ICOM's CIDOC-CRM and IFLA's FRBRoo (soon to be LRMoo) offers an effective basis for establishing an effective model to represent writers' archives in general, including the specific case of the FICLIT Department's writers' archives. This combined framework supports the creation of ever-evolving archival representations, accommodating multiple different readings of the archives.

Each of the following sections illustrate a different part of the model: Section 4.1 discusses ontology reuse in the FICLIT Digital Library conceptual model;[1] Section 4.2 illustrates the representation of the main types of manuscript material that populate Giuseppe Raimondi's personal archive and library – manuscripts of articles, notebooks, annotated printed volumes, letters, and postcards – using the FICLITDL conceptual model; Section 4.3 outlines the representation of manuscripts as physical objects, Section 4.4 focuses on the representation of manuscripts as texts, while Section 4.5 describes the modelling of the events in which archival documents participate. Finally, Section 4.6 addresses the description of archival order, while Section 4.7 discusses ways of accommodating multiple, perhaps conflicting, interpretations in the archival representation.

[1] The conceptual model reuses FRBRoo (specifically, Erlangen FRBRoo, an OWL DL 1.0 implementations of FRBRoo), which is about to evolve into a simplified version named LRMoo (the Library Reference Model object-oriented). Because a stable version of LRMoo has not yet been released, and IFLA does not provide an official OWL implementation of FRBRoo, the FICLITDL conceptual model employs Erlangen FRBRoo while reflecting on the changes that will be required to update the model to LRMoo once this is released.

4.1 Ontology Reuse

When developing ontologies, it is strongly advised to reuse existing models whenever possible.[2] Pinto defines ontology reuse as "assembling, extending, specialising, and adapting" existing ontologies to create a new one.[3] When existing ontologies are reused by importing them into the ontology being created, data from the imported ontology becomes available for reasoning and inference in the importing ontology.[4] Reuse may concern only specific classes or properties of an existing ontology.

Reuse has several advantages. It is not only cost-effective from the point of view of ontology engineering and maintenance, but it also plays an important role in the dissemination and collaborative validation of semantic models.[5] The FICLIT Digital Library conceptual model, which is implemented as an OWL ontology (FICLITDL-O), reuses and extends CIDOC-CRM and FRBRoo (FRBRoo too was developed by IFLA as an extension of CIDOC-CRM). The two models combined together provide a shared language for libraries and museums to integrate the representations of their collections. This harmonisation already covers part of the archival domain for concepts shared with libraries and museums (for example, the concept of curated holding, which is shared by all three types of institutions and is represented in CIDOC-CRM by the E78 Curated Holding class).[6]

The following sections illustrate the FICLIT Digital Library conceptual model and ontology. The model, as already specified, is based on and extends CIDOC-CRM and FRBRoo, while importing specific sets of classes and properties from the Provenance Ontology (PROV-O), the Dublin Core Terms (DCTERMS), the Publish-

[2] On the reuse of existing models in the process of ontology development see Silvio Peroni, 'A Simplified Agile Methodology for Ontology Development', in *OWL: Experiences and Directions – Reasoner Evaluation*, ed. Mauro Dragoni, María Poveda-Villalón and Ernesto Jimenez-Ruiz, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2017), 55–69; for a statistical analysis of the practice of ontology reuse see Mariano Fernández-López et al., 'Why are ontologies not reused across the same domain?', *Journal of Web Semantics* 57 (2019), https://doi.org/10.1016/j.websem.2018.12.010. On the advantages of ontology reuse see Deryle Lonsdale et al., 'Reusing ontologies and language components for ontology generation', *Data & Knowledge Engineering* 69, no. 4 (2010): 318–330.

[3] H. Sofia Pinto, 'Towards Ontology Reuse', *Proceedings of AAAI 99's Workshop on Ontology Management* 13 (1999): 67.

[4] See Dean Allemang and James Hendler, *Semantic web for the working ontologist: effective modeling in RDFS and OWL* (Elsevier, 2011).

[5] Elena Simperl, 'Guidelines for reusing ontologies on the semantic web', Publisher: World Scientific, *International Journal of Semantic Computing* 4, no. 2 (2010): 239–283; see also Piotr Sowinski et al., 'Ontology Reuse: the Real Test of Ontological Design' (14 September 2022), which considers ontology reuse to be the primary method for evaluating and improving ontology quality, https://doi.org/10.3233/FAIA220292.

[6] See the scope note for E78 Curated Holding in *Classes & Properties Declarations of CIDOC-CRM, version 7.1.2*, http://cidoc-crm.org/cidoc-crm/7.1.2/E78_Curated_Holding.

ing Requirements for Industry Standard Metadata (PRISM), and the Sequence Ontology Design Pattern (Sequence ODP). Constructs from the Web Annotation Ontology (OA), the Publishing Roles Ontology (PRO), and the Nanopublication Ontology (NP) are also used for addressing the representation of different aspects and characteristics of writers' archives as discussed throughout this chapter. Sections 4.3-7 will explain in more detail how each of these ontologies is employed within the FICLIT Digital Library conceptual model.

4.2 The Representation of Writers' Manuscripts

When examining Giuseppe Raimondi's archive, the presence of an exceptionally diverse range of writers' materials immediately stands out. As noticed by Crocetti, such a distinguishing feature can be observed in the archives of many twentieth-century writers, including those housed at the Ezio Raimondi Library of the FICLIT Department.[7] Indeed, Giuseppe Raimondi's manuscripts and personal papers represent all stages of literary production, from rough notes and first drafts to corrected proofs and correspondence with other writers. The archival collection includes loose sheets, each bearing a fair copy of a published article; notebooks; letters and postcards; and thematic albums. Several books in the writer's personal library also contain manuscript annotations.

The WEMI conceptual model, introduced in Chapter 2, that underpins FRBRoo, is directly applicable to the representation of non-unique bibliographic entities. Erlangen FRBRoo provides a specialised class for representing manuscripts, the F4 Manifestation Singleton class, which includes unique physical documents bearing text.[8]

Le Boeuf demonstrates the use of F4 for representing manuscripts.[9] In particular, he suggests modelling each manuscript as an instance of F4 Manifestation Singleton that *P128 carries* an instance of F2 Expression (the text), creating a relationship between a pair of FRBRoo entities (F4 and F2) via a CIDOC-CRM property (*P128*).[10] Le Boeuf's proposed model represents manuscripts effectively, including

[7] Crocetti, 'Parole introduttive a Conservare il Novecento', 23–26.

[8] Throughout this dissertation, the term 'text' is used in the broadest sense possible to include visual items and other forms of expression.

[9] Patrick Le Boeuf, 'Modeling rare and unique documents: using FRBROO/CIDOC CRM', Publisher: Taylor & Francis, *Journal of Archival Organization* 10, no. 2 (2012): 99–100.

[10] In CIDOC-CRM, the names of all classes begin with the letter 'E', while the names of properties begin with 'P'. The names of classes and properties in FRBRoo and in its updated version, LRMoo, begin with 'F' and 'R' respectively. All properties referred to in this chapter are it-

special categories of manuscripts such as notebooks or letters; however, further consideration is needed on how to represent the distinguishing characteristics of the different types of unique documents that can be found in Giuseppe Raimondi's personal archive and in the archives of similar writers.

4.2.1 Articles

Raimondi's collection of newspaper and journal articles includes fair copies of the texts, some of which have minor corrections and annotations. Some articles are simultaneously published in multiple newspapers, while others are republished in different venues years after their original publication, often with varying degrees of revision. However, the current finding aid lacks descriptions of the typescripts and does not establish links between typescript and manuscript versions of the same texts. It is also important to note that although these items are related to individual pieces in other archival series within Raimondi's collection, there are no cross-references provided in the existing description. Similarly, interarchival links to the published versions of the manuscript articles in newspaper and journal archives, such as *Il Resto del Carlino* or the *Corriere della Sera* (which published numerous pieces by Raimondi), are not available. Additionally, the archive contains a substantial number of newspaper clippings, most of which are conceptually related to Raimondi's correspondence. However, there are currently no established connections between these clippings and the manuscript article series.

The new archival knowledge base, built upon the conceptual model and ontology discussed in this chapter, aims to enhance the existing finding aid for the collection by providing answers to the following questions:

1. What is the title of the article?
2. What is the subject of the article?
3. When and where was the article published?
4. How many versions of the text are there in the archival collection?
5. Is there a typescript version of the manuscript?
6. Which version of the text came first?
7. How does the typescript differ from the manuscript?

alicised.

8. Has the article been republished in other venues, including monograph publications?
9. In the case of republication, how has the text changed through time?

Figures 4.1 schematises the modelling of manuscripts and typescripts of articles.[11] A manuscript article is represented as an individual of the F4 Manifestation Singleton class that *P128 carries* an instance of F22 Self-Contained Expression (the text of the article). The F4 class, defined in FRBRoo as comprising "physical objects that each carry an instance of F2 Expression and that were produced as unique objects, with no siblings intended in the course of their production",[12] combines the two otherwise distinct WEMI concepts of manifestation and item into a single class on the basis that, because no two manuscripts can be identical, the manifestation associated with a unique document consists of a singleton exemplar.[13]

To categorise manuscript types, an E55 Type is assigned to both F4 Manifestation Singleton and F22 Self-Contained Expression. For manuscript articles, the attributed type is 'loose-leaf papers' for the physical documents and 'newspaper article' or 'journal article' for the texts (see Section 4.3.1).[14]

The event of creation of a manuscript article is represented using the class F28 Expression Creation. However, when text is written down, a physical manuscript is also produced. As such, F28 is said to result in the creation of both the text (F2) and the physical manuscript (F4). By attributing a specific technique of creation to the expression creation event, handwritten articles can be distinguished from typewritten articles when querying the dataset. This is achieved via the relation *P32 used general technique*, which takes as a domain F28 and as a range an individual of the ficlitdlo:Technique class (in this case, the type of employed technique is either 'handwriting' or 'typewriting'), newly defined within the namespace of the FICLIT Digital Library Ontology as a subclass of E55 Type.

[11] The Graffoo graphical notation (Riccardo Falco et al., 'Modelling OWL ontologies with Graffoo', in *European Semantic Web Conference* (Springer, 2014), see) is adopted throughout the figures in this chapter to graphically represent features of the FICLITDL conceptual model.

[12] Chryssoula Bekiari et al., *Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism*, 2.4 (IFLA, 2015), 57.

[13] As explained in Pat Riva and Maja Žumer, 'FRBRoo, the IFLA Library Reference Model, and now LRMoo: a circle of development', 2018, 5, https://library.ifla.org/id/eprint/2130/1/074-riva-en.pdf , LRMoo deprecates the use of F4 in order to replace it with F5 Item, which will no longer be limited to the representation of physical documents that were industrially produced in multiple copies but will also be used for manuscripts. As a result, using the LRMoo conceptual model, it will be appropriate to assert that an instance of F5 Item (the manuscript) *P128 contains* an instance of F2 Expression (the article text).

[14] The F22 Self-Contained Expression class, like F4 Manifestation Singleton, will be deprecated in LRMoo in favour of the more generic F2 Expression.

In order to interconnect (although indirectly) all variant versions of the same article, each article text is linked to the work it realises via the *R3 is realised in* property, which "expresses the association that exists between an expression (F22) and the work that this expression conveys (F1)".[15]

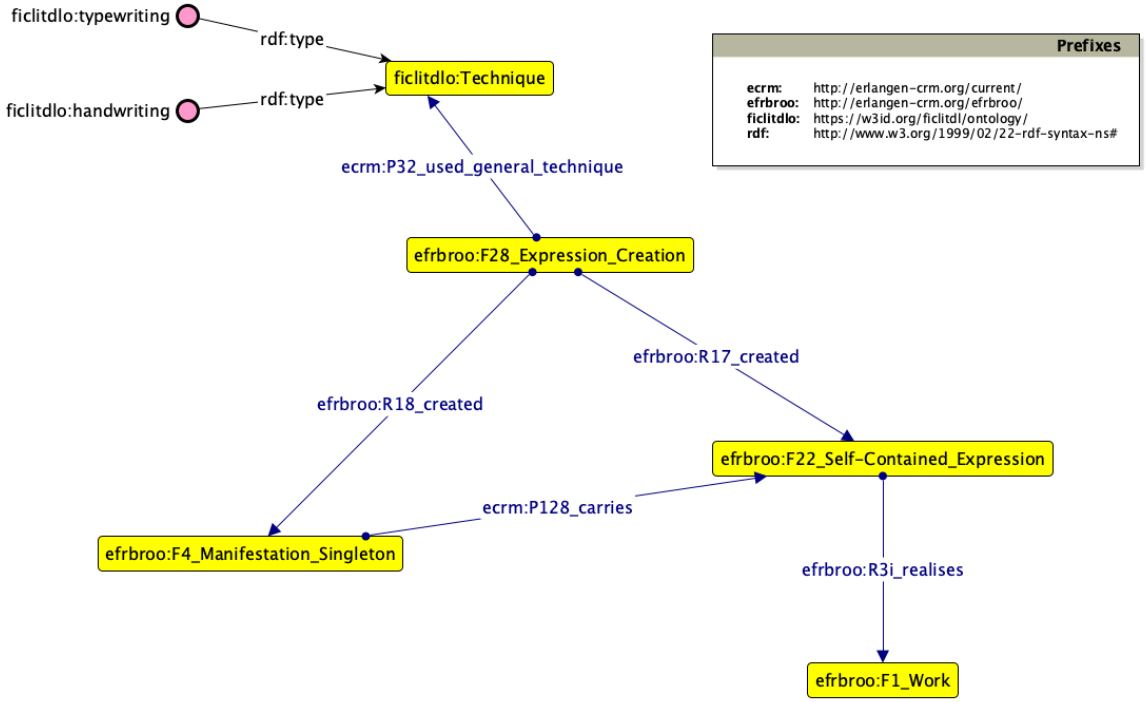

Figure 4.1: Manuscript and typescript articles.

Details on the date and venue of publication (e.g. 'first published on *ll Resto del Carlino* on 15 October 1954') are given for the published version of the articles, which are not present in Giuseppe Raimondi's archive but can be found in the archives of the publishing organisations. The texts of the published articles, as shown in Figure 4.2, are represented as members of F22 Self-Contained Expression, as suggested in PRESSoo, a conceptual model for representing continuing resources such as journals and newspapers that reuses and extends FRBRoo and is endorsed by IFLA.[16] Each F22 (the published text) is incorporated in an F24 Publication Expression (the published text, including its layout).[17]

Each published article is further described by leveraging a set of properties from the Dublin Core Terms (DCTERMS) and the Publishing Requirements for Industry

[15] Bekiari et al., *Definition of FRBRoo*, 87.

[16] Clément Oury et al., 'Definition of PRESSoo: A conceptual model for Bibliographic Information Pertaining to Serials and Other Continuing Resources (Version 1.3)', Publisher: International Federation of Library Associations and Institutions (IFLA), 2017, 10–11.

[17] PRESSoo also allows for the representation of issues and periodicals (also instances of F24 Publication Expression). However, this model limits its focus to the representation of individual articles and their features.

Standard Metadata (PRISM).[18] These properties, all taking as a domain the publication expression (F24) allow for the description of the publishing organisation (*dcterms:publisher* E40 Legal Body); the publication format (*dcterms:format* 'print');[19] the publication date (*prism:publicationDate*, which takes as a range a date expressed according to ISO 8601); the starting page and the ending page (*prism:startingPage* and *prism:endingPage*, which take as a range an integer number). The representations of the published articles within the archival knowledge base can be connected to the corresponding entities in the archives of the publishing organisations via the *owl:sameAs* property.

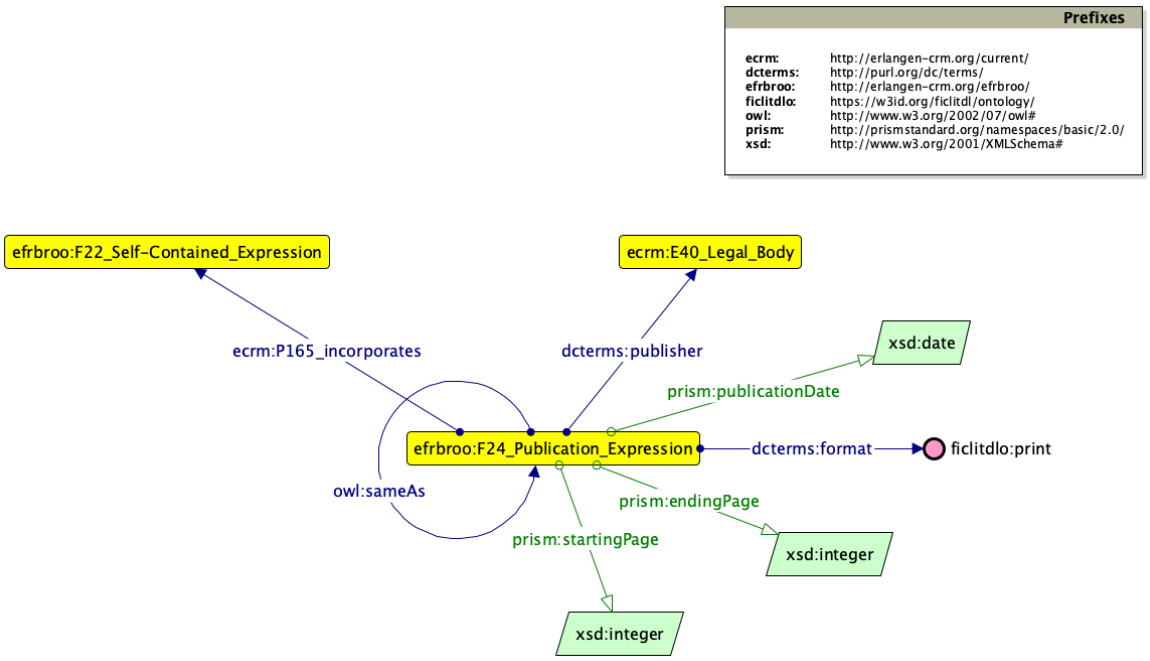

Figure 4.2: Published versions of the articles.

The relationship between each variant expression and the work it realises (F22 Self-Contained Expression *R3i realises* F1 Work) serves as an indirect link between the manuscript, typescript, and published versions of each article (see Figure 4.3). The manuscript and typescript versions of Giuseppe Raimondi's articles usually correspond to the fair copies that were submitted for publication, with a few minor changes occurring in the passage from manuscript to typescript. FRBRoo/LRMoo

[18] The SPAR Ontologies (Silvio Peroni and David Shotton, 'The SPAR Ontologies', in *The Semantic Web – ISWC 2018*, ed. Denny Vrandečić et al., Lecture Notes in Computer Science (Cham: Springer International Publishing, 2018), 119–136), a set of formal ontologies for the description of the publishing domain, adopts the same DCTERMS and PRISM properties for describing continuing publications (see 'Using FRBR for describing a bibliographic entity', http://www.sparontologies.net/examples#fabio_1).

[19] Note that the range for the *dcterms:format* relation is a member of the Format class, which is declared as a subclass of E55 Type within the FICLITDL-O namespace.

does not address expression-to-expression relations. As such, in order to further detail the connections between the versions, this model introduces a new generic relation between variant texts (F2 Expression *ficlitdlo:hasVariantVersion* F2 Expression) and borrows more specific relations from the Provenance Ontology (PROV-O), as Section 4.5.3 of this chapter, discussing the representation of textual variation, will show later.

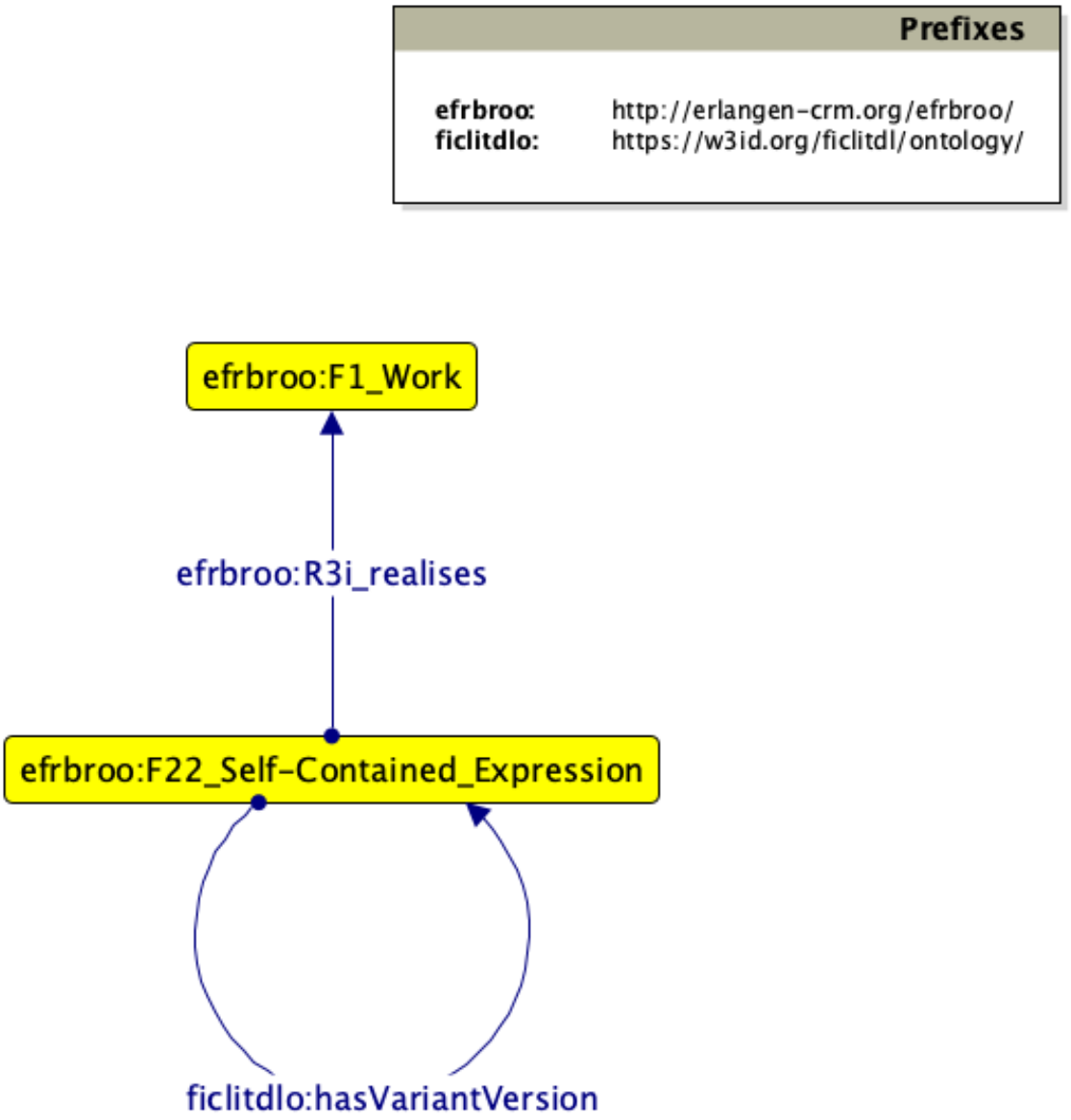

Figure 4.3: Relation between manuscript, typescript, and published articles.

4.2.2 Notebooks

Notebooks, which are a common type of manuscripts found in the archives of twentieth-century writers, have received limited attention in the archival literature.[20] Writers use notebooks in multiple ways for notes and drafts, corrected proofs and fair copies, for diary entries and personal annotations, for sketches and diagrams. As they can be adapted to many different writing functions and activities, providing a comprehensive definition of notebooks as well as establishing a shared and univocal way of representing them in the framework of writers' archives is a task fraught with difficulty.[21] Several text encoding projects have dealt with the representation of manuscript notebooks using the Text Encoding Initiative (TEI) vocabulary.[22] Only few projects, however, have addressed the representation of notebooks in graph-based rather than document-centric environments.[23]

Giuseppe Raimondi's personal archive comprises over 450 manuscript notebooks that bear autograph texts at different stages of development, from notes, to drafts, to fair copies bearing corrections and revisions. Most of the texts are short stories, genre with which Raimondi experimented extensively over the years. Some of the notebooks can contain multiple texts, while some of the texts – or groups of related texts – can span across multiple notebooks. The current finding aid describes the notebook collection at the item level, with each notebook having its own entry, with the exception of a set of 16 smaller notebooks containing preparatory notes. However, the descriptions do not cover the actual content of the notebooks, but rather report the information about content handwritten on the notebook covers by Raimondi himself or by subsequent custodians (the use of different pens and handwriting styles allows the different hands that contributed to the covers as well as author revisions to be distinguished and identified).

Due to the impossibility of thoroughly analysing the content of each notebook

[20] For instance, neither the Society of American Archivists' Dictionary of Archives Terminology (DAT) nor the International Council on Archives' Multilingual Archival Terminology (MAT) feature an entry for notebooks as a type of archival record.

[21] To gain an understanding of the diverse range of notebooks, one can explore the digital scholarly editions and experiments of notebooks that have been published online, such as Darwin Online, http://darwin-online.org.uk/EditorialIntroductions/vanWyhe_notebooks.html; Paolo Bufalini's Notebook, https://projects.dharc.unibo.it/bufalini-notebook/; Dickens Notes, https://www.dickensnotes.com/; Proust Prototype, http://peterstokes.org/elena/proust_prototype/.

[22] The TEI Guidelines are available at http://www.tei-c.org/Guidelines/P5 (see TEI Consortium, *TEI P5 Guidelines*).

[23] One of these is Paolo Bufalini's Notebook Digital Edition (see Daquino et al., 'Enhancing semantic expressivity in the cultural heritage domain').

owing to budget and time constraints, the current description most likely limits its scope to covers. As a result, users of the description – mostly textual and literary scholars – cannot be certain what each notebook contains and how a notebook may relate to other documents in the same or in a different record set without physically visiting the archive. For example, two distinct notebooks may feature alternate versions of the same work or may be related because they both contain texts sharing the same themes or subjects (e.g. the reception of Leopardi in Raimondi's production). In the case of notebooks, because the reconstruction of the archival context of a document can only emerge from the analysis and understanding of its content, it is especially important for textual scholars to be able to augment the archival representation with textual interpretation by adding to the existing finding aid.

Such representation, once enriched with supplemental readings, should be able to answer at least the following questions:

1. What and how many texts are contained in the notebook?
2. When was the notebook text created?
3. Is there a different text version carried by another notebook or a different type of document, including published material?
4. What are the differences between the text versions?
5. Does the notebook serve as a storage unit for additional documents?
6. Is there more than one title for the text? How has the title evolved over time?
7. In the case of multiple versions of the same text, which version was created first?

Generally speaking, the LRMoo work-expression-manifestation-item chain can be used to describe a notebook as an item, i.e. a material object carrying signs intended to convey intellectual content; as a manifestation, i.e. a set of all carriers that are assumed to share the same characteristics as to intellectual content and aspects of physical form (in this case the set comprises a single, unique item); as an expression, i.e. a distinct combination of signs conveying intellectual content (an aggregate expression when the notebook comprises multiple distinct texts); and finally as a work, i.e. the intellectual content of a distinct creation.

Each text in a notebook can then be described as a self-contained expression that is related to the work it realises as well as to possible variant expressions, regard-

less of their carrier type. The whole of the texts contained in a notebook form an aggregate expression.

However, if we were to use LRMoo, notebooks as material objects could not be treated simply and consistently as members of the F5 Item class. A notebook may indeed carry multiple texts or one single text may fill up more than one notebook as schematised in Figure 4.4. In the first case, a notebook can be regarded as both an individual document and an archival unit, namely a file, whereas in the second case, the archivist could look at a set of notebooks as one unitary document composed of different pieces.

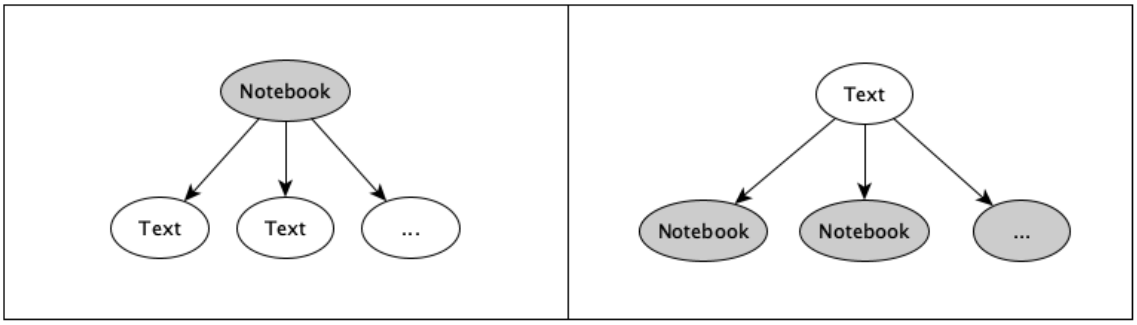

Figure 4.4: A notebook containing multiple texts (left) and a text spanning multiple notebooks (right).

Giuseppe Raimondi's notebooks are frequently organised chronologically, with no explicit interdependency between the texts. As a result, a notebook containing multiple texts would need to be modelled in LRMoo as an instance of E22 Human-Made Object that *P46 is composed of* multiple instances of F5 Item, whereas a collection of notebooks carrying a single text would need to be modelled as a single instance of F5 Item that *P46 is composed of* multiple instances of E22 Human-Made Object. If a notebook contains only one text, it can be regarded as an E22 Human-Made Object and an F5 Item at the same time because the boundary of the object and that of the item overlap. In this case, a manuscript notebook is instantiated as a member of both E22 and F5. As anticipated, we use F4 Manifestation Singleton instead of F5 Item for practical reasons, with the intention of updating this conceptual model in the near future, but the remainder of the conceptualisation outlined in this paragraph remains the same (see Figure 4.5).

Using the WEMI model, we can continue the description of a manuscript notebook from item to expression to work (see Figure 4.6).[24] As noted earlier, the item

[24] In the case of unique documents, the manifestation level is less important because, whereas the

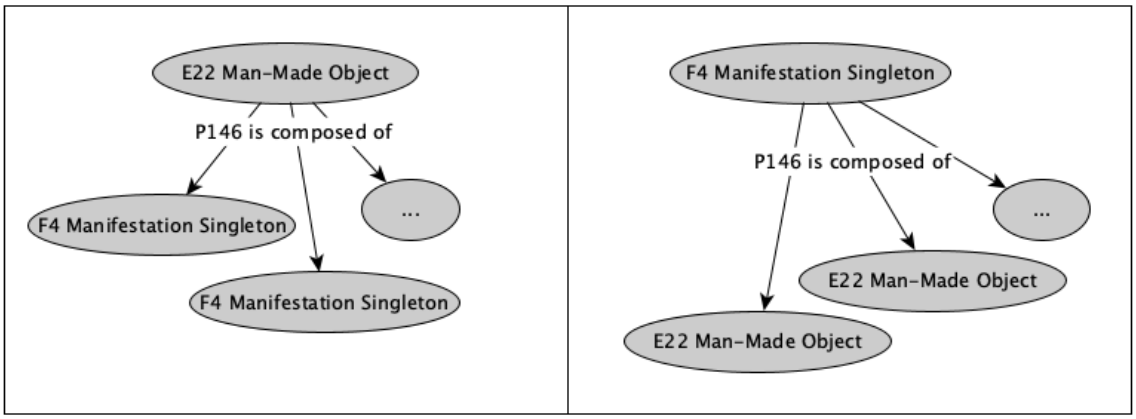

Figure 4.5: Representing a notebook containing multiple texts (left) and a text spanning multiple notebooks (right).

level corresponds to the notebook or to a part of it seen as a physical document "carrying signs intended to convey intellectual content", the expression level corresponds to the notebook or to a part of it seen as "a distinct combination of signs conveying intellectual content", and the work level corresponds to the notebook or to a part of it seen as "the intellectual content of a distinct creation".[25] Finally, it is important to note that all characteristics attributed to E22 also apply to F5, and it is possible for a notebook to belong to both E22 and F5.

When a notebook contains multiple texts, each text realises an individual work. Formally speaking, a work is realised in an expression (F1 Work *R3 is realised in* F22 Self-Contained Expression); an expression can incorporate n expressions (F2 Expression *P165 incorporates* F2 Expression), which are carried by a notebook (E22 Human-Made Object/F4 Manifestation Singleton *P128 carries* F22 Self-Contained Expression). It is worthy to note the use of the *P128 carries* property. The property is defined in CIDOC-CRM and can be used as a shortcut to link a notebook and an expression directly as suggested by Le Boeuf.[26] FRBRoo does not allow you to skip any of the four levels of representation of a resource because it lacks expression-item, work-item, and work-expression links. However, because items and expressions are subclasses of E18 Physical Thing and E90 Symbolic Object respectively, the *P128 carries* property can be used to link an item to the expression it bears. Notebooks in

manifestation of a non-unique document includes all carriers that share the same characteristics in terms of intellectual content and physical form, the manifestation of a unique document consists of a single, unique carrier corresponding to the item itself (and it is for this very reason that, in FRBRoo, the concepts of manifestation and item, in the case of unique documents, were merged into one additional F4 class).

[25] Pat Riva, Patrick Le Boeuf and Maya Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information* (2017), 21–27.

[26] See Le Boeuf, 'Modeling rare and unique documents'.

Giuseppe Raimondi's archives are sometimes used as physical containers for other loose leaf documents such as manuscripts, typescripts, press clippings, and notes, or even for other notebooks, which usually have some kind of relationship with the texts of the notebooks themselves. This scenario can be modelled using a new CIDOC-CRM property which has been introduced with version 7.0.1. The property, *P198 holds or supports*, "relates one instance of E18 Physical Thing which acts as a container or support to a supported or contained instance of E18 Physical Thing. Typical examples of E18 Physical Things which are intended to function as a container or support include shelves, folders or boxes".[27]

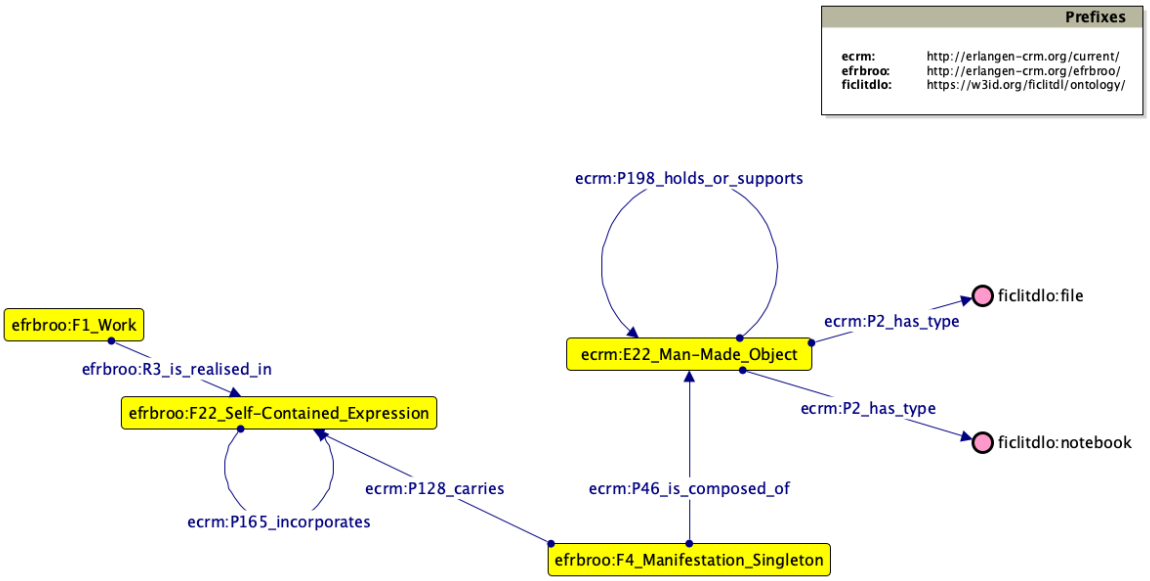

Figure 4.6: Representation of a manuscript notebook.

4.2.3 Annotated Printed Books

Giuseppe Raimondi's personal library contains over 6,000 literary and art publications, including about 50 exemplars of his own monographs. Some of these books feature manuscript marginalia, underlines, highlights, and folded page corners, which provide evidence of Raimondi's activity as a reader. Despite the important information that writers' libraries can provide, the logical divide between writers' papers and libraries is frequently visible in archival representation. In this regard, in a 2006 article on private libraries Attilio Mauro Caproni defined an author as

[27] Chryssoula Bekiari et al., *Definition of the CIDOC Conceptual Reference Model (Version 7.1.2)* (2022), 218.

someone who does not only write and publish, but also reads.[28] Caproni's statement was intended to draw scholars' attention not only to autograph manuscripts in the archive, but also to book collections, particularly if they bear signs of the author's reading activity, and to highlight the critical role that writers' libraries can play in the scholarly process of reconstructing the archive and providing context for paper records.

The existing finding aid for the Giuseppe Raimondi Archive provides no connections between the author's papers and library. However, Maria Chiara Tortora, a PhD student researching the reception of Giacomo Leopardi and Carlo Emilio Gadda's works in Giuseppe Raimondi's literary production, has reconnected most of the notebook texts to their published versions in the OPAC SNB online catalogue. On the basis of her reconstruction, the archival knowledge base can be supplemented with links connecting the texts in the notebooks (author who writes) to the corresponding published versions in the author's library (author who publishes) and, if further scholarly reconstructions becomes available, to the publications that influenced the process of authorial creation (author who reads).

Taking all this into account, the new archival knowledge base, once enriched with further archive-library connections, should be able to answer at least the following questions:

1. Which books contain annotations?
2. Which books feature specific types of annotations (e.g., underlines)?
3. Who is the author of the annotation? Is the author of the annotation the same as the author of the annotated publication?
4. When was the annotation created?
5. What is the content of the annotation?
6. What part of the text does the annotation refers to?
7. What page does the annotation appear on?
8. Which books contain other documents?
9. Which books contain specific types of documents (e.g., a typescript)?
10. Is the book linked to different documents in the archive (e.g., a book attached to a letter sent to Raimondi)?

[28] "L'autore è una persona che scrive, legge, pubblica" (Caproni, 'Biblioteca privata: ipotesi di definizione', 22–28).

11. What is the title, the publication year, the author, the publisher, the place of publication, the edition number, the shelfmark and inventory number, the place of conservation of the book?

This list of competency questions will be used as a litmus test to evaluate the conceptual model and ontology later on in Chapter 5. An annotated printed volume bears an expression (the published text) which is augmented with additional manuscript material that was not intended as part of the original work. The presence of such additions on a printed volume conceptually transforms it from one of many identical copies to a unique record.[29] The author of the annotation may or may not be the same as the author of the publication (Raimondi reads and annotates the works of others but he also returns on his owns).

In FRBRoo (and LRMoo), which is specifically designed for bibliographic description, we can easily model a printed volume as an item (the physical copy) that materialises a manifestation (the specific edition) which, in turn, embodies the expression (the publication text). Such expression realises the work. Nonetheless, applying the WEMI chain to the representation of annotated volumes poses some challenges. A printed book bearing annotations can be regarded as an item that carries multiple expressions, namely the publication text as well as the additional manuscript annotations (these can take the form of marginalia, interlinear notes or text revisions, and markup such as underlines or other meaningful symbols).

The LRMoo documentation, however, suggest that members of the F5 Item class that are "enhanced through the addition of manuscript annotations, or any material that was not intended by the publisher, such as press clippings, dried flowers, etc." be regarded as archival holdings exceeding the boundaries of the F5 Item class.[30] The workaround proposed by LRMoo is to model such holdings as storage units composed of distinct multiple entities. The storage unit would be an instance of E19 Physical Object, with each component linked to it via the *P46i forms part of* property.

This solution is suitable for representing objects physically stored within the pages of books (for example, a typescript that has been folded and included among the pages of a volume), but it is insufficient for representing annotations. Applying

[29] Renzo Cremante, 'Introduzione', in *Conservare il Novecento: convegno nazionale, Ferrara, Salone internazionale dell'arte del restauro e della conservazione dei beni culturali e ambientali, 25-26 marzo 2000: atti*, ed. Maurizio Messina and Giuliana Zagra (Roma: Associazione italiana biblioteche, 2001), see; Luigi Crocetti, 'Indicizzare la libertà', *Biblioteche oggi* 20, no. 2 (2002): 8–11; De Pasquale, 'Private archives in the library. Types, acquisition, treatment and description'.

[30] Bekiari et al., *Definition of the CIDOC Conceptual Reference Model (Version 7.1.2)*, 22.

the concept of storage unit to manuscript annotations would result in a representation that treats the annotations as separate entities from the text to which they refer, disregarding that annotations acquire meaning only when placed in relation to the annotated text with which they form and interdependent system. The same observation holds true for objects whose storage in a book has a specific meaning, such as a bookmark used to mark a location in the book.

How should annotated printed volumes be then represented as unique documents using CIDOC-CRM and FRBRoo/LRMoo? The LRM conceptual model introduces the concept of aggregates. An aggregate is a manifestation that embodies multiple expressions, that is, a publication that contains multiple distinct texts. LRM defines three different categories of aggregate. Among these are aggregates that result from an activity of augmentation, such as annotating. An augmented aggregate consists of a "single independent work that has been supplemented with one or more dependent works [...]. Forewords, introductions, illustrations, notes, etc. are examples of augmenting works".[31] The manuscript annotations are the augmenting works in the case of annotated printed books: they do not produce an augmented publication, but they do augment a specific copy of a book, transforming it into a unique document.

The LRM proposed model for aggregates resulting from augmentation can be applied to annotated volumes. As previously stated, an LRM aggregate is a new manifestation (F3) that *R4 embodies* multiple expressions (each is a distinct individual of F2 or of any of its subclasses). The expressions altogether, i.e. the text of the publication and the manuscript annotations, constitute a new aggregate expression that is embodied in a new manifestation and that is carried by a material copy of the publication (a member of the F5 Item class).[32]

Figure 4.7 schematises the modelling of annotated printed books. The annotated exemplar is an instance of F4 Manifestation Singleton that *P128 carries* and aggregate expression (F2). The aggregate expression (F2) comprises (*P165 incorporates*) the text of the publication (F22 Self-Contained Expression) as well as any annotation (E90 Symbolic Object). The model allows for specifying annotation types: 'note' for manuscript annotations such as marginalia; 'markup' for symbols used to highlight a specific passage such as asterisks; 'highlight' for highlights; 'underline' for underlines; 'fold' for folded down page corners serving as bookmarks; 'bookmark' for

[31] Riva, Le Boeuf and Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*, 93.

[32] For a comprehensive model for aggregates see Riva, Le Boeuf and Žumer, 94.

bookmarks.

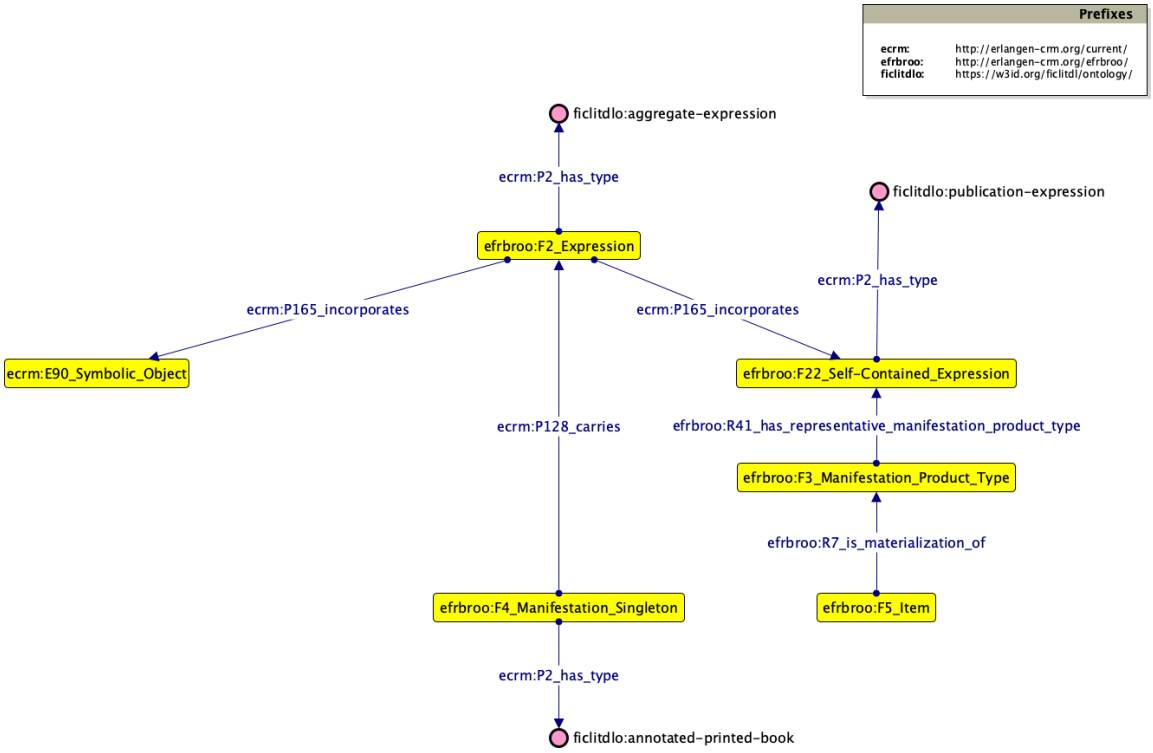

Figure 4.7: Annotated printed books.

The model described above can represent both publication texts and annotations. It does not, however, allow annotations and the passages to which they refer to be linked together, nor does it allow the author's activity of annotating to be described. This gap can be filled by integrating CIDOC-CRM/FRBRoo with the Open Annotation Data Model (OA). OA allows for linking annotations to the textual passages they refer to using the Annotation class. Each Annotation has a body (the text of the annotation) and a target (the text being annotated). The body of the annotation is an E90 Symbolic Object. The target of the annotation is also an E90 Symbolic Object that represents a delimited fragment of the publication text. Only annotations of type 'note' or 'markup' have a body.

Annotations result from an E65 Creation activity that involves an annotator (the activity of annotating is *P14 carried out by* an E39 Actor) and is characterised by a specific time span (the activity of annotating *P52 has time-span* E52 Time-Span).[33] The *P32 used general technique* property can be used to describe annotating as an

[33] The beginning and the end of the time span can be specified using the *P82a begin of the begin* and *P82b end of the end* properties, which take as a range a primitive value of type xsd:dateTime. In the RDF version of CIDOC-CRM, these properties replace the *P82 at some time within* property. Because they represent an interval, CIDOC-CRM considers it invalid to assign the same value to both *P82a* and *P82b*.

handwriting activity. The *P125 used object of type* property can be used to associate a writing tool, such as a pencil, with the annotation activity. To aid in identifying the exact location of annotations, the *prism:startingPage* and *prism:endingPage* datatype properties can be associated to the Annotation instance to specify the annotation page number.

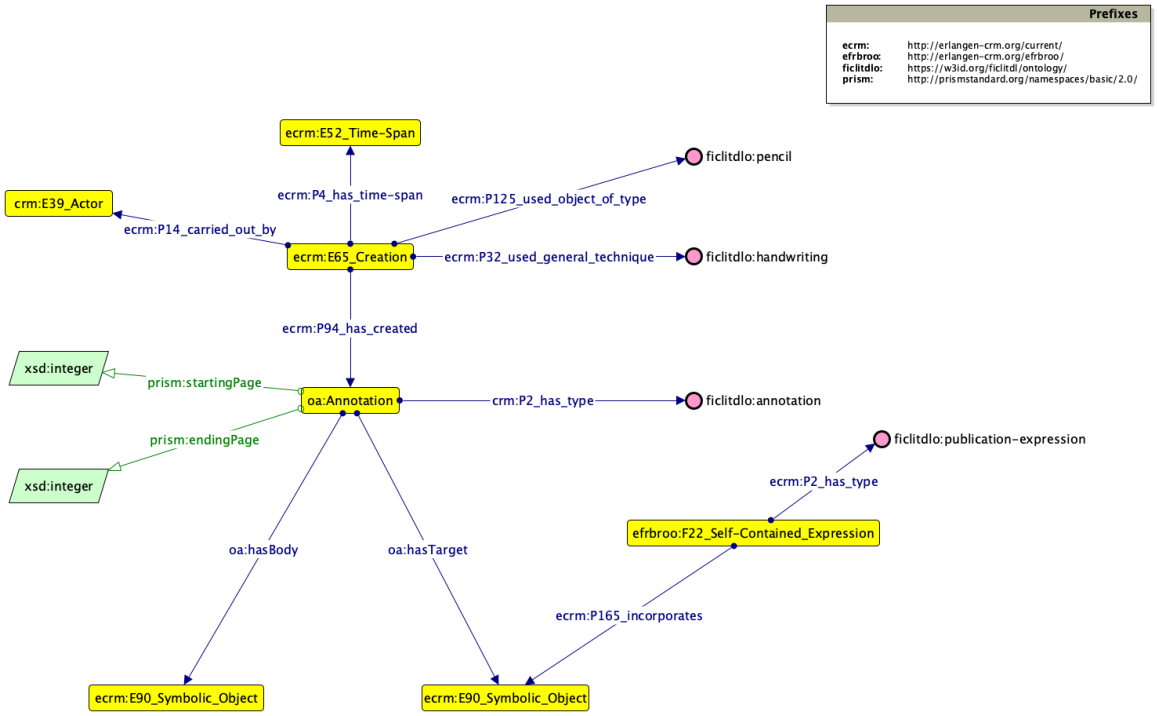

Figure 4.8: Annotation of a printed passage.

4.2.4 Correspondence

Giuseppe Raimondi's archive includes letters from over 800 correspondents. The Correspondence Series is further divided into files, each of which is dedicated to a different correspondent. Along with letters, the archival files also contain heterogeneous papers and material including postcards, manuscripts, typescripts, newspaper clippings, and photographs. To give readers an idea of the range of document types that make up Raimondi's correspondence, it is sufficient to list part of the contents of the correspondence files dedicated to the Italian poet Giuseppe Ungaretti, with whom Raimondi entertained a lifetime-long relationship after their meeting during World War One. In fact, this file contains, among other items, 14 letters, 2 responses by Raimondi, 4 postcards, 1 invitation card, 1 typescript article by Raimondi about Giuseppe Ungaretti, and 1 autograph by Ungaretti himself of an early draft of the

poem "La Madre".[34]

To represent complex scenarios in which, for example, a letter is sent with a draft manuscript or accompanies a copy of a printed publication, the representation of correspondence must first be viewed as a set of related events. Stadler, Illetschko, and Seifert, in proposing a conceptual model for the representation of correspondence in TEI-encoded documents, emphasise the importance of focusing not only on the material description of a piece of correspondence and the text of the carried message, but also on "the 'eventness' of the communicative act".[35] Such a communicative act comprehends at least three different events or activities: the creation of the text of the message (which produces, at the same time, a unique manifestation), the mailing of the piece of correspondence, and the event of receiving a letter. The ability to represent the communicative act as an event, allows for the representation of situations where multiple pieces of correspondence are sent as a unitary dispatch.

FRBRoo/LRMoo, combined with CIDOC-CRM, allows for the representation of the materiality of a piece of correspondence as well as the text of the message, but it does not feature a specific class for representing communicative acts. The same can be said about ICA's Records in Contexts, which proposes a simplistic representation of a piece of correspondence as a material document that has a sender and a receiver, while failing to place such a representation in the context of a communicative act. How, for instance, can it be asserted that that someone, upon receiving a letter, forwarded that letter to someone else, or that a record viewed as a piece of correspondence in one context becomes something different when placed in another context (one example is a book sent to Raimondi, a piece of correspondence, that annotations transform into a rich source of information about the writer's reading activity)? Given these considerations, an effective representation of correspondence should be able to answer at least the following questions:

1. What are all the pieces of correspondence in the archive, regardless of the series they belong to?
2. What is the list of all correspondence of a specific type?
3. What correspondence was sent by a specific sender?

[34] See BFICLIT FR.A CORRISP UNGAG.

[35] Peter Stadler, Marcel Illetschko and Sabine Seifert, 'Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>', Number: Issue 9 Publisher: Text Encoding Initiative Consortium, *Journal of the Text Encoding Initiative*, Issue 9 2016, 9, accessed 24 June 2021, https://doi.org/10.4000/jtei.1433, http://journals.openedition.org/jtei/1433.

4. What correspondence was sent from a specific location?
5. On which specific date was a piece of correspondence mailed (the date the item was mailed may differ from the date the message was written and the date the item was delivered)?
6. Which dispatches involved multiple items (e.g. a manuscript draft sent along with an accompanying letter)?

Figure 4.9 illustrates the representation of correspondence as a communicative act in the FICLIT Digital Library conceptual model. The new class Correspondence Activity is declared as a subclass of E7 Activity. The addition of a subclass to E7 as opposed to creating a new type of E7 Activity to represent correspondence activities is justified by the need to declare a set of specific properties for the individuals populating this class. This class contains activities that result in someone sending an instance of E22 Human-Made Object to someone else. Each Correspondence Activity can be associated with one or more item being sent (Correspondence Activity sent E22, F4 or F5),[36] a place of origin and a place of destination (Correspondence Activity had place of origin some E53 Place and had place of destination some E53 Place), a sender and a receiver (Correspondence Activity had sender E39 Actor and Correspondence Activity had receiver E39 Actor),[37] a date of dispatch and a date of delivery (Correspondence Activity had date of dispatch and Correspondence Activity had date of delivery; both ranges are XSD DateTime values).[38] The model distinguishes between the dispatch date (the date and time when the item was sent by the sender) and the delivery date (the date and time when the item was actually received by the receiver). The date of creation of the text of the message is recorded using F28 Expression Creation as the next sections discussing the representation of letters and postcards will show.

4.2.4.1 Letters

Letters and postcards are the main types of correspondence documents in Giuseppe Raimondi's archive. The conceptual model for letters, which is schematised in Figure 4.10, should be able to answer at least the following questions:

[36] The FICLITDL-O *sent* property is a subproperty of E5 Event *P12 occurred in the presence of* E77 Persistent Item.

[37] The FICLITDL-O *had sender* property is a subproperty of E7 Activity *P14 carried out by* E39 Actor; while had recipient is a subproperty of E7 Activity *P11 had participant* E39 Actor.

[38] The FICLITDL-O *had date of dispatch* and *had date of delivery* are subproperties of E2 Temporal Entity *P4 has time-span* E52 Time Span.

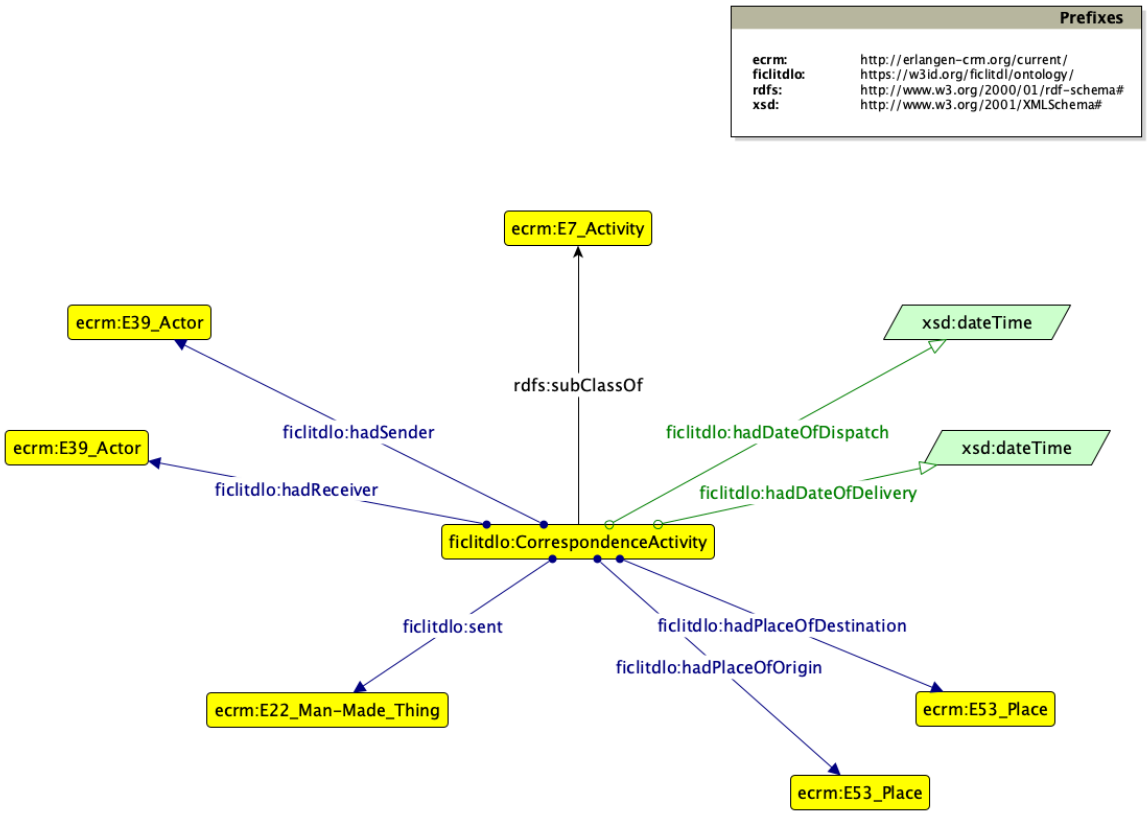

Figure 4.9: Correspondence as a communicative act.

1. What are all the letters in the archive?
2. Which letters were sent by a specific sender?
3. What are Raimondi's available responses?
4. Which letters were sent from a particular location?
5. Which letters were mailed on a specific date (considering the date the letter was mailed may differ from the date the message was written)?
6. Can you provide the facsimile and transcription of a specific letter?
7. Which letters contain a specific string of text?
8. What are the dimensions (height and width) of a specific letter?

A letter is a unique manifestation that carries text (F4 Manifestation Singleton *P128 carries* F22 Self-Contained Expression). Letters are attributed a document type using the *P2 has type* property (the value is 'letter', a subcategory of 'correspondence'). The creation of a letter is represented as an instance of F28 Expression Creation, which creates the text of the message and, at the same time, the letter as a physical document (F28 Expression Creation *R17 created* F22 Self-Contained Expres-

sion, and F28 Expression Creation *R18 created* F4 Manifestation Singleton).[39] The writing technique is specified using the *P32 used general technique* property, which takes as a domain F28 and as a range the FICLITDL-O class Technique Type (a subclass of E55 Type). The correspondence activity that provides context for a letter is linked to the letter itself via the FICLITDL-O property sent, which takes as a domain the Correspondence Activity and as a range the letter as a physical document.

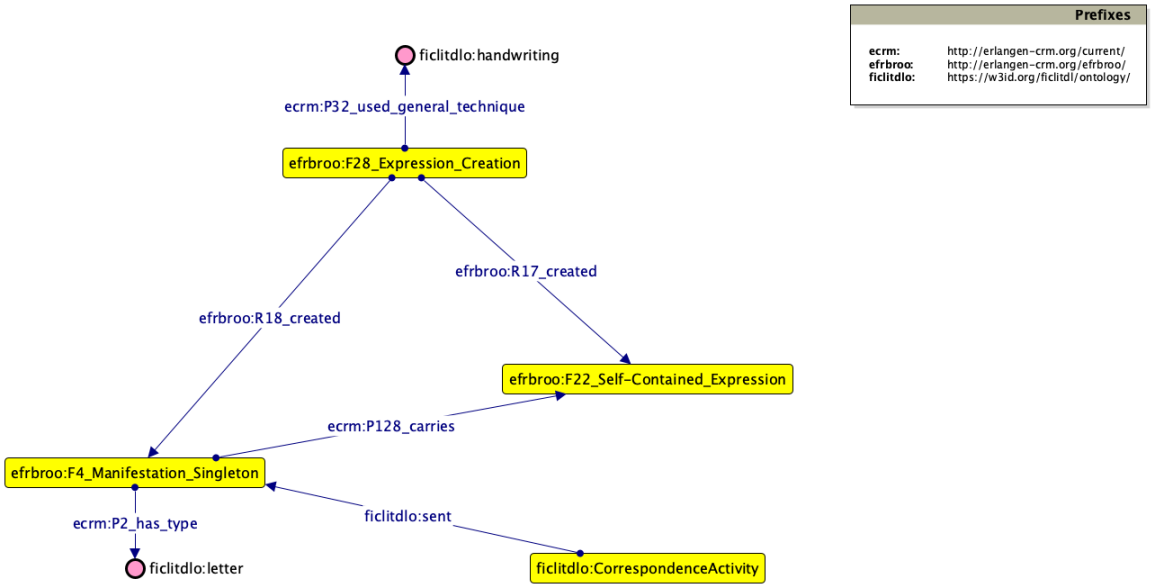

Figure 4.10: Letters.

4.2.4.2 Postcards

There are numerous approaches to representing the various types of correspondence (letters, postcards, telegrams, invitations, but also books, manuscripts, and so on), and each must convey the object's inherent complexity. Postcards, as highly complex pieces of correspondence, demonstrate this point clearly. These double-sided documents may include text as well as photographs or illustrations, which are frequently representations of places or works of art and contribute to the overall understanding of the message's context. Text on a postcard has a specific structure (a sender, a message, an addressee, and so on), and illustrations can also contain text, making the structure of the representation of postcards similar to that of a Chinese-box.

[39] In LRMoo letters too will be represented as individuals of the F5 Item class (see Chryssoula Bekiari et al., eds., *LRMoo (formerly FRBRoo) object-oriented definition and mapping from IFLA LRM (version 0.7)*, June 2021).

A single postcard frequently involves multiple creators: the author of the message (who may or may not be the sender), the author of the photograph or illustration, the author of any text that may be incorporated into the illustration, and so on. The archival representation should be capable of answering at least the following questions:

1. What are all the postcards in the archive?
2. Which postcards in the archive are illustrated?
3. Which postcards were sent by a specific sender?
4. Which postcards were sent from a particular location?
5. Which postcards were mailed on a specific date (considering the date the postcard was mailed may differ from the date the postcard text was written)?
6. Can you provide the facsimile and transcription of a specific postcard?
7. Which postcards contain a specific string of text?
8. What are the dimensions (height and width) of a specific postcard?

Figure 4.11 shows the representation of postcards in the FICLITDL conceptual model. Postcards are members of F4 Manifestation Singleton (if the postcard has not been compiled, F5 Item would be the appropriate class of belonging). A postcard can contain both text and visual items, which are represented as instances of F22 Self-Contained Expression. One postcard may carry multiple expressions (this is the case of illustrated postcards which contain text on the back side and visual items on the front side; see Section 4.4 for a more detailed discussion of content description).

Postcards as document types are categorised as correspondence artefacts according to the following taxonomy, implemented as part of FICLITDL-O. The taxonomy, which can be expanded to meet other projects' needs, is based on and aligned with the Getty Art and Architecture Thesaurus (AAT) classification of postcards:[40]

- correspondence
  - postcard
    - picture postcard
      - photographic postcard

[40] See the entry for 'Correspondence artefacts' in Getty AAT at http://vocab.getty.edu/page/aat/300207872.

- illustrated postcard
* postal card

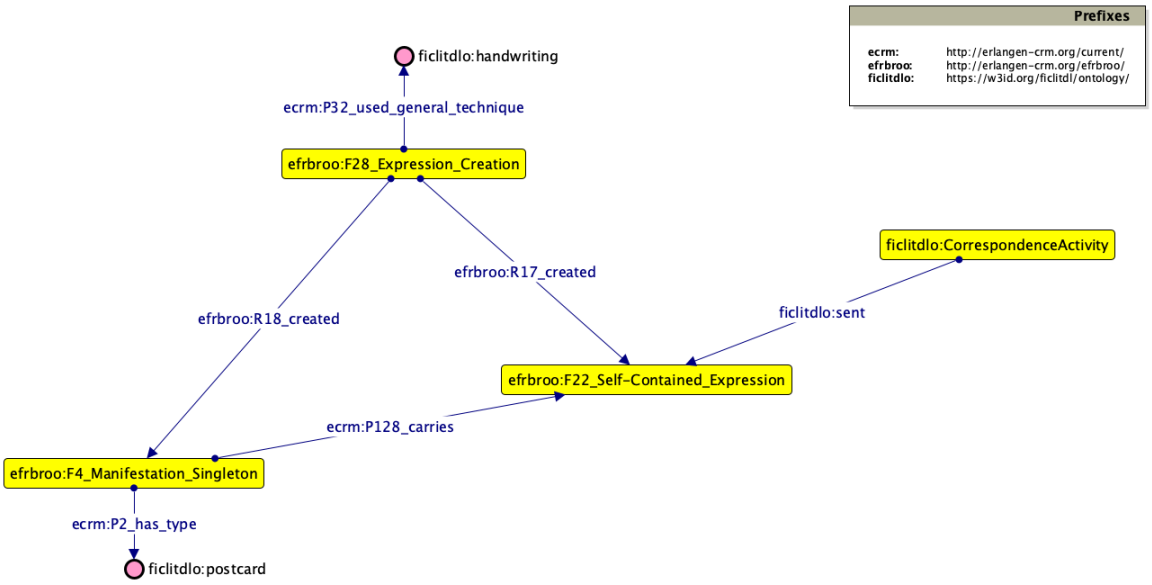

Figure 4.11: Postcards.

4.2.5 Albums

The Giuseppe Raimondi Archive includes a total of 14 albums featuring thematic collections assembled by Raimondi on the occasion of the exhibition "Giuseppe Raimondi fra poeti e pittori" (Bologna, Museo Civico, 28 May-30 June 1977).[41] Each collection gathers existing documents, particularly correspondence, related to Raimondi's personal relationships with fellow writers, artists, and intellectuals from Italy and abroad. The Getty AAT defines albums as follows:

> Books with blank pages used for assembling and presenting a collection, or any unpublished sets of pages that are bound or loose-leaf and have other materials affixed to them or writing or other images made on them. The volume may have been constructed with blank pages that were intended to have materials affixed, or it may have been assembled after having had materials affixed to separate pages.[42]

[41] See Giuseppe Raimondi and Clemente Mazzotta, *Giuseppe Raimondi fra poeti e pittori: mostra di carteggi* (Edizioni Alfa, 1977).
[42] Entry for 'Albums' in Getty AAT, http://vocab.getty.edu/page/aat/300026690.

Albums, as the above definition suggests, are complex objects composed of multiple pieces having heterogeneous characteristics. Albums present similar modelling challenges to notebooks. In particular, one collection can span multiple albums, or one album can contain multiple collections. Moreover, they can carry text (a photograph, for example, may be accompanied by a caption), and can be containers for other objects. For example, the collection dedicated to the artist and writer Filippo De Pisis (Ferrara, 1896-1956) encompasses two albums and includes different records arranged chronologically that retrace the intellectual exchange between De Pisis and Raimondi: 15 photographs, 41 original drawings, 27 reproductions of drawings, 5 manuscripts, 1 book, 9 between invitation notes and leaflets, 1 poster, 87 between letters and postcards.[43] The heterogeneous nature of the records of the albums provides an interesting modelling test case because it mimics, on a smaller scale, that of the entire archive. Moreover, like personal archives, albums feature collections of documents which were intentionally and purposefully aggregated by a creator. In LRM, such a process of aggregation is considered as an intellectual or artistic endeavour resulting in the creation of a new aggregating work whose essence are the adopted criteria of selection and arrangement.[44] In this context, the role of the creator of an album is analogous to that of an editor preparing an aggregate publication, e.g. an anthology, by evaluating, connecting, and assembling texts that were created independently.

A representation of albums should be able to answer at least the following questions:

1. What items are included in the album?
2. What is the theme of the album?
3. When was the items in the album created?
4. What is the sequence of the items in the album?
5. What record set did the items originate from?
6. Does the album store any other documents?
7. When was the album created?
8. Which item comes first in sequence?

[43] BFICLIT, FR.A ALBUM DEPISI.

[44] Riva, Le Boeuf and Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*, 94.

Figure 4.12 illustrates the representation of albums. In LRM terms it is possible to assert that an album or, better, a collection, realises an aggregate work, i.e. that a collection is an instance of F22 Self-Contained Expression which *R3i realises* an F1 Work. The aggregate expression, in turn, is *P128i carried by* an instance of F4 Manifestation Singleton. Such a manifestation can either correspond to a single album or encompass a set of multiple albums (in the De Pisis example, two). As with notebooks, indeed, if a collection (or expression) spans multiple albums, the set of albums is an instance of the F4 Manifestation Singleton class that is composed of multiple objects (F4 Manifestation Singleton *P46 is composed of* E22 Human-Made Object). On the other hand, when the boundary of the aggregating expression coincides with that of the album, the album is an instance of both the F4 Manifestation Singleton and E22 Human-Made Object classes. Each physical album, an E22 Human-Made Object, is further categorised as an archival unit of type 'file' and as an individual document of type 'album' through the *P2 has type* object property.

The possibility of representing albums on the work, expression, and physical levels allows for the description of the album itself as a newly created expression having attributes of its own (e.g. a title, a date, a creator, etc.). The creation of an album is represented using the F28 Expression Creation class. Such an activity of creation corresponds to the transformation of a set of distinct heterogeneous entities (e.g. a photograph, a letter, a drawing, etc.) into an aggregate expression (F28 Expression Creation *R17 created* F22 Self-Contained Expression). At the same time, the expression creation generates the album itself as a physical object (F28 Expression *R18 created* F4 Manifestation Singleton).

As seen at the beginning of this section, in addition to carrying texts or expressions in the same way that a notebook would, albums can be used as supports for physically aggregating individual documents (this is actually their primary use in Giuseppe Raimondi's archive). As a result, whereas in LRM "the aggregation happens on the expression level, because only expressions can be combined (or aggregated)," aggregation can also happen on the item level in the FICLITDL conceptual model. When it comes to album representation, this is the main difference between the two models.

The *P46 is composed of* property is used to model the relationship between the album and the individual documents it aggregates. The album and the documents, which are the domain and range of the relation, are both instances of E22. It should be noted that the *P46* property is preferred over the *P198 holds or supports* relation-

ships because the latter implies that the held or supported document is not a conceptual component of the container or support. The individual documents that compose an album, as well as their interconnections, can then be described separately, depending on their document type.

Because the arrangement of the documents in an album is the essence of an aggregating work along with selection, users may be interested in retrieving the exact sequence in which documents appear within an album. This is done through two object properties offered by the the Sequence Ontology Design Pattern (Sequence ODP): *seq:follows* and *seq:precedes*. In particular, an E22 Human-Made Object either *seq:follows* or *seq:precedes* another E22 Human-Made Object.[45]

Individual documents in Giuseppe Raimondi's albums hold a relationship not only with the album in which they have been collected, but also with the record set to which they originally belong (for example, some letter that were originally part of Raimondi's correspondence with Giuseppe Ungaretti were moved to the Ungaretti Album in the 1970s). When exploring the record set of which the item was originally part, users should be able to find that the item has been moved to constitute the album.

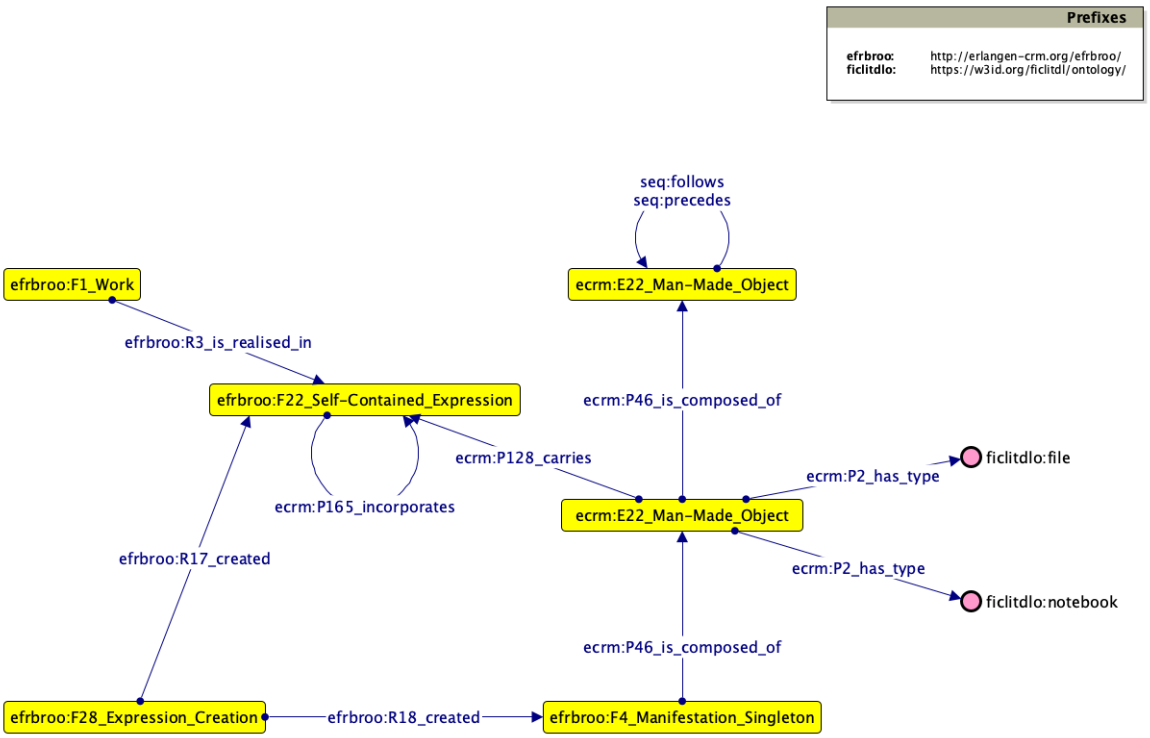

Figure 4.12: Albums.

[45] Sequence ODP is available from http://ontologydesignpatterns.org/cp/owl/sequence.owl.

4.3 Records as Physical Objects

Any record in a paper-based archive is inherently shaped by its materiality. The representation of archival records as physical objects is logically separate from that of archival records as texts, although the two dimensions closely determine one another. As previously mentioned, the key classes used for representing the materiality of manuscripts and other items are E22 Human-Made Object from ICOM's CIDOC-CRM and F4 Manifestation Singleton from IFLA's FRBRoo (in LRMoo, as pointed out above, F4 will be substituted by F5 Item). E22 Human-Made Object is the subject of all relations describing the physicality of archival documents. An E22 Human-Made Object is an individual item in the archive, as opposed to an F4 Manifestation Singleton, which can comprehend multiple items if they realise one single work (e.g. a text spanning multiple notebooks or a thematic collection of items occupying multiple albums), though most of the time the same item is instantiated as both an E22 and an F4. The *P128 carries* property, which takes the physical document (F4) as a domain and the text that such a document contains as a range, is used to connect physical archival documents to the texts they bear (F22). The sections that follow each illustrate a different aspect of physical record description.

4.3.1 Document Type

Records are assigned a document type via the *P2 has type* property, which takes as a domain a member of E22 Human-Made Object, the physical individual item, and as a range a member of the Document Type class, which is declared as a subclass of E55 Type within the FICLIT Digital Library Ontology namespace. Members of the Document Type class represent document type categories that are frequently found in the FICLIT Department's paper-based writers' archives. The following taxonomy is used to classify document types. Given the significant heterogeneity of document types that populate the archives of twentieth-century writers, the provided classification of document types is inevitably partial and will require ad hoc integration in order to become applicable to other archival collections:

- book
  - album
  - notebook

  - printed volume
- card
  - postcard
    - picture postcard
      - photographic postcard
      - illustrated postcard
    - postal card
- clipping
  - newspaper clipping
- letter
- manuscript
- typescript
- telegram

The Simple Knowledge Organization System (SKOS) ontology and, in particular, the has close match and has exact match properties are employed to match each document type to a related concept in the Getty Art and Architecture Thesaurus (Getty AAT). The property *P127 has broader term (has narrower term)* allows an instance of Document type to be associated with another instance of Document Type that has a broader meaning (e.g., 'postal card' *P127 has broader term* 'postcard'). Labels and descriptions in English and Italian are provided for each document type. The goal of this classification is to allow archive users to filter specific document types, possibly by including additional filters such as production techniques (e.g., handwriting and typewriting) and text genres (article, short story, essay, review, etc.). Figure 4.13 shows the main concepts and relationships in the FICLITDL-O document type scheme. It is possible to further describe the classification of physical documents by means of the E17 Type Assignment class (this is discussed in more details in Section 4.4.1).

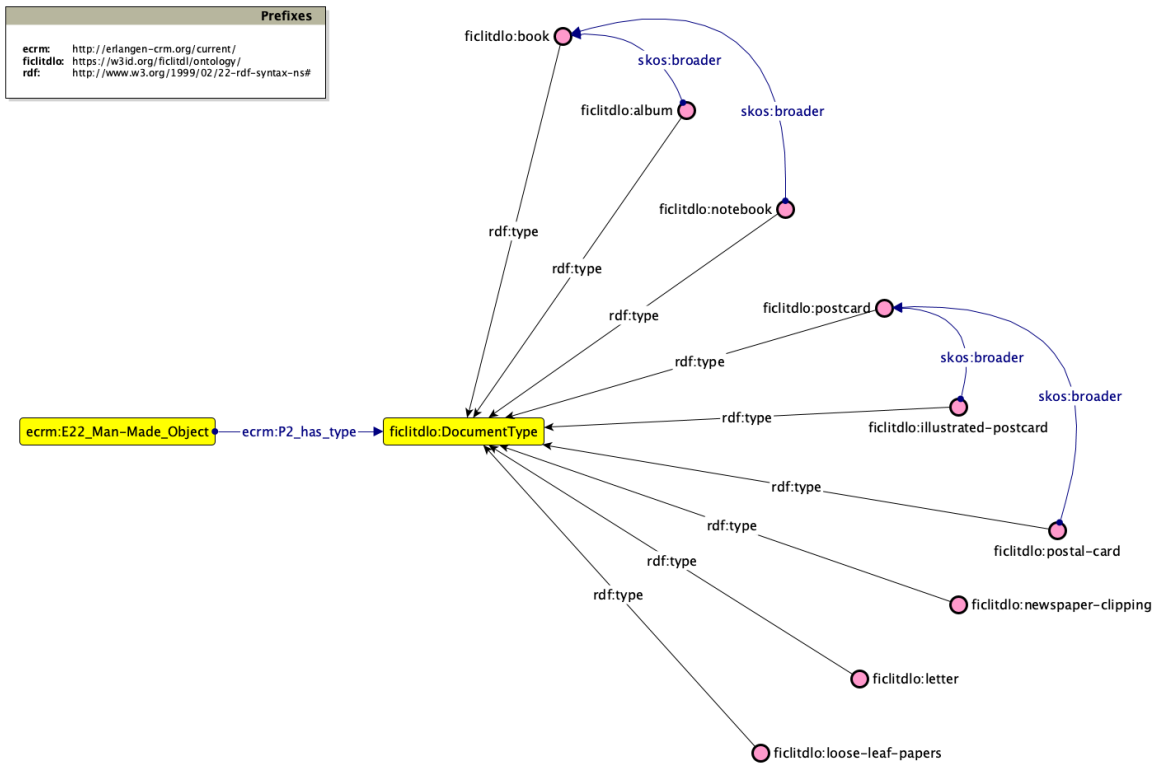

Figure 4.13: Document types.

4.3.2 Dimensions and Extent

In the FICLITDL conceptual model, the description of the relationship between a physical document and one of its three dimensions can be entirely represented using CIDOC-CRM by asserting that a document (E22) *P43 has dimension* E54 Dimension, as illustrated in Figure 4.14. Dimensions can be further described by a type of dimension, a numerical value, and unit of measurement for the numerical value. The Dimension Type class is declared within the FICLITDL conceptual model as a subclass of E55 Type to represent various types of dimensions. A notebook with a height of 22 cm, for example, would be described as an object that has dimension '22' (E54 Dimension *P90 has value* an xsd:decimal). Such a dimension would have type 'height' and measurement unit 'centimetre' (E54 Dimension *P2 has type* Dimension Type; E54 Dimension *P91 has unit* E58 Measurement Unit). Individuals of the Dimension Type class are aligned to the corresponding concepts in the Getty AAT.

In addition to size, the physical description of bound documents such as notebooks and albums includes the extent seen as the total number of pages that compose the object (a page corresponds to one side of a leaf of paper, either the recto or verso). Extent is represented as an instance of E54 Dimension with a numerical value that corresponds to the total number of pages characterising the document,

the measurement unit (the page), and a dimension type which in this case corresponds to the concept of 'extent'. The conceptual model distinguishes between the total extent of a document and the extent of the text, that is, the number of text-bearing pages in the document (see Section 4.4).

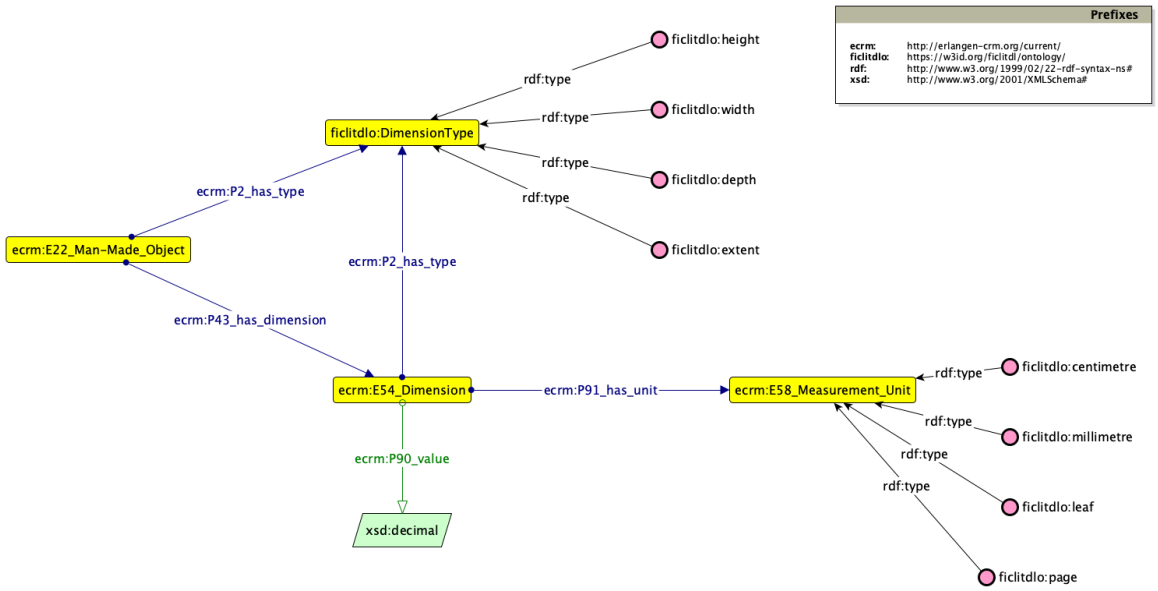

Figure 4.14: Dimensions and extent.

4.3.3 Facsimiles

One crucial feature of the FICLITDL conceptual model is that each physical document is tied to a IIIF manifest (an instance of E73 Information Object, which is a URL at which the manifest is available online) pointing to large-scale, zoomable facsimiles of the object.[46] The physical document (E22) is linked to the manifest using the FICLITDL-O has IIIF manifest property, which is declared as a subclass of *P67 refers to* (see Figure 4.15).[47]. Simultaneously, as detailed in Section 4.4, each expression can be linked to one or more transcriptions, which can be encoded in XML/TEI. The direct linking of facsimiles and transcriptions to the archival knowledge base supports further research and can act as a basis for the production of new digital scholarly editions of the texts of the archive.[48]

[46] For an overview of the International Image Interoperability Framework (IIIF) see https://iiif.io.
[47] The decision to create a new property is motivated by the limitation of not being able to use P138. This is due to the fact that the domain of P138 represents is restricted to visual items only, but a IIIF Manifest is not. As a result, I chose to declare a sibling relationship with P138, which is a subpropery of P67 as the newly introduced property.
[48] On the topic see Section 1.6 of this dissertation.

4.3.4 Identifying and Locating Records

Figure 4.15 schematises the representation of identifiers in the FICLITDL conceptual model. All physical records can be given an inventory number, a shelfmark at the fonds, series, file, or item level to allow for identifying and locating records unambiguously. The current location of each document in the archive is also provided (E22 Human-Made Object *P55 has current location* E53 Place (a place can be further described by the *P89 falls within*, and the *P76 has contact point* properties to specify a broader area within which the place is located, as well as a contact point that provides access to the holding institution; in the case of the FICLIT collection of archives this is the Ezio Raimondi Library). Places are associated with Geonames coordinates.[49]

Additionally, as the finding aid is transformed into a knowledge base, all entities in the graph are assigned a unique identifier taking the form of a permalink.[50] All identifiers in the FICLITDL conceptual model are members of the CIDOC-CRM E42 Identifier class and are linked to the record they identify via the *P1 is identified by* property. Identifiers are individuals of the Identifier Type class, which is declared as a subclass of E55 Type in the FICLITDL ontology namespace to differentiate between the various types of identifiers available for each record (current values are 'inventory number', 'shelfmark', and 'permalink').

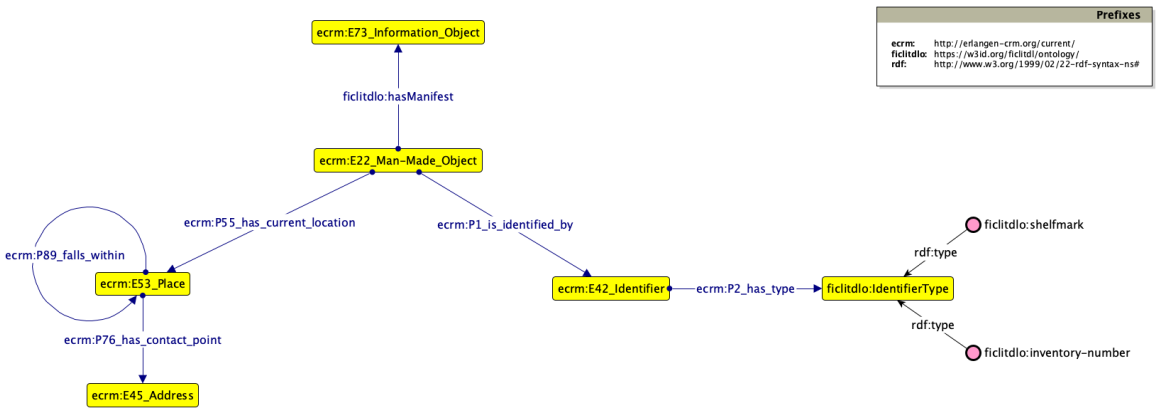

Figure 4.15: Facsimiles and identifiers.

[49] Geonames, https://www.geonames.org.
[50] On the URI scheme followed in the project, see Section 3.2.3 of this dissertation.

4.3.5 Legal and Physical Custody

As archival literature emphasises, one essential piece of information that finding aids must provide users is information about the history of physical custody of collections. Figure 4.16 illustrates the main classes and relationships employed represent custody and rights information in the FICLITDL conceptual model. The relationship between a physical document and the person or organisation that maintains it can be represented using CIDOC-CRM (E22 Human-Made Object *P50 has current keeper* E39 Actor).

It is possible and desirable to provide a more detailed representation of an object's history of custody by describing any change of custody event that affected the object over time. This can be accomplished by using the E10 Transfer of custody class to represent events of custody change. Each transfer of custody is attributed an actor who surrendered custody of the object (E10 Transfer of Custody *P28 custody received by* E39 Actor), an actor who receives custody of the object (E10 Transfer of Custody *P29 custody surrendered by* E39 Actor), and the object whose custody is being transferred (E10 Transfer of Custody *P30 transferred custody of* E22 Human-Made Object). Finally, the date of the change of custody event is specified (E10 Transfer of Custody *P4 has time-span* E52 Time-Span). In the case of Giuseppe Raimondi's archive, custody passed from the creator to his heirs to the university library, resulting in a sequence of two distinct transfers of custody.

Custody information is accompanied by ownership information using the relationship E22 Human-Made Object *P52 has current owner* E39 Actor (Giuseppe Raimondi's archive is kept at the Ezio Raimondi's library of the University of Bologna, but it is legally owned by the Istituto per i beni artistici, culturali e naturali della Regione Emilia-Romagna).[51]

4.3.6 Rights and Licences

Each archival document is also instantiated as a member of the E72 Legal Object class, which includes material and immaterial objects "to which instances of E30 Right, such as the right of ownership or use, can be applied".[52] In the context of

[51] The Istituto per i beni artistici, culturali e naturali della Regione Emilia-Romagna (IBC) was renamed Settore Patrimonio Culturale in 2020 (see https://patrimonioculturale.regione.emilia-romagna.it). The RDF descriptions of all organisations are aligned to VIAF and Wikidata authority records and feature direct links to the corresponding organisation homepage.

[52] Bekiari et al., *Definition of the CIDOC Conceptual Reference Model (Version 7.1.2)*, 98.

research and reuse, an explicit declaration of rights and licencing becomes critical. A significant advantage of using a multidimensional approach to archival description is the ability to specify differentiated legal information for physical documents, their content, facsimiles and transcriptions, and any record component in general. E72 Legal Object *P104 is subject to* E30 Right is the relationship used to specify rights information (see Figure 4.16). All rights are reserved for items in Giuseppe Raimondi's archive, but other types of licencing information, such as Creative Commons licences, which provide an RDF schema, can also be adopted.[53]

For each record and record set in the archival knowledge base, instructions for citing the record are provided using the property *dcterms:bibliographicCitation* (for example, for one of the notebooks, the preferred form of citation would be "Raimondi, Giuseppe. 1963. Quaderno manoscritto. Fondo Giuseppe Raimondi, Biblioteca Ezio Raimondi, Università di Bologna. RDq 179, FR.A QUADERNI.1 1963 10").

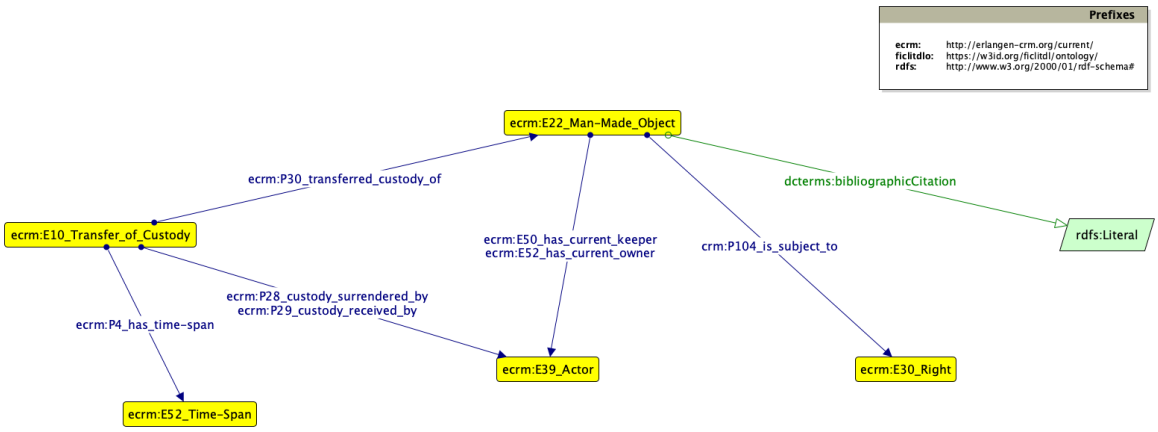

Figure 4.16: Custody and legal information.

4.4 Records as Texts

As previously seen and outlined in Figure 4.17, texts and images are modelled using FRBRoo as instances of F22 Self-Contained Expression, which "comprises the intellectual or artistic realisations of Works in the form of identifiable immaterial objects, such as texts, poems, jokes, musical or choreographic notations, movement pattern, sound pattern, images, multimedia objects, or any combination of such forms" (in LRMoo the class for texts will simply change to F2 Expression).[54] In order to distinguish between textual and visual contents, expressions can be simultaneously

[53] On using Creative Commons to describe licences in RDF see https://creativecommons.org/ns.
[54] Bekiari et al., *LRMoo (formerly FRBRoo) object-oriented definition and mapping from IFLA LRM (version 0.7)*, 19.

instantiated as members of the E22 Linguistic Object class (for texts in the strict sense) or as members of the E36 Visual Item class (for images). The linking between documents as physical objects and the texts they bear is realised through the *P128 carries* property from CIDOC-CRM (the physical object, an F4 that can be made up of multiple E22, carries the text, which is an F22).

Often times, the physical document bears multiple expressions. This is the case of notebooks as well as illustrated postcards. For example, a notebook can contain distinct texts and an illustrated postcard can contain texts on one side and multiple images on the other. In such cases it is possible to assert multiple times, one for each carried expression, that the physical document carries an expression. Alternatively, the content carried by a physical document may be treated as one single expression that incorporates other expressions (F22 *P165 incorporates* F22).

Moving deeper into the text, it is possible to describe circumscribed portions of text or images using the E90 Symbolic Object class.[55] Like self-contained expressions, members of E90 too are also instantiated as members of E33 Linguistic Object or E36 Visual Item depending on their nature. Textual or image fragments are linked to the texts or images to which they belong using the *P106 is composed of* property, which takes as a domain an instance of F22 and as a range an instance of E90.[56]

The ability to represent textual fragments opens up new possibilities for archival practise by allowing arbitrary portions of text to become the subject of additional archival description and editorial interpretation. For example, by viewing archives as webs of intertwined texts rather than just documents, one can represent authorial changes from one version of a text to another or identify cases of text reuse between different works as Section 4.5.3 will further detail.

Expressions can be further described using the *P72 has language* property to document the language of the text. Languages are instantiated as members of the E56 Language class, a subclass of E55 Type (E33 Linguistic Object *P72 has language* E56 Language). Individuals of E56 Language are taken from the Library of Congress ISO 639-2, which assigns each language a permanent URI ending with a three-letter code representing the language (for example, the URI for English is http://id.loc.gov/vocabulary/iso639-2/eng).[57] Like physical documents, expressions

[55] LRMoo deprecates the use of F23 Expression Fragment as it aims to simplify the process by promoting the use of CIDOC-CRM terms instead of introducing new ones. Specifically, when dealing with fragments of texts, LRMoo suggests employing E90 as it is already considered suitable for this purpose.

[56] In LRMoo, it will be possible to use the more specific property *R15 has fragment* with range E90.

[57] The FICLITDL-O model adopts ISO 639-2 and later versions language codes for which the Lib-

too are also members of the E73 Legal Object class. As such, it is possible to express rights and licencing information that specifically apply to content.

In addition, it is possible to specify an extent for each text. The extent of a text is represented as an instance of E54 Dimension with a value (*P90 has value*) that corresponds to the total number of words composing the text, a type (*P2 has type*) that specifies the type of extent being represented ('text length'), and a measurement unit (*P91 has unit*) that specifies the unit of measurement ('word count'). Other measurement units, such as character count, can be added to the model.

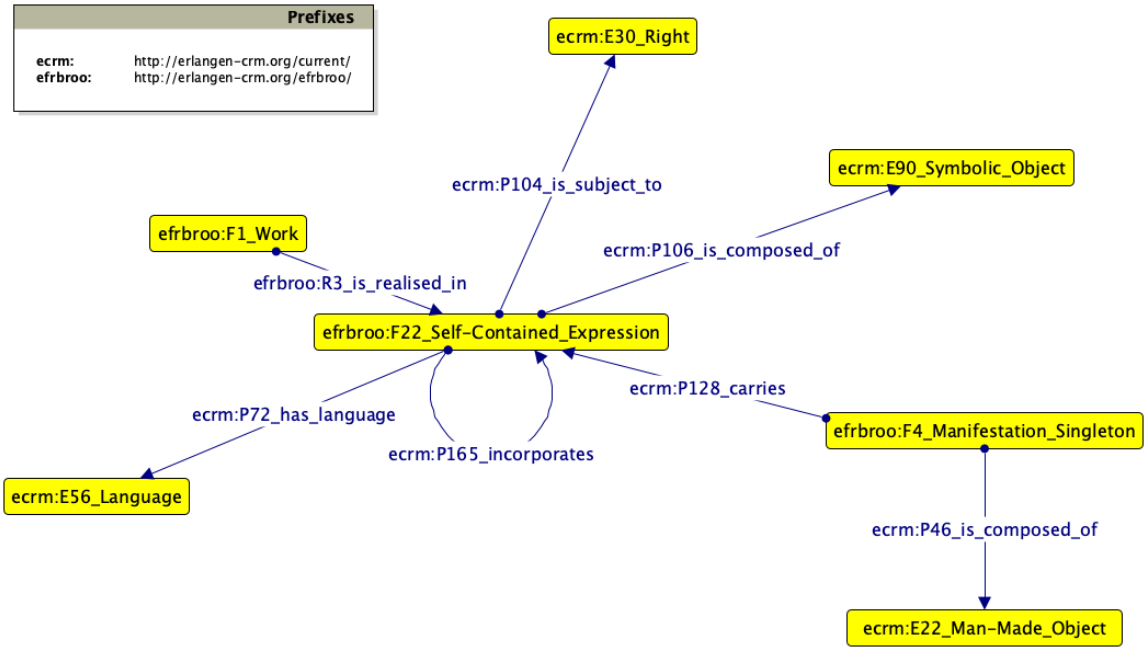

Figure 4.17: Records as texts.

4.4.1 Text Type

Texts, like physical documents, are assigned a category in order to facilitate automatic processing and querying. Texts can be classified according to their type. The Text Type class is declared as subclass of E55 Type within the FICLITDL-O namespace. Possible text types include:

- article
  - journal article

rary of Congress has been designated the registry authority. In the context of its Linked Data Service, the Library of Congress offers a SKOS Concept Scheme of ISO 639-1, ISO 639-2, and ISO 639-5 language codes.

  - newspaper article

- lecture notes
- literary work
  1. short story
  2. novel
  3. poem
- essay
- preface
- homage
- correspondence
- invitation

The above classification covers the bare minimum of text types needed to describe Giuseppe Raimondi's archival collection, but it can be expanded as more types in the FICLIT Departments' writers' archives are discovered. Text types, like document types, are aligned with related concepts in the Getty AAT (see Figure 4.18).

Classification of documents and texts is the result of critical interpretation, particularly when working with draft manuscripts. The E17 Type Assignment class from CIDOC-CRM provides further details about the act of classifying entities in the dataset, including archival records. The E17 class is defined as an activity that assigns a Document or Text Type to an object, either the physical document or the text it contains. The person responsible for the classification and the time when the classification was completed can also be specified, as shown in Figure 4.19.

4.4.2 Title Attribution

The *P102 has title* property, which takes as a domain F22 and as a range E35 Title, is used to attribute titles to the texts. The attribution of a title to a text as well as authorship attribution are described as an E13 Attribute Assignment activity. In the case of title attribution, the attribute assignment activity assigns an instance of E35 Title (the attribute) to the text (E13 *P140 assigned attribute to* F22; E13 *P141 assigned*

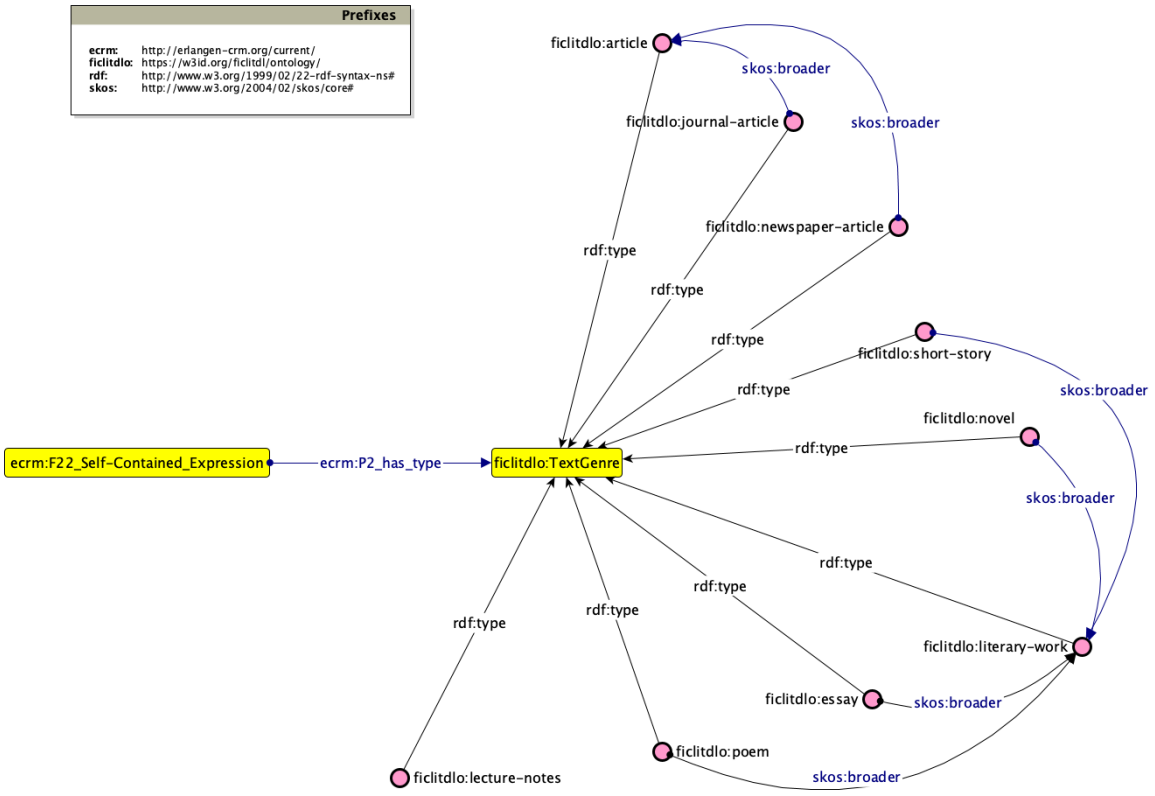

Figure 4.18: Text type.

E35). The agent responsible for the title attribution is described using the *P14 carried out by* property which takes as a domain the E13 and as a range an E39 Actor. It is possible to specify the date when the title attribution took place using *P4 has time-span* E52 Time-Span (see Figure 4.20).[58]

4.4.3 Text Transcription

Each expression in the archival knowledge base can be linked to one or more transcriptions, which can be encoded in XML/TEI. The direct linking of facsimiles and transcriptions to the archival knowledge base facilitates further research and can serve as the foundation for the creation of new digital scholarly editions of the texts of the archive. In particular, using the Open Annotation standard, users can annotate facsimiles and transcriptions, further augmenting the archival knowledge base.[59]

Figure 4.21 displays the modelling of text transcriptions. Transcriptions are represented as individuals of the E73 Information Object class. The format of each

[58] Time spans are further parsed into xsd:dateTime primitives to facilitate automatic processing and querying. For each time span an interval start date and an interval end date are specified using the Time Interval Ontology. Values are xsd:dateTime values.

[59] Refer back to Chapter 1 for a more detailed discussion on the relationship between archive and edition.

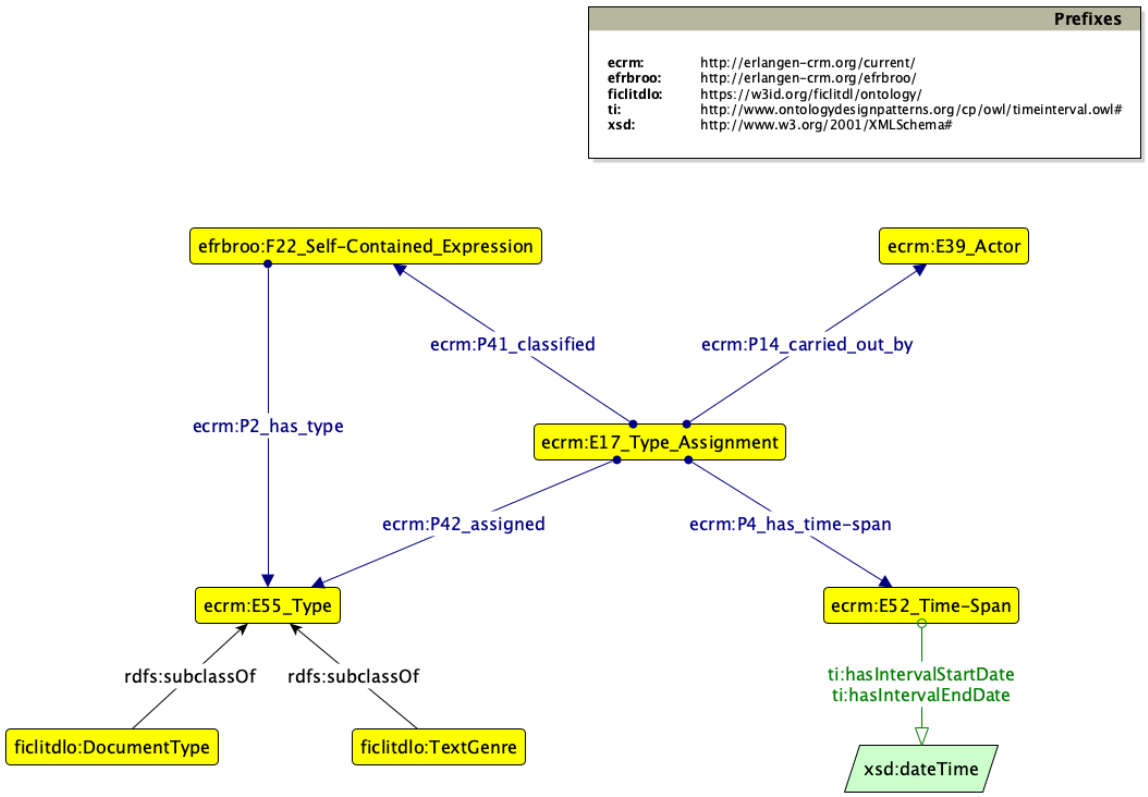

Figure 4.19: Type attribution.

transcription is expressed through the relationship *dcterms:format*, which takes the transcription (E73) as a domain and an individual of the ficlitdlo:Format class as a range (for XML/TEI transcriptions this is 'xml-tei'). The *ficlitdlo:hasTranscription* property connects texts (F2) to their corresponding transcriptions (E73).

In addition to linking texts (F2) and transcriptions (E73), specific fragments of texts represented as instances of E90 Symbolic Object can be linked to their specific locations in the transcription. There are various mechanisms for connecting an entity in an RDF graph (in this case, a textual fragment) and a location in an XML/TEI file. One possibility is to provide unique identifiers (xml:id attributes) for each XML element in the transcription. The elements to which the identifiers are assigned may not markup the exact location of a textual fragment but readers may reconstruct such a location autonomously, once they have identified the relevant line as they would do with a printed scholarly edition.

Xml:ids attributes can then be used to specify the targets of annotations on text transcriptions. Annotations on text transcriptions are achieved using the Open Annotation data model introduced in Section 4.1.4. Each annotation is constituted of a body that represents the content of the annotation and a target that represents the text for which the annotation is provided. The location of the target of an annota-

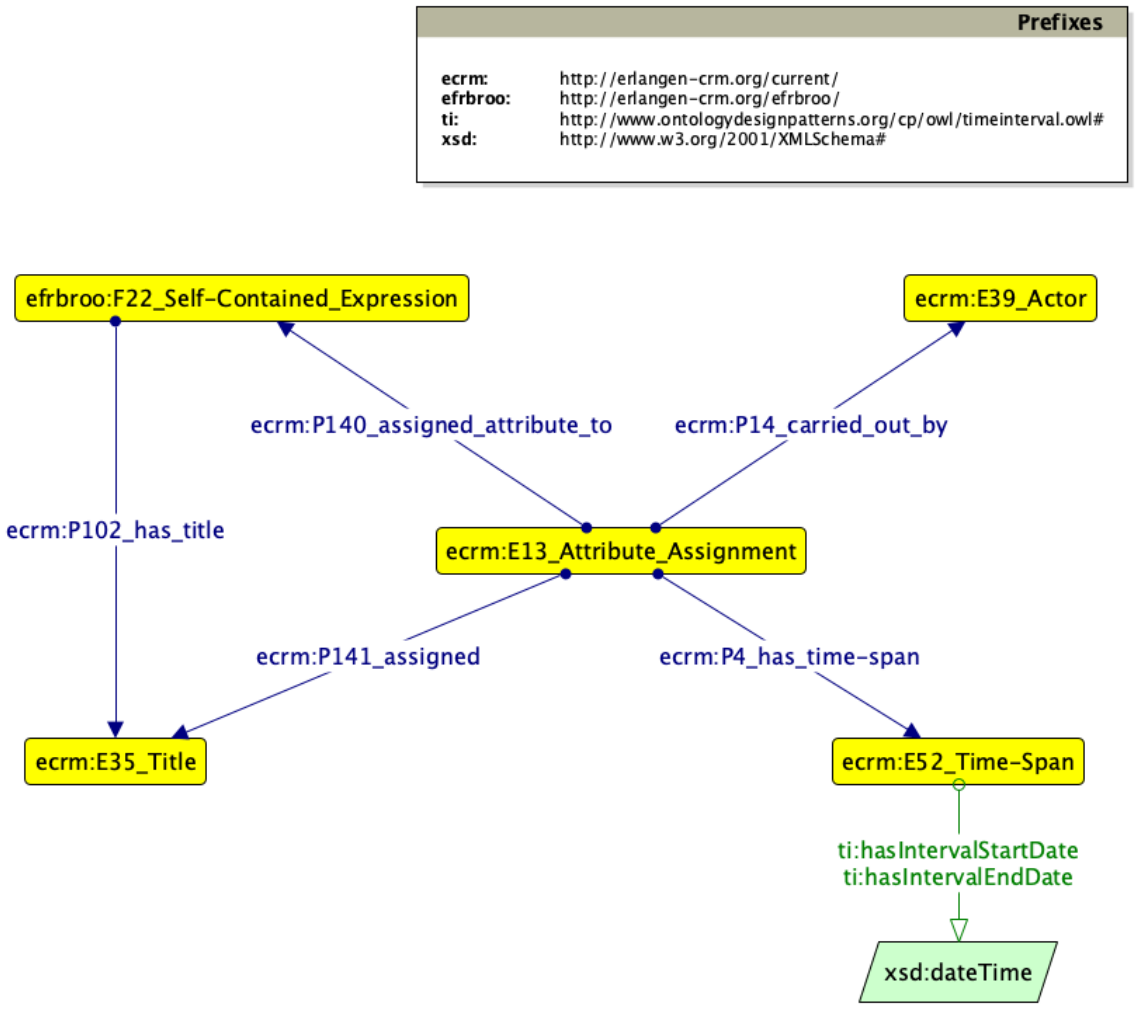

Figure 4.20: Title attribution.

tion can be expressed using an XPath selector that specifies the exact position of the element that markup the text that is the target of the annotation in the XML/TEI hierarchy.

4.4.4 References to Named Entities

References to named entities in records, such as people, places, and organisations, can serve as alternative access points to an archival collection. The *P67 refers to* property can be used to model the relationship between a text and a named entity (an E21 Person, an E53 Place, an E40 Legal Body, or even a another text). The domain of the relationship can be an F22, i.e. the entire text, or a specific fragment within that text. The fragment can be linked to the location in the transcription where it appears.

Persons, places and organisations are further described using existing proper-

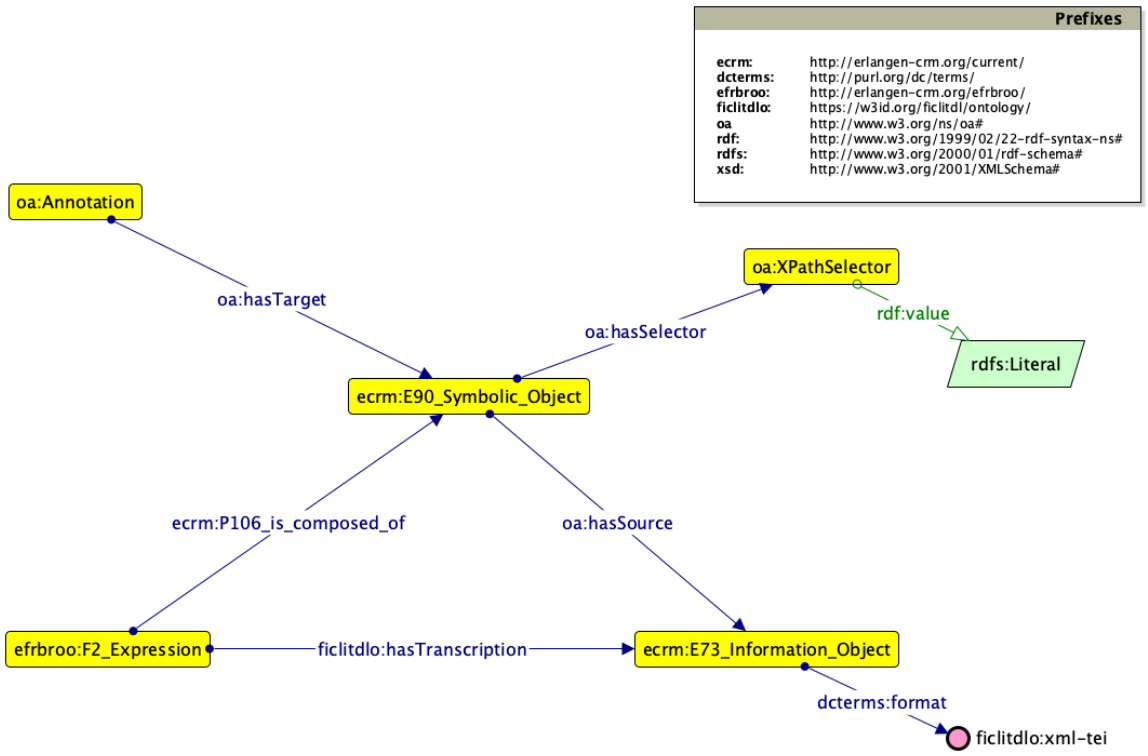

Figure 4.21: Text transcription.

ties from CIDOC-CRM and other data models. Figure 4.22 illustrates some among different possible relationships that can be used to describe persons, places and organisations. All named references have standardised labels. Using the *owl:sameAs* property, all named references are linked to existing authority records such as VIAF as well as widely used datasets such as Wikidata and DBpedia. Furthermore, organisations can be given a homepage, and places can be contextualised by using the E89 falls within property, which takes as a range another E53 Place.

4.5 Records as Events

As discussed in Chapter 1, several scholars have emphasised the benefits of having multiple access points to the archive. This entails not only being able to explore archival records by following the archival hierarchy, which, in the case of personal archives, originates from a single named creator, but also by leveraging other types of archival data. The events in which documents participate and related information, such as dates, people, and locations, are one example. Events in which documents participate are one example of alternate access points.

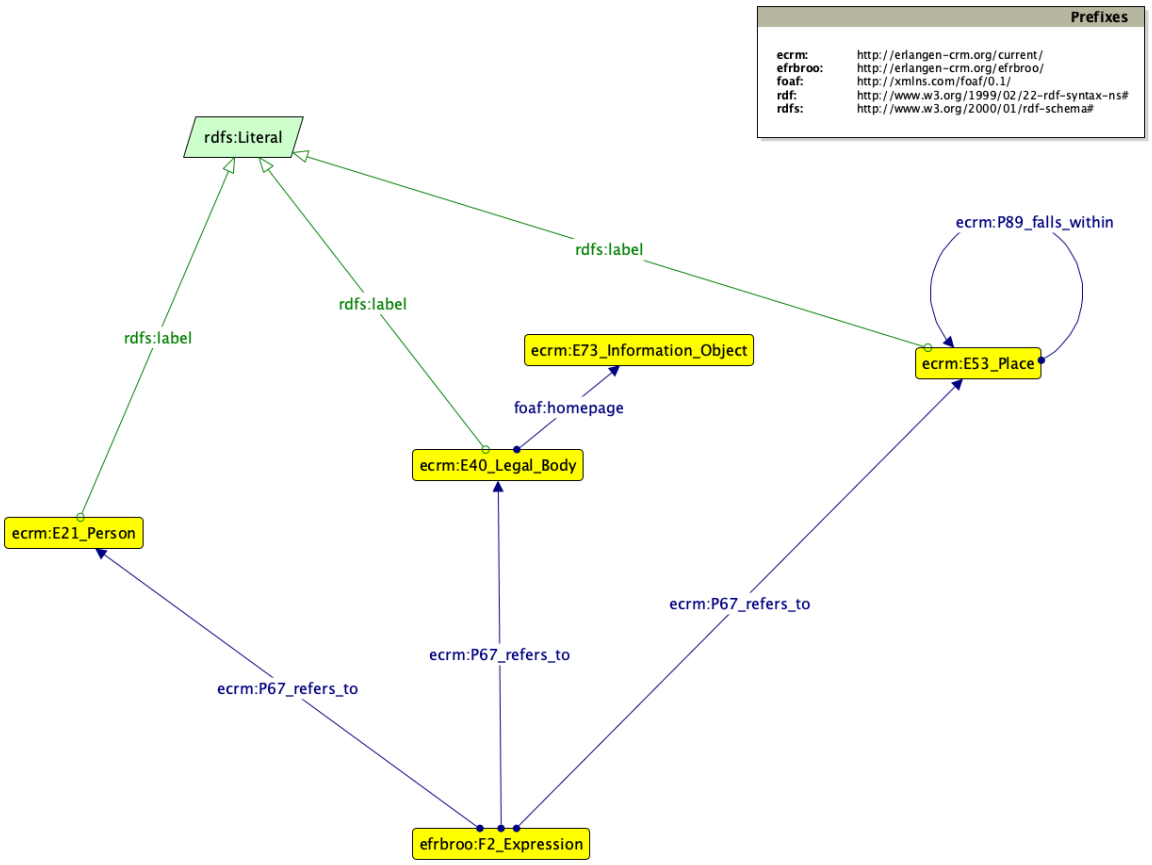

Figure 4.22: References to named entities and their description.

4.5.1 Manuscript Creation

Writing a manuscript results in the creation of a new expression as well as a new item, as explained in the scope note for the F28 Expression Creation class which clarifies that "the creation of an expression inevitably also affects the physical world" as it "causes a physical modification of the E18 Physical Thing that serves as the carrier".[60] As a result, the FRBRoo/LRMoo model provides two homonymous but distinct properties for defining the relationship between a single expression creation event and the resulting text and physical document (F28 Expression Creation *R17 created* F2 Expression, and F28 Expression Creation *R18 created* F4 Manifestation Singleton).[61]

In certain circumstances, a manuscript text is not new but a copy, although in the context of modern manuscripts this is a rare event, especially if even subtle changes, such as modified punctuation, are regarded as new features that determine the cre-

[60] Bekiari et al., *Definition of FRBRoo*, 27.

[61] In LRMoo, the range of the *R18 created* relation will be F5 Item rather than F4 Manifestation Singleton as the scope of F5 will be expanded to include unique documents, making the F4 class redundant.

ation of a new expression or version of the text. Nonetheless, The FICLITDL conceptual model allows for the modelling of an expression as an exact copy of another expression in cases where this is required: a copy is produced, rather than created, by a production activity (E12 Production *P108 has produced* F4 Manifestation Singleton).

In cases where a single document contains multiple separate texts, as with Giuseppe Raimondi's manuscript notebooks, it is possible to describe the corresponding creation event in greater detail by dividing it into further sub-events, one for each text, so that each text creation can be described separately (F28 Expression Creation *P9 consists of* F28 Expression Creation). Each F28 Expression Creation also creates a realisation of a work (F20 Expression Creation *R19 created a realisation of* F1 Work). This enables indirect linking of works to physical documents containing expressions that realise such works.

One or more authors, a date, and the technique used to generate the expression can all be specified for each expression creation (F28 Expression Creation *P14 carried out by* E39 Agent; *P4 has time-span* E52 Time Span; and *P32 used general technique* Technique). Handwriting and typing are examples of possible writing techniques available within FICLITDL-O.

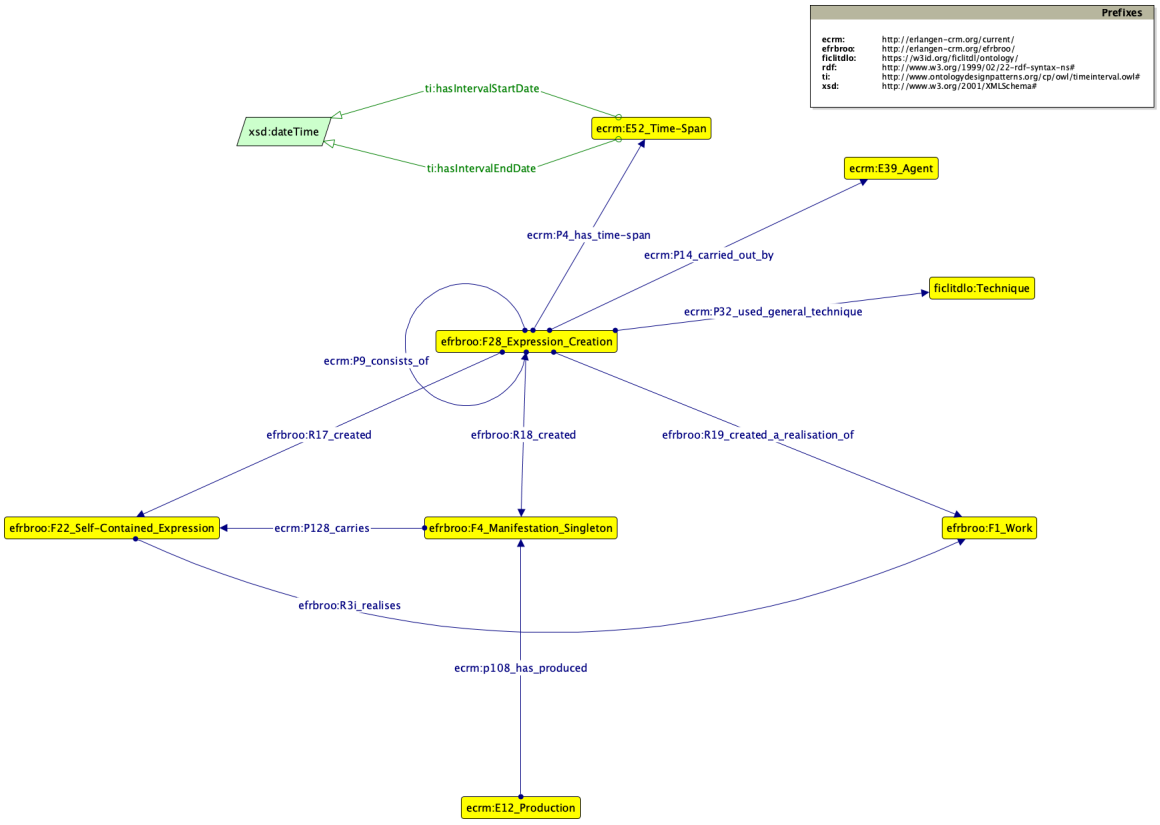

Figure 4.23: Manuscript creation.

4.5.1.1 Authorship attribution

The E13 Attribute Assignment class can be used to represent authorship attribution. The attribute assignment activity assigns the role of author (the attribute) to a person (E13 Attribute Assignment *P140 assigned attribute to* E21 Person; and *P141 assigned* Role In Time). The agent responsible for the attribution, either an individual or an individual acting on behalf of an institution, and the date and time of attribution are attached to the attribute assignment event.

The description of roles in the FICLITDL conceptual model reuses concepts from the Publishing Roles Ontology (PRO), as illustrated in Figure 4.24.[62] The PRO conceptual model allows for a representation of roles in which each role is bound to a specific context. For example, the role of author is not generically assigned to Giuseppe Raimondi, but he is described as holding the role of author with respect to a specific work. The Role In Time class interconnects the assigned role, the person who holds it, and the entity to which the role is related.

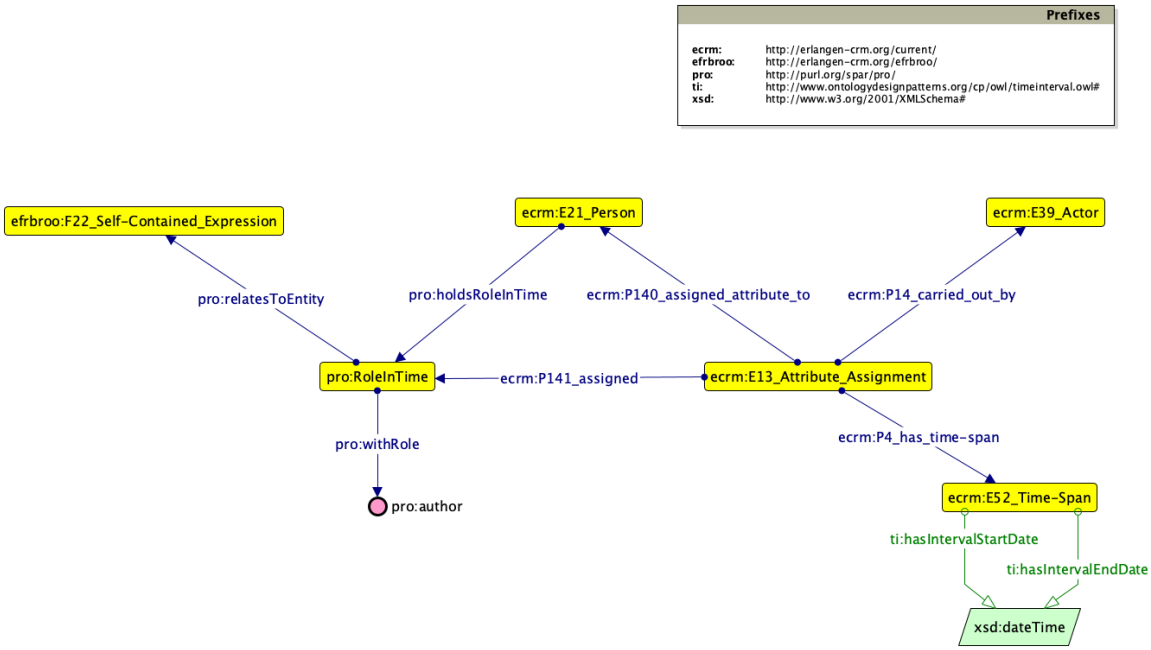

Figure 4.24: Authorship attribution.

4.5.2 Influence Between Texts

Influence between texts is expressed indirectly using the *P15 was influenced by* property, which "captures the relationship between an E7 Activity and anything that may

[62] Peroni, Shotton and Vitali, 'Scholarly publishing and linked data'.

have had some bearing upon it".[63] As such, instead of asserting that a text was influenced by another text, the activity of creation of a new text is said to be influenced by an F2 Expression (F28 Expression Creation *P15 was influenced by* F2 Expression).

As discussed before in this chapter, the FICLITDL conceptual model focuses on the role of writers not only as authors but also as readers. An E7 Activity of type 'reading' can be used to represent the reading activity connected to a text that had an impact on the creation of another (F28 Expression Creation *P15 was influenced by* E7 Activity *P2 has type* 'reading'; relates To Text F2 Expression). The PRO model is employed to associate a reading activity to a person in the role of reader (E21 Person *pro:holdsRoleInTime* pro:RoleInTime; pro:RoleInTime *pro:relatesTo* Entity E7 Activity; with Role 'reader'). A time frame can be specified for each reading activity, allowing a writer's activities as a reader and as an author to be linked (E7 Activity *P4 has time-span* E52 Time-Span). For example, it is possible to query the knowledge base to find out which texts an author read before creating a specific work.

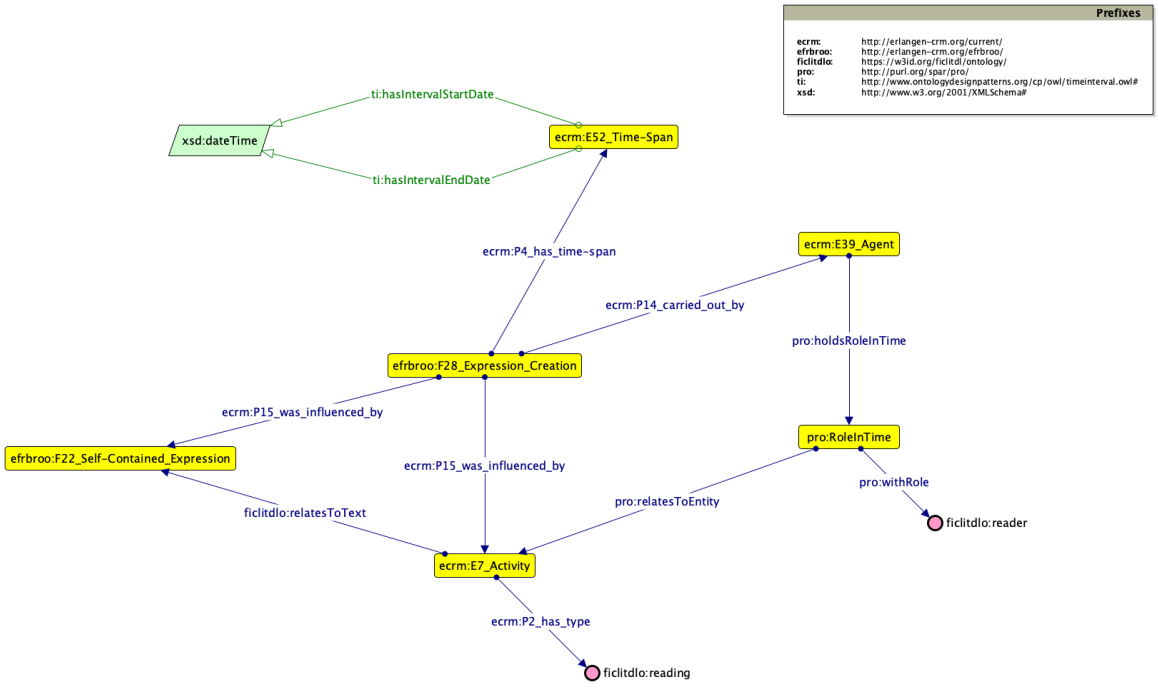

Figure 4.25: Influence between texts.

4.5.3 Textual Variation

One of the goals of the FICLITDL conceptual model is to allow for the representation of interrelationships between archival records, with a particular emphasis on

[63] Bekiari et al., *Definition of the CIDOC Conceptual Reference Model (Version 7.1.2)*, 124.

connecting texts. The personal archive and library of Giuseppe Raimondi contains different versions, either manuscript copies or print editions, of the same works. Raimondi's notebooks, for example, bear manuscript texts that frequently contain internal variation in the form of deletions, additions, substitutions and transpositions. Textual variation between versions of one work is frequently observed externally as well. The majority of the texts in Raimondi' notebook, to continue the example, were later published as part of collections of short stories, novels or articles, and, occasionally, Raimondi's library contains the author's personal copies of such publications. Additional copies, sometimes bearing annotations, can also be found in other writers' and academics' archives held at the FICLIT Department.

The existing finding aid does not provide connections between different versions of a work within Giuseppe Raimondi's archive and library, despite the fact that such a feature would greatly benefit the community of researchers exploring Raimondi's literary production. Modern schools of textual research, in fact, focus on the text as a process, adopting a dynamic rather than static perspective in which each text is examined in relation to previous versions.[64]

When compared with the existing finding aid, the new archival knowledge base, should eventually be able to answer questions about how a text is connected to other texts, including at least the following:

1. What are all the versions of a specific work in the archive?
2. Which variant versions of a specific text are available?
3. Which variant versions of a specific text are carried by a notebook?
4. If there are two distinct versions of the same work in the archive, what are all the variants?
5. For a given text, what are all the immediate revisions made by the author?

Figure 4.26 illustrates the representation of textual variation. As anticipated early in this chapter, FRBRoo/LRMoo does not provide expression-to-expression relations. As such, in order to interconnect variant versions of one work, the FICLITDL conceptual model introduces the generic symmetric relation has Variant Version which takes as a domain, as well as a range, an F2 Expression. In addition to that, this model reuses properties from the Provenance Ontology (PROV-O). Possible relationships between texts are 'text B derives from text A' (prov:wasDerivedFrom);

[64] See Giulia Raboni and Paola Italia, *Che cos' è la filologia d'autore*, Publisher: Carocci (Roma: Carocci, 2010), 9–33.

'text B is a revision of text A' (prov:wasRevisionOf); 'text B is an exact copy of text A' (*ficlitdlo:isExactCopyOf*). The latter property belongs to the namespace of the FICLIT Digital Library Ontology and is declared as a subproperty of *prov:wasDerivedFrom*. This set of relationships is based on research conducted on the Raimondi case study and will need to be expanded if new types of interconnections will be identified.

In addition to representing relations between texts, scholars can express intertextual connections and, in particular, textual variation at a finer level of granularity. Textual variation is seen as an event involving at least two alternative variants, either internal or external. The Textual Variation class is declared as a subclass of E5 Event in the FICLITDL-O namespace. Textual variants are represented as instances of E90 Symbolic Object that participate in the textual variation (ficlitdlo:TextualVariation *ficlitdlo:hasVariantReading* E90 Symbolic Object).[65] For each variant text, a reading is provided by means of the rdf:value data property which takes as a range a string. Each variant reading is linked to the witness to the text (an individual of F2 Expression or one of its subclasses, including F24 Publication Expression to represent printed editions of the work) by means of the *P106 is composed of (forms part of)* property, which relates a textual fragment to the text it belongs to. Each text, in turn, is linked to the physical document that bears it using the *P128 carries* property, as already described in Section 4.3.3. Researchers can use this construct to express the presence of variation between two or more texts, specifying which parts of the text differ and how.

In the context of Giuseppe Raimondi's archive, and of writers' archives in general, variation is the result of an activity of text revision performed by the author. Taking that into consideration, the FICLITDL conceptual model allows for the description of the act of text revision, which is seen as an activity performed by an agent at a specific point in time, in addition to asserting that there is a case of textual variation between two or more variant expressions. The ficlitdlo:TextRevision class, which is declared as a subclass of E7 Activity within the FICLITDL-O namespace, can have a time span (ficlitdlo:TextRevision *P4 has time-span* E52 Time-Span) and may be associated with the agent responsible for the revision (ficlitdlo:TextRevision *P14 carried out by* E21 Person). An act of text revision results in an event of textual variation (ficlitdlo:TextRevision *ficlitdlo:resultedIn* ficlitdlo:TextualVariation).

Text revision is classified by type (ficlitdlo:TextRevision *P2 has type* ficlitdlo:Re-

[65] The FICLITDL-O *has variant reading* object property is declared as a subproperty of *P12 occurred in the presence of*, which takes as a domain an E5 Event and as a range an E77 Persistent Item.

visionType). Types of text revision (which are members of ficlitdlo:RevisionType, a subclass of E55 Type) include deletions (the type for a text revision that is a deletion, i.e. it was marked or somehow indicated by the author as deleted), additions (the type for a text revision that is an addition, i.e. a syntactic and/or semantic expansion), substitution (the type for a text revision that is a substitution, i.e a deletion followed by an addition), omissions (the type for a text revision that is an omission, i.e. the author or scribe did not include the reading in the witness); transposition (the type for a text revision that was transposed from another location in the text, including the flipping of the order of two words or phrases).

If a text revision is a substitution, the text revision act deletes one textual fragment while adding another (ficlitdlo:TextRevision *ficlitdlo:deleted* E90 Symbolic Object; *ficlitdlo:added* E90 Symbolic Object). If the text revision is a simple addition or deletion only the *ficlitdlo:added* or *ficlitdlo:deleted* relation may be established. In case of a transposition (which could be seen as composed of two separate activities of deletion and addition), the *ficlitdlo:deleted* and *ficlitdlo:added* relations are used but the text revision activity is typed as a transposition to distinguish it from other categories of revision.

When textual variation involves more than two readings (either internal, i.e. being layers of the same witness, or external, i.e. being part of different witnesses to one work) it is possible to reconstruct the chronological order of variation. This can be achieved by asserting that a particular authorial act of text revision precedes or follows another one (ficlitdlo:TextRevision *seq:precedes* ficlitdlo:TextRevision; or *seq:follows* ficlitdlo:TextRevision).

As already said, readings are modelled as instances of E90 Symbolic Objects. When transcriptions of the texts are available in an XML/TEI format as part of the archival knowledge base, readings can be linked to the exact location in the text where they appear using the Open Annotation data model: the body of the annotation is the textual variation event, while the targets are the textual variants; each variant can be linked to its location in the transcription, as described already in Section 4.4.2 and as shown in Figure 4.21.

4.6 Archival Order in the Graph

The personal archive of Giuseppe Raimondi is divided into two subfonds (the library and the papers). The latter is divided further into four record sets, which in-

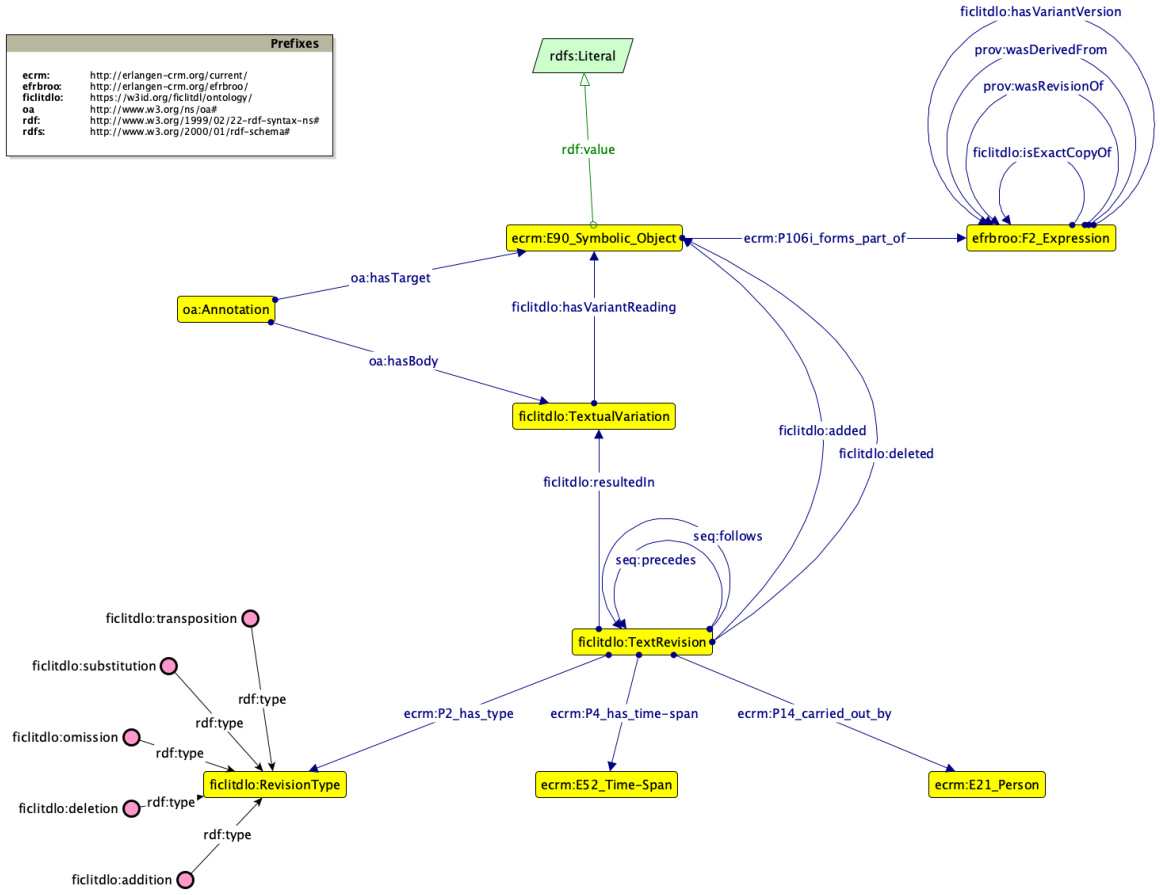

Figure 4.26: Textual variation resulting from authorial revision.

clude manuscript articles, notebooks, correspondence, and albums. Each of these record sets has additional subsets. Articles, for example, are organised by year of publication, whereas correspondence is organised by correspondent. The current organisation was purposefully established by Giuseppe Raimondi himself, though changes probably occurred in the at least two transfers of custody that the collection underwent — first when the author's heirs inherited it, and then when they handed the material over to the Ezio Raimondi Library of the FICLIT Department. Giuseppe Raimondi, as previously discussed in Section 4.2.5, also reorganised a selection of his correspondence in the 1970s, creating 14 thematic albums.

The representation of Giuseppe Raimondi's archive reflects the hierarchical organisation outlined above. However, as Bailey points out, the notion of archival order as a hierarchy is "a product of the nature of the archival record as a physical document to be sorted, grouped, and described after receipt from a single donating entity".[66] In the digital realm, by contrast, and particularly in a linked open

[66] Jefferson Bailey, 'Disrespect des Fonds: Rethinking Arrangement and Description in Born-Digital Archives', *Archive Journal*, 3rd ser., 2013, accessed 16 March 2021, http://dev.archivejournal.net/?p=4722.

data framework, order becomes multidimensional and takes the form of a network of heterogeneous relationships: the archival hierarchy, whose representation is the centre of this section, becomes just one out of many different contextualisations of a record within a graph.

Before illustrating the modelling of hierarchical order in FICLITDL-O, it is important to note that the physical order of a collection does not necessarily correspond to its conceptual order. Yeo develops this argument further, as already discussed in Chapter 1. According to Yeo, the concepts of collection and fonds, which are traditionally considered to be diametrically opposed in archival theory, are actually complementary. He contends that the notion of collection corresponds to the physical aggregation of documents in an archival repository, whereas the fonds is a conceptual entity, that is, a critical reconstruction of the interrelationships between documents. The boundary of a fonds may correspond to that of the collection or may extend beyond the collection to include external material (letters sent to others are an example of archival material that, despite not being physically present in the collection, could be considered part of an author's conceptual fonds). Based on these assumptions, Yeo suggests that multiple reconstructions or, more accurately, multiple fonds can coexist and overlap.[67]

The archival knowledge base should be able to answer at least the following questions about order:

1. Which physical aggregation is the document a part of?
2. Which record set(s) is the document part of?
3. Which record set(s) is the record set part of?
4. What is the next record in the sequence?
5. What is the previous record in the sequence?

The E78 Curated Holding class in CIDOC-CRM represents physical aggregations of documents that are curated and preserved by an institution.[68] This class, in conjunction with the set of relations shown in Figure 4.27, enables the description of physical archival collections. The *P46 is composed of (forms part of)* property allows for the representation of how documents are physically organised in the archive (for example, the manuscript articles in Giuseppe Raimondi's archive are stored in

[67] Yeo, 'The Conceptual Fonds and the Physical Collection'.
[68] Bekiari et al., *Definition of the CIDOC Conceptual Reference Model (Version 7.1.2)*, 101.

folders divided by year of publication), but it cannot be used to represent hierarchical relationships from a conceptual point of view. As such, the FICLITDL conceptual model specifies a new relationship between physical records that allows for asserting that a record (an E22 Human-Made Object) forms conceptually part of (*ficlitdlo:formsConceptuallyPartOf*) a particular record set (an E89 Propositional Object).

The introduction of this relationship allows for the organisation of archival material into conceptual record sets in addition to physical aggregations, allowing users to access and understand archival documents and data in alternative ways. For example, researchers could conceptually combine archival material into an additional set of records to document exhibitions involving such material. This type of record set may overlap with other physical or conceptual record sets so that the same document may belong to multiple conceptual sets at the same time. Documents chosen by Raimondi from the correspondents folders on the occasion of the "Giuseppe Raimondi fra poeti e pittori" exhibition, for example, are physically part of the albums but are conceptually part of both the albums and the correspondents files.[69]

For representing changes to a physical aggregation, such as moving a document from one folder to another, CIDOC-CRM provides the E79 Part Addition and E80 Part Removal classes. In this way, it is possible to document the evolution of a physical archive over time. The *seq:follows* and *seq:precedes* properties can be used to represent sequences of items within a physical aggregation or a conceptual record set. In order to specify the level of description to which the entity being represented belong, the FICLITDL conceptual model introduces the Archival Unit Type class as a specialisation of E55 Type. Members of Archival Unit Type include files, series and fonds. The concepts of subfile, subseries and subfonds are not represented as they can be derived from the way units are nested into one another.

4.7 Multiple Provenances

One of the goals of the FICLIT Digital Library is to enable researchers to supplement archival representations with additional contextualisations. Each contextualisation is represented by a set of assertions that form a graph.

Each graph of statements must be associated with provenance information in

[69] See Raimondi and Mazzotta, *Giuseppe Raimondi fra poeti e pittori*.

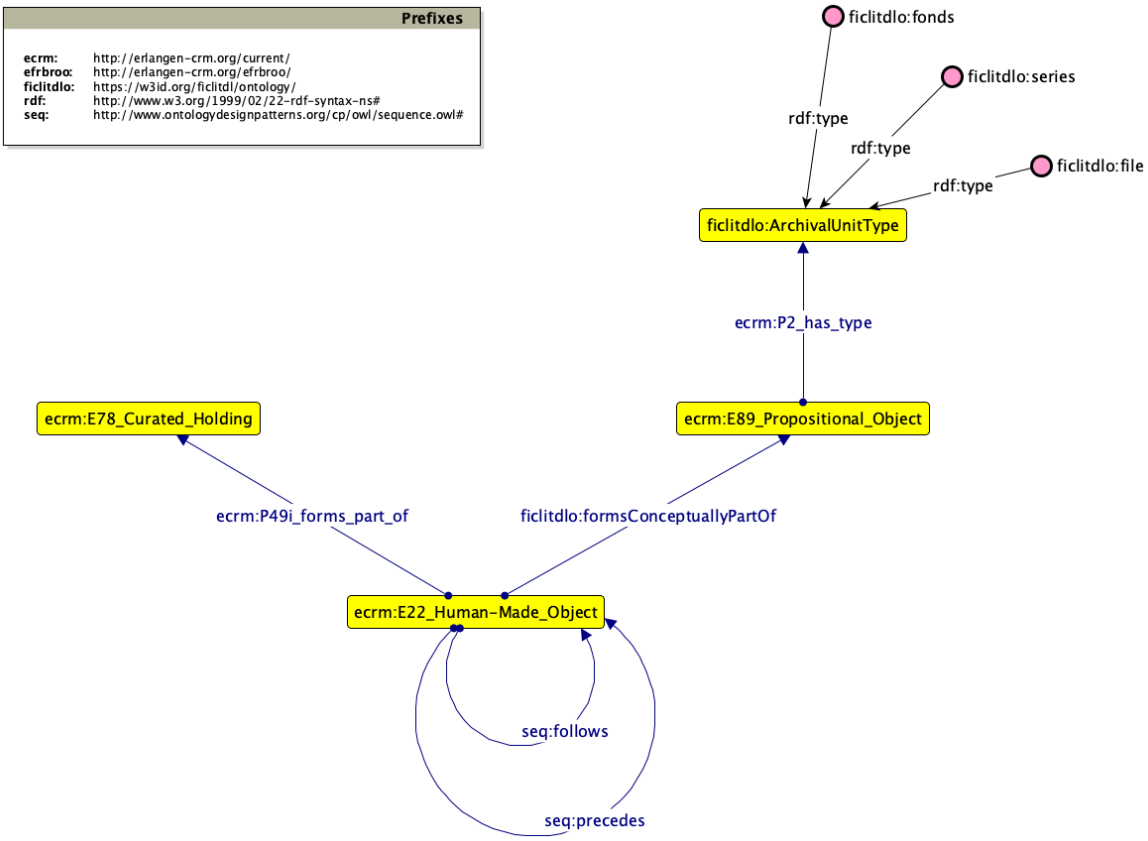

Figure 4.27: Archival order (hierarchy).

order to accommodate multiple interpretations in the finding aid. This enables the archival knowledge base to describe not only the archive but also the process of archival representation, and to incorporate more graphs of statements conveying scholarly interpretations over time. To provide information about the provenance of each set of assertions, the FICLITDL conceptual model and ontology use named graphs, specifically the Nanopublication model, which provide a means of describing a graph of assertions.[70]

The graph of scholarly assertions (for example, a researcher's contextualisation of a group of records) is assigned a unique resource identifier, as shown in Figure 4.28. Every nanopublication is made up of four graphs. The first graph contains the researcher's assertions; the second graph describes the provenance of the first graph (the date of creation and the person responsible for the creation); the third graph describes the provenance of the nanopublication itself (the person entering the nanopublication in the archival knowledge graph may differ from the researcher proposing the interpretation); and the fourth graph combines the first three graphs

[70] For an in-depth introduction to the Nanopublication framework see Groth, Gibson and Velterop, 'The anatomy of a nanopublication'.

into a single graph.

A nanopublication, as shown in Figure 4.28, is composed of three named graphs, represented as rounded squares. Each graph is assigned a unique resource identifier. The first graph contains the assertions (for example, a researcher's contextualisation of a group of records); a second graph provides information about the provenance of the first graph (the date of creation and the person responsible for the creation); and a third graph describes the provenance of the nanopublication itself (the person entering the nanopublication in the archival knowledge graph may differ from the researcher proposing the interpretation). The three graphs are connected together to form a nanopublication that represents the scholarly interpretation as a citable entity.

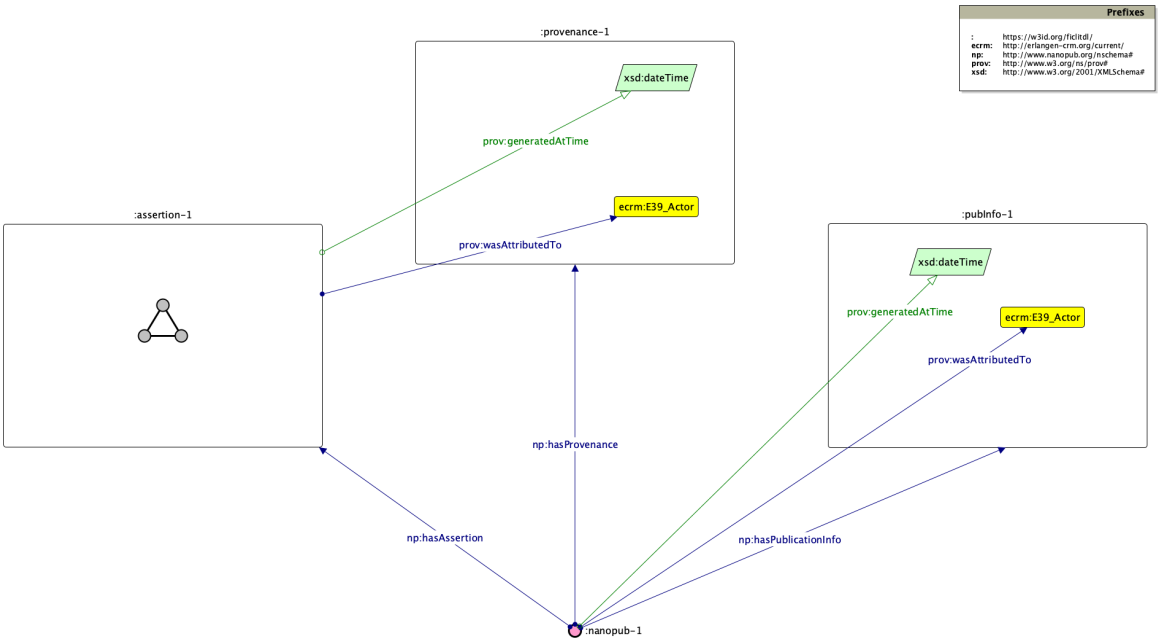

Figure 4.28: Multiple provenances.

Using a practical example from the Giuseppe Raimondi Archive, the next chapter will evaluate the FICLIT Digital Library conceptual model by applying it to an heterogeneous selection of case studies taken from the Giuseppe Raimondi's personal archive.

Chapter 5

Evaluating the FICLITDL Conceptual Model: Four Use Cases

Using the personal archive and library of Giuseppe Raimondi as a case study, this chapter aims to demonstrate how the FICLIT Digital Library conceptual model introduced in Chapter 4 applies to the Department's collection of writers' archives. In particular, the chapter presents four use cases involving different types of archival documents and varying levels of representation complexity from the Giuseppe Raimondi Fonds and other archival collections. Each use case provides an example of enrichment of the archival knowledge base by subsequent researchers contributing to the representation with supplemental interpretations. Each individual addition is associated with provenance information as is the base archival description, allowing the contributions expanding the archive over time to be distinguished and attributed to the respective authors. The agent who added the assertions to the archival knowledge base is also associated with provenance information. Indeed, by converting the researcher's reconstruction into RDF triples, the agent adds an extra layer of interpretation to the representation that should be made explicit. The use cases will serve as a test for evaluating and assessing the expressivity of archival representations based on the FICLITDL conceptual model.

Each use case is divided into three main parts: 1. the original ISBD archival note, accompanied by a natural language description and analysis of the use case; 2. the representation of the use case according to the FICLITDL conceptual model; 3. a sample set of competency questions based on the questions defined for each document type in Chapter 4 and formalised in SPARQL, the standard query language for RDF, with results. The descriptions of the representations of the use cases contained

in this chapter are aided by illustrations showing the reader the entities and relations involved. The source code of the representations (in TriG, an extension of the Turtle language that defines an RDF syntax for multiple graphs, as well as N-Quads, a line base format for encoding RDF quadruples) and the SPARQL formalisation of the queries are available online in the dedicated Github repository that is an integral part of this chapter.[1]

5.1 Use Case 1: Heterogeneous Documents Bearing Variants of a Work

The first use case focuses on the representation of the interconnections between heterogeneous documents — including archival records and published material — that contain different versions of a work.[2] The entry below, taken from the existing ISBD finding aid for the Giuseppe Raimondi Archive, describes one of Raimondi's manuscript notebooks:

> RDq 303
> FR.A QUADERNI.1 1976 03
> 07/03/93
> Testo manoscritto
> 3627063
> Una forca per il poeta Francois Villon / [Giuseppe Raimondi]. - 1976. - 1 quaderno (7 p. ms., di cui alcune numerate irregolarmente, su 10 c.) ; 21 cm. ((In cop., di altra mano, anche: (Gelo invernale e nostalgia di legna accesa) Il giorno 7.6.76; I dattiloscritti sono inseriti ne "I tetti sulla citta"; a c. [3v]: 24.5.1976; a c. [5v]: 3 luglio 1976. - Contiene anche: A proposito di tegole, di tetti e di fantasmi. – Ms. FA PARTE DI Mat. document.-manoscr 3621049 [Quaderni manoscritti]. 1954-[1976] / Giuseppe Raimondi. - 1954-1976. - 308 quaderni ; 22 cm. ((Prima serie di quaderni di appunti, minute di racconti, recensioni, articoli. Serie ordinata dallo stesso Giuseppe Raimondi in carpette per anno.[3]

The description begins with the manuscript inventory number (RDq303), its col-

[1] The Github repository containing all code produced during the course of this project can be accessed from https://github.com/fgiovannetti/giuseppe-raimondi-lod.

[2] This section expands on a paper presented at the Linked Archives International Workshop, Online, September 13, 2021 (Francesca Giovannetti and Francesca Tomasi, 'The linked finding aid as a platform for textual research: The case study of the Giuseppe Raimondi Archive', in *Proceedings of the Linked Archives International Workshop* (2021), 104–113).

[3] Giuseppe Raimondi, 'Una forca per il poeta François Villon' (Quaderno ms. BFICLIT FR.A QUADERNI.1 1976 03, RDq 303, 1976).

location in the archival hierarchy (FR.A QUADERNI.1 1976 03, which indicates that notebook is part of the Giuseppe Raimondi Archive, Notebooks Series, Notebooks Subseries 1, Folder 1973 and that, within this folder, the notebook is third in sequence), the date of creation of the finding aid (March 7, 1993), the record type (a manuscript text), and a unique identifier (3627063). The archival note that accompanies the description provides a title for the notebook, "Una forca per il poeta François Villon". The text is attributed to Giuseppe Raimondi (authorship attribution is given between square brackets when the author's name does not appear on the manuscript, and the same convention is used for dates or other types of metadata that are not explicitly declared on the manuscripts).

The note continues with the date of the creation of the text. It then specifies the document type (a notebook), the length of the text in page numbers, and the total length of the manuscript notebook, followed by a measurement (in cm) corresponding to the height of the document, although this is note explicitly stated.

The last section of the entry assigns the notebook to the Manuscript Notebook series titled "[Quaderni manoscritti]. 1954-[1976]", which contains 308 notebooks that are further arranged by year. In particular, the notebook is part of the file containing all the notebooks dated 1976. The note concludes by stating that this series was organised by Giuseppe Raimondi himself.

In addition to "Una forca per il poeta François Villon" (hereafter identified as T1), the notebook description mentions three other titles: "Gelo invernale e nostalgia di legna accesa" (T2), "I tetti sulla città" (T3) and "A proposito di tegole, di tetti e di fantasmi" (T4).[4] However, the note does not indicate how the texts relate to one another, so users will need to visit the Giuseppe Raimondi Archive in person to clearly interpret the information provided. The note also contains three full dates that are not explicitly linked to the corresponding texts.

Closer examination has revealed that the notebook contains only two of the texts listed in the archival note: T1 (dated 24 May 1976) and T4 (3 July 1976). Both texts were later included in T3, a collection of short stories by Giuseppe Raimondi first published in September 1977.[5] The archive contains the author's personal copy of T3 (with the ownership note "Giuseppe Raimondi. 1927"). In addition, the Ezio Raimondi Library owns two other copies of the same volume in the personal archives of Clemente Mazzotta and Ezio Raimondi. One of these copies bears a dedication

[4] The cover of the notebook appears to have been compiled at different times by different hands, including Raimondi's, as evidenced by the use of at least two different pens and handwritings.
[5] Giuseppe Raimondi, *I tetti sulla città* (Milano: Mondadori, 1977).

by Giuseppe Raimondi to Clemente Mazzotta on the title page, while the other has pencil underlining by Ezio Raimondi.[6] As for T2, it is a variant version of T1 that appeared on *Il Giorno* one month after T1 was added to the notebook (7 June 1976).[7] The current archival representation, being hierarchically structured, does not provide a link between the individual documents that contain versions of the same work.

The Giuseppe Raimondi Archive features three further texts about François Villon: 1. "Le nevi di Villon", dated 12 December 1972, which is located within the manuscript articles series and was published in *Il Resto del Carlino* on 2 January 1973, and *La Nazione* on 4 January 1973;[8] 2. "L'uomo dal bicchiere di vino", dated 24 August 1967; and 3. "Il 'mostro' François Villon", dated 25 May 1966, which are also part of the notebook series. In particular, the second text, "L'uomo dal bicchiere di vino", associates the Louvre portrait *Man with a Glass of Wine* with the French poet, constituting one of many examples of connection between art and literature in Raimondi's production.[9]

This emerging path of interrelationships between heterogeneous documents and data across different record sets and collections illustrates the nature of scholarly research within archives, which revolves around clusters rather than hierarchies of records.[10] How can this scenario, reconstructed through subsequent research within the archive, be represented in the finding aid to facilitate and enrich future searches?

5.1.1 Use Case Representation According to the FICLITDL Conceptual Model

Figures 5.1-5.4 show the representation of the use case according to the FICLITDL conceptual model. Different colours are used throughout the illustrations to distinguish between different contributors to the archival knowledge graph: the base archival description is black, while sets of relations added later are displayed in

[6] BFICLIT, FR.A TETTIS, RL7137; BFICLIT, RAIMONDI II 03 G 0037, RAI 5105; BFICLIT, MAZZOTTA C 0571, 465 41448.

[7] Giuseppe Raimondi, 'Nelle ballate di François Villon. Gelo invernale e nostalgia di legna accesa', *Il Giorno*, 7 June 1978.

[8] For the manuscript of the article see Giuseppe Raimondi, 'Le nevi di Villon' (Articolo ms. BFICLIT FR.A ARTICOLI1 1972 12.22, RDa 349, 22 December 1972). For the published versions of the article see Giuseppe Raimondi, 'Le nevi di Villon', *Il Resto del Carlino*, 2 January 1973 and Giuseppe Raimondi, 'Le nevi di Villon', *La Nazione*, 4 January 1973.

[9] The painting titled *Man with a Glass of Wine*, an oil on panel on display at the Louvre Museum, is dated 1450-1475 and attributed to an anonymous artist. The persistent identifier of the painting is https://collections.louvre.fr/ark:/53355/cl010066409.

[10] See Yakel, 2003, 4-5.

different colours. Punctuated and dashed lines are used to highlight further layers of assertions by researchers. The prefixes shown in Listing 5.1 below apply to all figures in the chapter.[11]

```
@base <https://w3id.org/giuseppe-raimondi-lod/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ecrm: <http://erlangen-crm.org/current/> .
@prefix efrbroo: <http://erlangen-crm.org/efrbroo/> .
@prefix ficlitdlo: <https://w3id.org/ficlitdl/ontology/> .
@prefix ficlitdl-np: <https://w3id.org/ficlitdl/nanopub/nanopub2/> .
@prefix grlod-np: <https://w3id.org/giuseppe-raimondi-lod/nanopub/
    nanopub2/> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix prism: <http://prismstandard.org/namespaces/basic/2.0/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix pro: <https://purl.org/spar/pro/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix seq: <http://www.ontologydesignpatterns.org/cp/owl/sequence.owl#>
    .
@prefix ti: <http://www.ontologydesignpatterns.org/cp/owl/timeinterval.
    owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

Listing 5.1: List of prefixes valid for all figures in this chapter (Turtle syntax).

Figure 5.1 illustrates the representation of the notebook as a physical document.[12] The notebook is modelled as an instance of E22 Human-Made Object that carries an F22 Self-Contained Expression incorporating two texts, T1 and T4 (centre). Since it contains multiple distinct texts, the notebook is classified both as a single document of the type 'notebook' and as an archival 'file' by the *P2 has type* property (top left). The notebook is part of the archival unit "QUADERNI.1 1976", which is represented both as a physical collection of documents (an instance of E78 Curated Holding) and as a conceptual record set (an instance of E89 Propositional Object).

The creation of the notebook is modelled as an instance of F28 Expression Creation, linked to a specific date and technique (handwriting). The role of author is

[11] The scripts created for the transformation from ISBD to RDF of the archival descriptions of Giuseppe Raimondi's notebooks, written in Python, are available in the Github repository of the project at https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/scripts.

[12] All of the illustrations in this chapter are also available in high resolution in the same repository at https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/chapter-figures.

attributed to Giuseppe Raimondi using the Publishing Role Ontology (PRO) model, which allows for the reification of roles in such a way that each role is always linked to a specific context. In our case, Raimondi holds the role of author in the context of the creation of the notebook. The role of author, the person holding such a role and the created document are linked together using the Role In Time class (Figure 5.1, bottom right). The E13 Attribute Assignment class is used to model authorship attribution. The attribute assigned is the Role In Time, and the assignee is Giuseppe Raimondi. The agent responsible for the attribution (in this case the Soprintendenza archivistica e bibliografica dell'Emilia Romagna, an organisation rather than an individual) and the time of attribution are linked to the instance of E13 (Figure 5.2, bottom right).

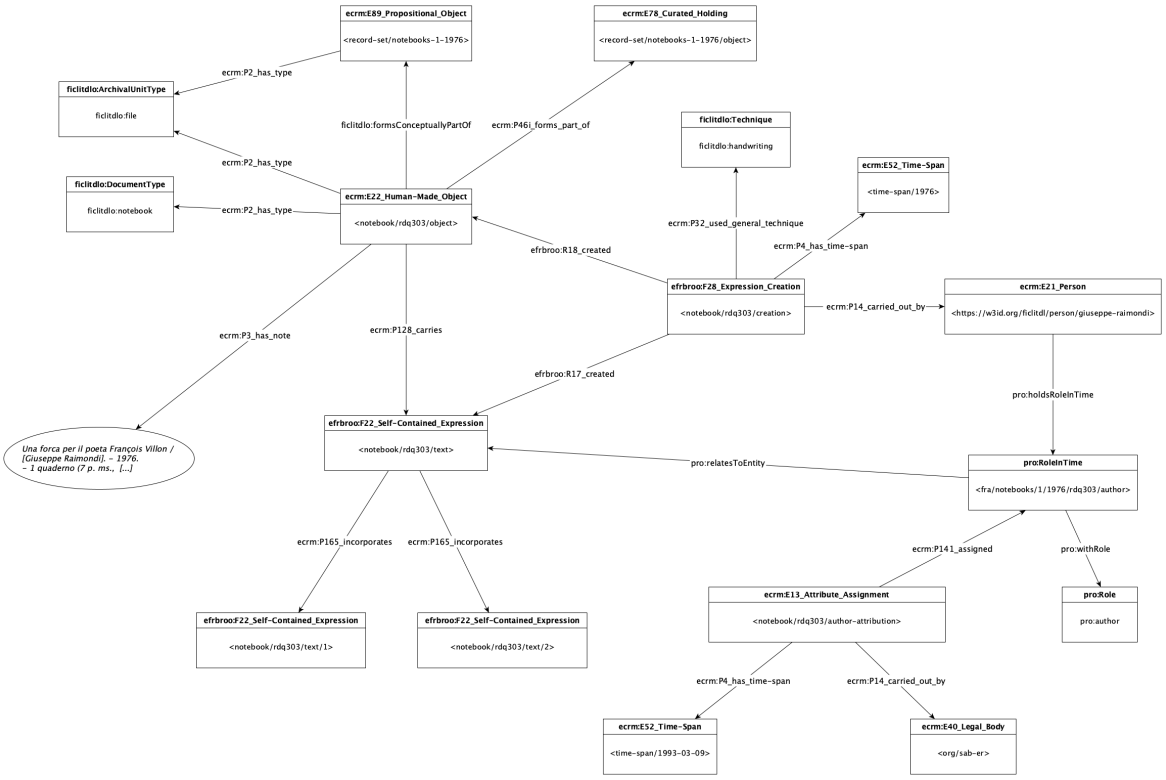

Figure 5.1: RDF representation of UC1, extracted from the existing ISBD description (part 1).

Continuing Figure 5.1, Figure 5.2 shows how the Expression Creation activity is divided into two sub-activities, one for each text, each linked to a specific date, via the *P9 consists of* property (Figure 5.2, top left). The association of the creation activities with dates is the output of subsequent research belonging to a separate additional graph (dashed).

Among the objectives of the FICLITDL conceptual model is to support the incorporation of subsequent scholarly interpretations within the finding aid. Figure

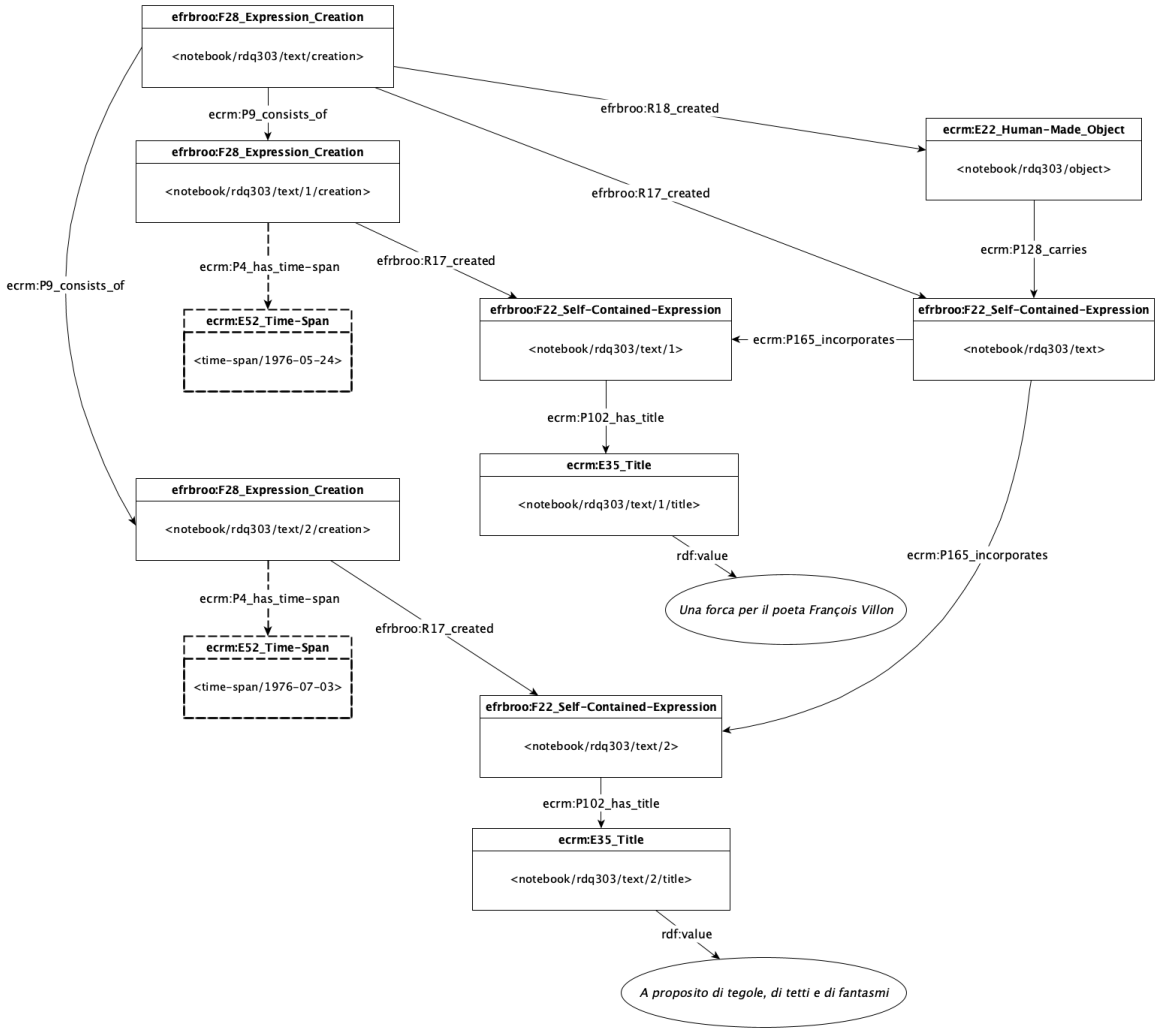

Figure 5.2: RDF representation of UC1, extracted from the existing ISBD description (part 2). Additional assertions by Researcher A are distinguished by dashed lines.

5.3 shows a graph of statements that reconstructs the relationships between T1, T2 and T3 as established by subsequent researchers analysing Giuseppe Raimondi's literary production. All three texts are modelled as instances of F22 Self-Contained Expression realising the same work. The texts are linked to their carriers, if these are contained in the archive: T1 is carried by the notebook, while T3 is carried by multiple printed volumes available in the personal archives of Giuseppe Raimondi, Ezio Raimondi and Clemente Mazzotta. Because the volumes are from different archival collections, connecting them within the finding aid represents a fundamental step towards dismantling archival data silos.[13]

The relationships *ficlitdlo:isPublishedVersionOf*, a specialisation of *prov:wasDerivedFrom*, provide an interpretation of the relationship between the alternate versions of the text. The physical documents containing such versions are also linked together

[13] See Nichols, 2009.

via a generic *dcterms:relation* to facilitate search and retrieval.

The same generic property, *dcterms:relation*, is used as a shortcut to link each notebook to the corresponding OPAC permalink of the published version so that users can navigate directly from manuscript to edition. The OPAC permalink is linked to the edition it represents (an individual of F3 Manifestation Product Type) via the property *ficlitdlo:hasSbnUboPermalink*, which was created ad hoc to accommodate the reconstruction of links between notebooks and editions.[14]

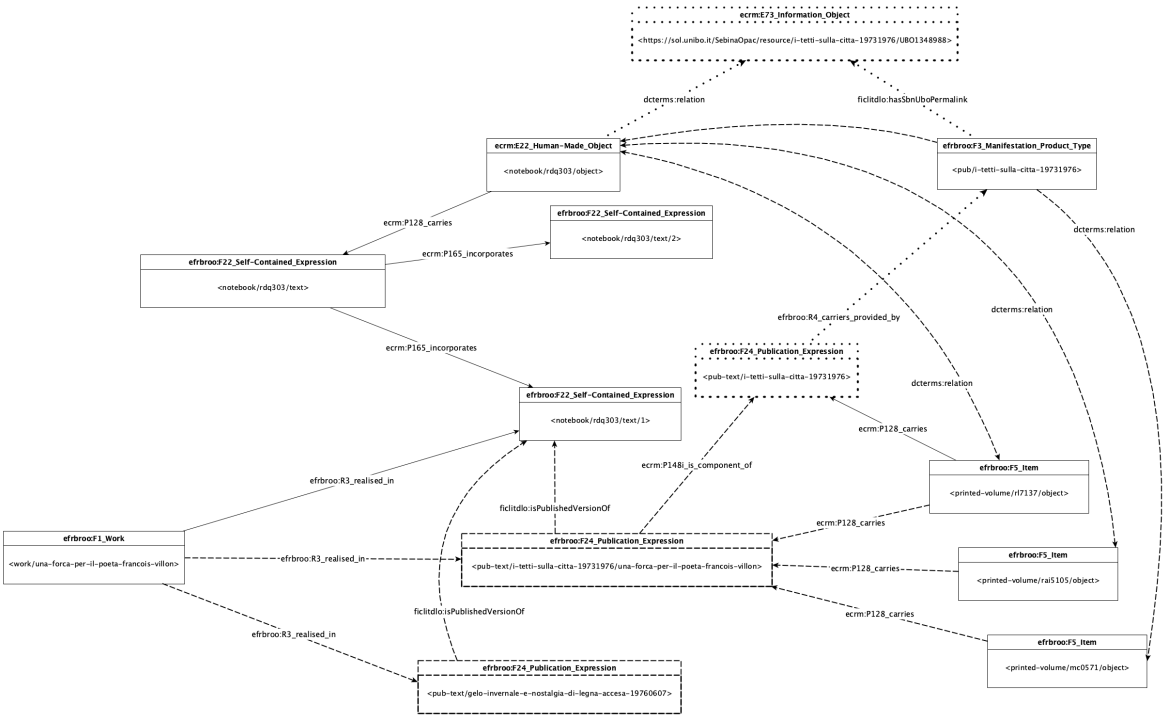

Figure 5.3: RDF representation of UC1, extracted from the existing ISBD description (part 3). Additional assertions by Researcher A are displayed using dashed lines, while additional assertions by Researcher B are displayed using punctuated lines.

Figure 5.4 completes the illustration of the use case. The notebook text is linked to other archival documents mentioning Villon via the *P67 refers to* property.[15] The same property is used to connect the notebook text referring to the artwork Men with a glass of wine to the permalink of the painting provided by the Louvre Museum.

In order to accommodate multiple perspectives in the finding aid, all collections of statements (in the case discussed above, there are four distinct graphs) must be

[14] The reconstruction of part of the links between the notebook texts and the corresponding published versions was carried out by Maria Chiara Tortora (FICLIT Department, University of Bologna) in the framework of her PhD project on Leopardi, Pascal and Galilei in Giuseppe Raimondi's production.

[15] References to persons have been extracted semi-automatically from the archival descriptions of the notebooks through Named Entity Recognition (NER) via OpenRefine, as described in Chapter 3.

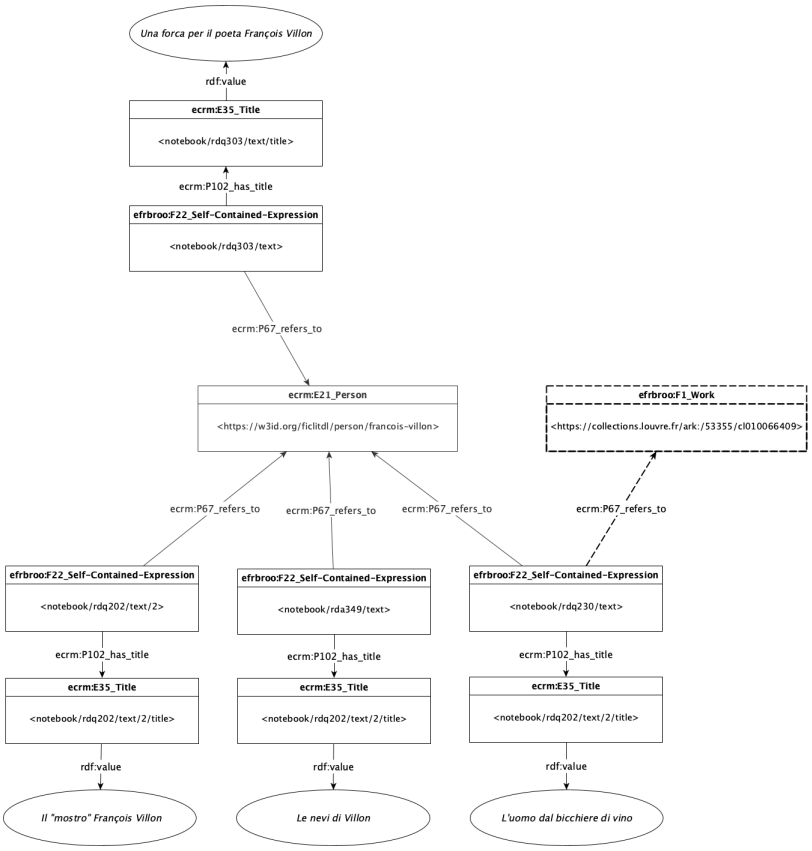

Figure 5.4: RDF representation of UC1, extracted from the existing ISBD description (part 4). Additional assertions by Researcher A are displayed using dashed lines.

associated with provenance information. This allows the archival knowledge base to describe not only the archive but also the process of archival representation and to integrate more collections of statements over time.

Provenance information is modelled using nanopublications. Listing 5.2 below shows the basic structure of the nanopublication encapsulating Researcher A's interpretation (dashed) from Figure 5.3.[16] It is composed of four graphs: 1. the graph of assertions being made by the researcher; 2. a graph describing the provenance of the assertions, 3. a graph describing the provenance of the publication itself; 4. the top graph combining the previous three graphs into a single nanopublication.

```
# Graph 1: Researcher's assertions.
grlod-np:assertion {
    # The researcher's reconstruction of the relationships between the
    texts (Figure 5.3, dashed).
}
# Graph 2: Provenance of the assertions.
grlod-np:provenance {
```

[16] The full representation of provenance information for each distinct graph can be found in the project's Github repository.

```
    grlod-np:assertion prov:generatedAtTime "2023-01-30"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> .
}
# Graph 3: Provenance of the nanopublication.
grlod-np:pubinfo {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub2>
        prov:generatedAtTime "2023-01-30"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> .
}
# Graph 4: the nanopublication and its components.
grlod-np:head {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub2>
        a np:Nanopublication ;
        np:hasAssertion grlod-np:assertion ;
        np:hasProvenance grlod-np:provenance ;
        np:hasPublicationInfo grlod-np:pubinfo .
}
```

Listing 5.2: RDF representation of provenance information for one of the supplemental graphs of statements according to the nanopublication data model (Turtle syntax).

5.1.2 Queries and Results

When accessing the enriched archival representation, users should be made aware of the wealth of contextual relationships that link the notebook to other documents, collections, and data, including people, places, events, and dates. For example, they should be able to ask for — as well as find serendipitously — the following information:

CQ1 All notebooks, with titles and page count.

CQ2 All notebooks that contain multiple texts.

CQ3 Texts contained in a specific notebook (RDq 303) with their titles and dates of creation.

CQ4 Authors of the texts and agents responsible for authorship attribution.

CQ5 Notebook texts that mentions one or more persons.

CQ6 All variants of the text “Una forca per il poeta François Villon” available in the archive.

CQ7 All archival documents that mention François Villon.

CQ8 All published versions of the notebook text "Una forca per il poeta François Villon".

CQ9 Relationships involving the notebook text "Una forca per il poeta François Villon", as reconstructed by a named researcher.

CQ10 Texts that mention one or more works of art.

The formalisation of these competency questions in SPARQL, the standard query language for RDF databases, is available from the project's Github repository. Each SPARQL query is run against the enriched archival knowledge base using Python. The generated results are organised into distinct CSV tables.[17]

5.2 Use Case 2: Author Revision of Short Story Title

This second use case focuses on a series of three notebooks, each of which contains multiple short story drafts featuring author revisions throughout the writing process and up to publication. The titles of the short stories went through different changes before reaching their final form in the short story collection *Mignon: racconti*, published in 1955.[18] For example, the title of the second short story draft contained in the second of three notebooks was revised by the author at least two times before eventually being published for the first time in 1954 as "Dei colombi in una facciata" in the monthly magazine *Comunità* (three copies of which are available in Giuseppe Raimondi's personal library)[19] and a second time as part of *Mignon*: the initial title "L'anno '43" was deleted by strike-through, while "Dei colombi in una facciata" was added at the top of the manuscript page (Figure 5.5, right). The notebook cover features one additional title for the same text, "I piccioni di Santa Lucia", which likely represent an intermediate stage between the initial "L'anno '43" and the final "Dei colombi in una facciata" (Figure 5.5, left). Below is the archival description for the notebook:

> Argomenti per i racconti : 2. volume. 1954 / [Giuseppe Raimondi]. - 1954. - 1 quaderno (25 p. ms. numerate irregolarmente su 22 c.) ; 22 cm. ((In cop. anche:

[17] The queries are available from https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/sparql, while results can be found at https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/results.
[18] Raimondi, 1955.
[19] BFICLIT, FR PER COMUNI 1954, RL 6683.

> 1. I quattro santi coronati; 2. L'anno '43 / I piccioni di Santa Lucia; 2. quaderno. - A c. [7]: (24.5.1954); a c. [14v]: (17.5.54). - Sul verso della cop. e a c. [10] incollati foglietti ms. - Ms. Tit. a c. [9]: Dei colombi in una facciata.[20]

The archival description mentions all three titles but does not go into detail about the revision process or the connections between them.[21] By reading this description, users are not informed about the publication of the short story in "Comunità" as well as "Mignon" as "Dei colombi in una facciata", but are only informed that this reworked version of the title appears on page 10 *recto*. Furthermore, there is no direct link between the notebook and Raimondi's personal copies of "Comunità"[22] and "Mignon".[23] How can the archival knowledge base be updated with the reconstruction of such a network of relationships, including those between the variant titles of the short story?

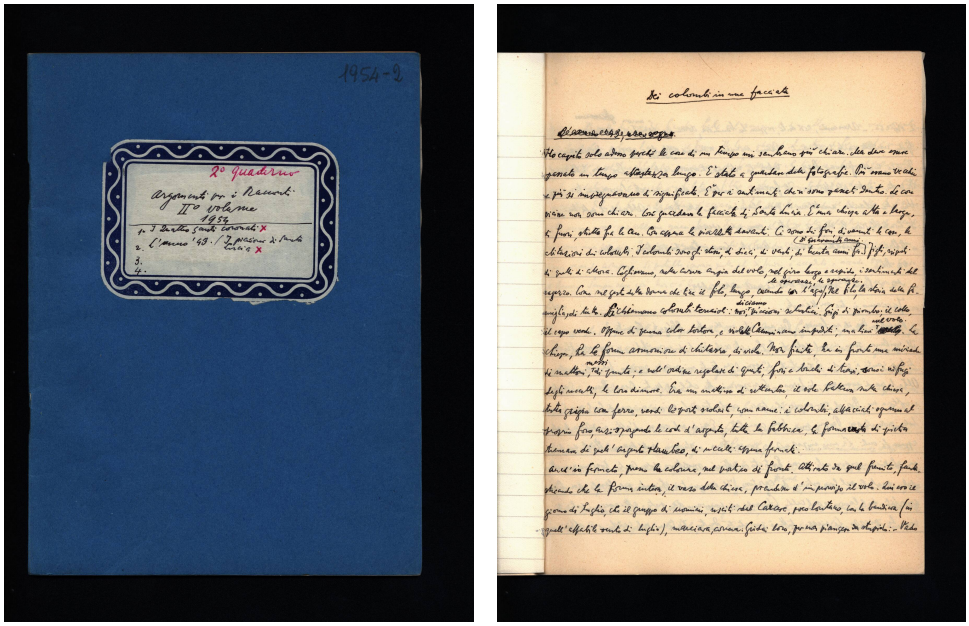

Figure 5.5: Notebook cover (left, 1 *recto*) and notebook page (right, 10 *recto*) featuring the different stages of title revision for one of the short stories that was eventually published as "Dei colombi in una facciata" in *Mignon: racconti* (1955).

[20] BFICLIT, FR.A QUADERNI.1 1954 02, RDq 2.

[21] A finding aid is not expected to contain elements of textual analysis and representation. However, one purpose of this study is to demonstrate how the transition of the finding aid from document-centric description to data-centric representation allows for the scope of the finding aid to be expanded beyond the domain of archival practice and into that of textual scholarship.

[22] See fn. 20 above.

[23] BFICLIT, FR.A MIGNON, RL7139.

5.2.1 Use Case Representation According to the FICLITDL Conceptual Model

Figures 5.6–5.9 show the representation of the use case according to the FICLITDL conceptual model. Solid lines represent entities and relationships that form part of the base graph, which contains all of the statements that were originally part of the ISBD finding aid and were extracted as part of this research using a series of Python scripts. Dashed lines, on the other hand, represent all of the statements conveying a subsequent enrichment of the archival data, as a result of a scholarly act of interpretation performed on the archive and later added to the base graph.

The short story is incorporated into the overall notebook text as an instance of F22 Self-Contained Expression. Each of the three short story titles is represented by an instance of E35 Title, a subclass of E90 Symbolic Object.[24] The additional reconstruction connects the short story to its published versions, which are then linked to their titles, allowing users to query the dataset for the published title of the short story draft contained in the notebook (Figure 5.6).[25]

Using the Textual Variation class, the researcher's reconstruction indicates the presence of an internal variation event within the notebook text. The *has variant reading* property links the instance of textual variation to the three different variant readings of the short story title. Using classes and properties from the Web Annotation data model, each reading, namely a title, is linked to the TEI-encoded transcription of the notebook text and its precise location within that transcription. For each reading, an XPath selector pointing to the @xml:id attribute that uniquely identifies each written line in the manuscript is provided (Figure 5.7).

The researcher elaborates further on the textual variation event, describing it as a result of the a text revision activity carried out by the author. Such an activity occurred in three stages, each represented as a distinct text revision event: in the first stage, a deletion, the author deleted "L'anno '43" from the notebook text; in the allegedly second stage, an addition, the author added "I piccioni di Santa Lucia"

[24] In version 0.7 of LRMoo, E90 Symbolic Object replaces the concept of F23 Expression Fragment in the spirit of avoiding the duplication of concepts that already exist in CIDOC-CRM.

[25] It is worth noting that the open-world assumption (OWA) is used in knowledge representation based on semantic web technologies, as opposed to a closed-world assumption (CWA). This means that what is just not in the dataset is simply unknown rather than false. In other words, the presence of only two published versions of the short story in the reconstruction does not imply that there are no more. For a detailed explanation of the concept of open-world assumption see C. Maria Keet, 'Open World Assumption', in *Encyclopedia of Systems Biology*, ed. Werner Dubitzky et al. (New York, NY: Springer New York, 2013), 1567–1567, https://doi.org/10.1007/978-1-4419-9863-7_734.

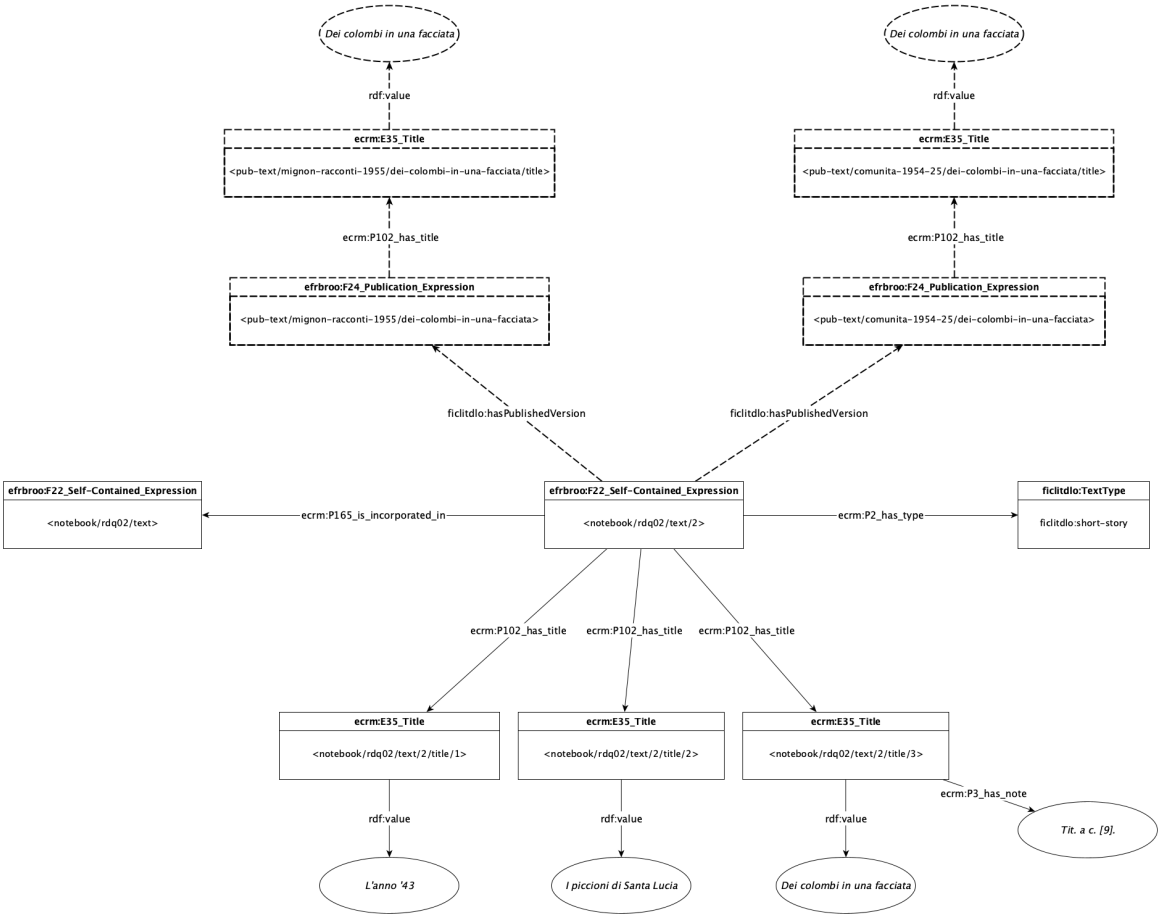

Figure 5.6: RDF representation of UC2, Part 1. Additional assertions by the researcher are displayed using dashed lines.

on the notebook cover; in the third stage, the author added "Dei colombi in una facciata" to the notebook text, above the crossed-out first version of the title, "L'anno '43". The *directly follows* and *directly precedes* properties are used to connect and order the three distinct revision phases (Figure 5.8).

Finally, the published versions of the short story draft manuscript are linked to the author's personal copies of such publications, which are housed in his library. The volume of *Mignon*, in particular, bears the author's underlines, which may provide users with additional information on his writing process (Figure 5.9).

```
# Graph 1: Researcher's assertions.
grlod-np:assertion {
    # The researcher's reconstruction of the relationships between the
   texts (Figures 5.6-5.9).
}
# Graph 2: Provenance of the assertions.
grlod-np:provenance {
    grlod-np:assertion prov:generatedAtTime "2023-02-22"^^xsd:date ;
```

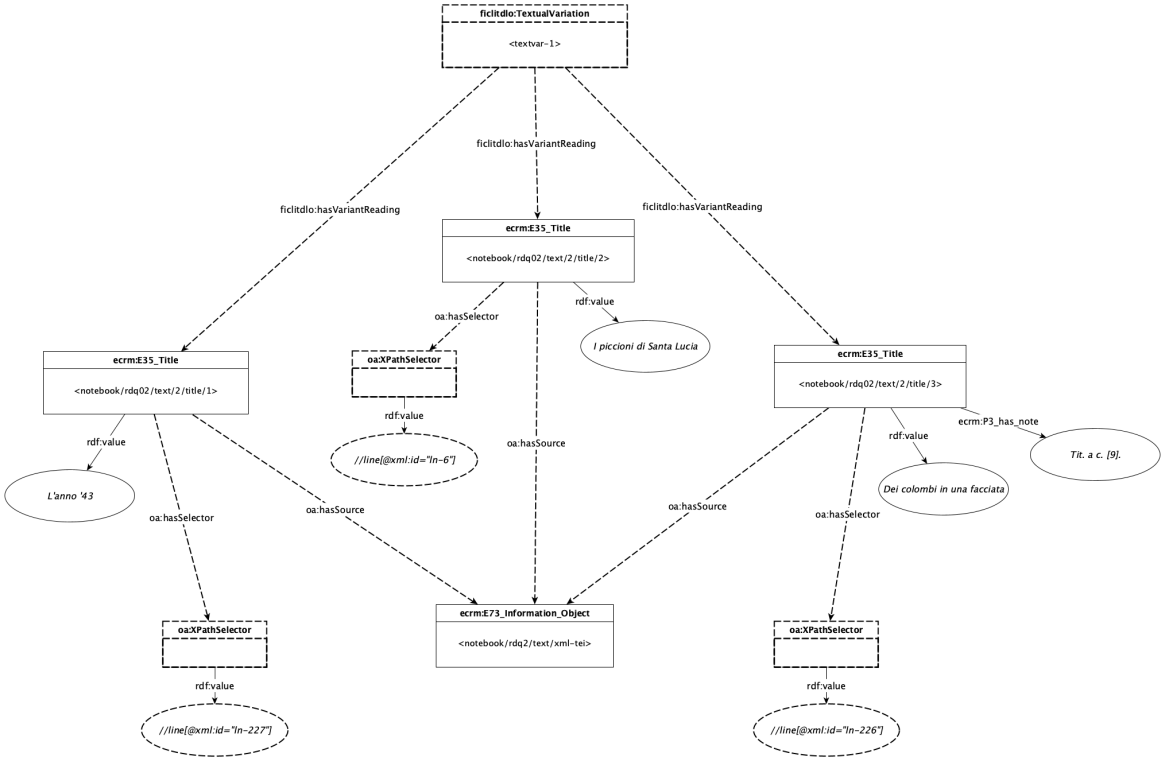

Figure 5.7: RDF representation of UC2, Part 2. Additional assertions by the researcher are displayed using dashed lines.

```
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> .
}
# Graph 3: Provenance of the nanopublication.
grlod-np:pubinfo {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub4>
        prov:generatedAtTime "2023-02-22"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> .
}
# Graph 4: The nanopublication and its components.
grlod-np:head {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub4>
        a np:Nanopublication ;
        np:hasAssertion grlod-np:assertion ;
        np:hasProvenance grlod-np:provenance ;
        np:hasPublicationInfo grlod-np:pubinfo .
}
```

Listing 5.3: RDF representation of provenance information for the supplemental graph of statements according to the nanopublication data model (Turtle syntax).

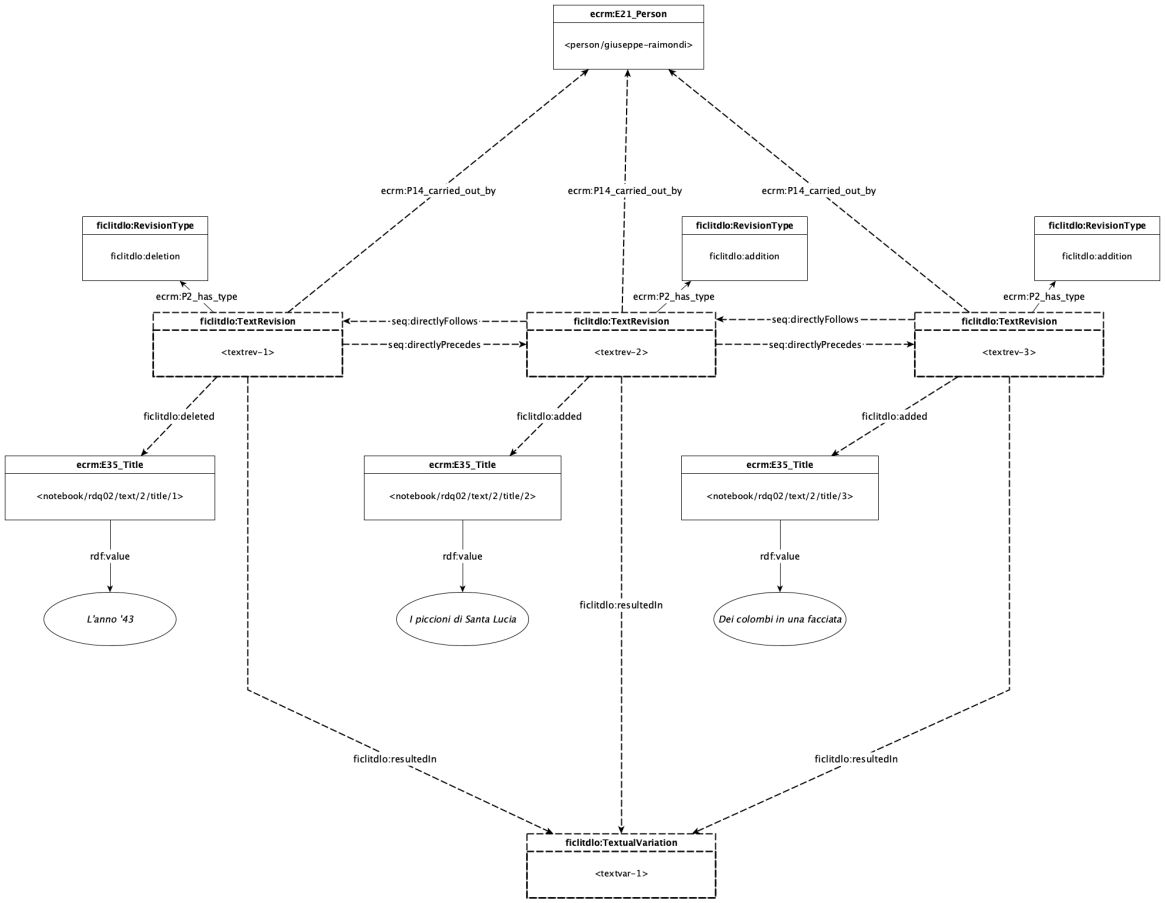

Figure 5.8: RDF representation of UC2, Part 3. Additional assertions by the researcher are displayed using dashed lines.

5.2.2 Queries and Results

Below is a list of possible queries for the archival knowledge after it has been augmented with reconstructions of the relationships between texts and text fragments. Each query has been formalised into SPARQL and run against the archival knowledge base. For each query, a tabular representation of the results is provided. [26]

CQ1 All titles are attributed to the notebook text.

CQ2 All published versions of the notebook text: title, date, and publisher.

CQ3 Copies of the published versions available within the Giuseppe Raimondi Archive.

CQ4 Textual variations (either internal or external) that involve the notebook text.

CQ5 For each textual variation in which the notebook text is involved, return all variant readings.

[26] The repository contains the formalisation of the queries in SPARQL for UC2 as well as the results. See https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/sparql and https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/results, respectively.

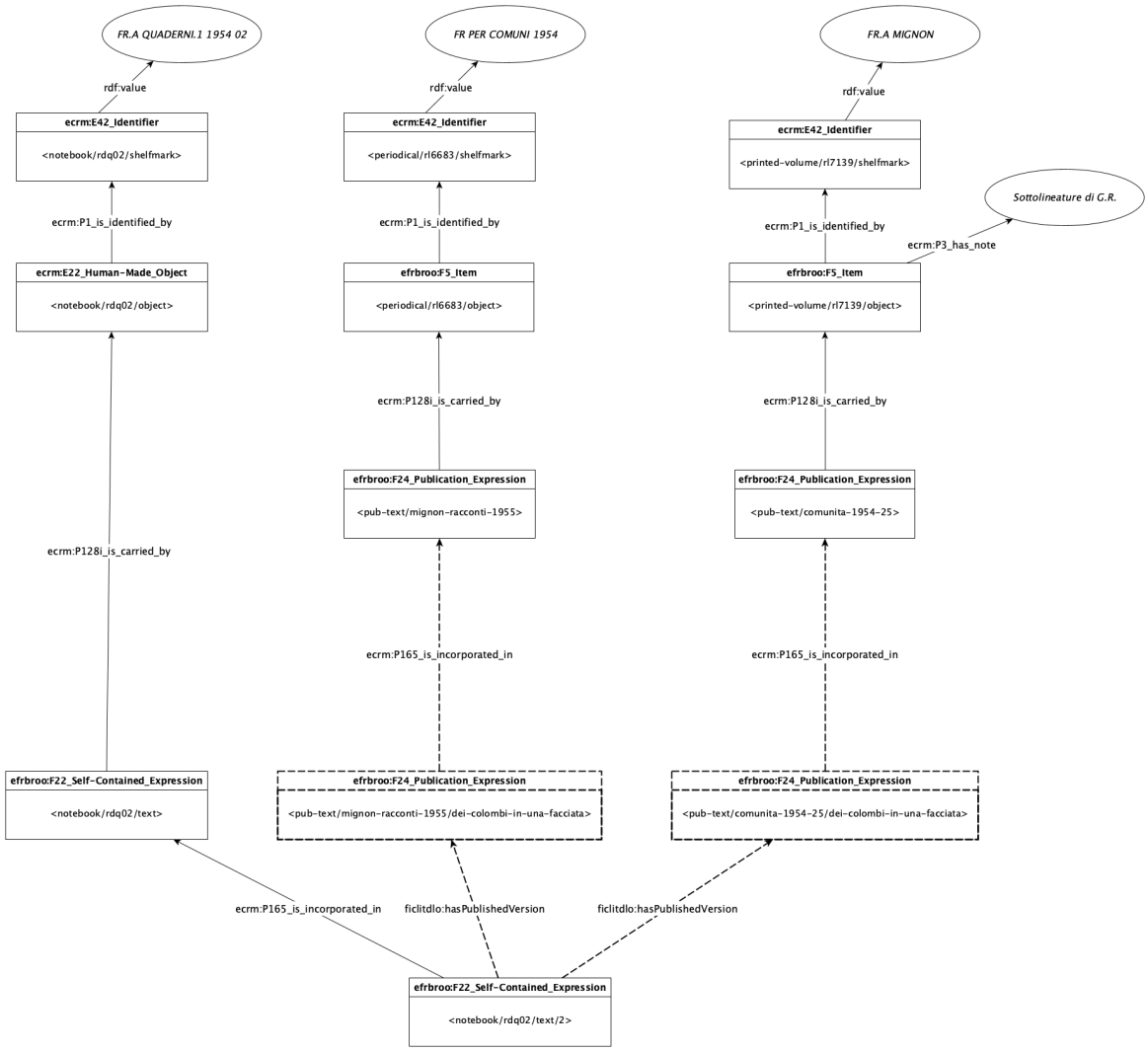

Figure 5.9: RDF representation of UC2, Part 4. Additional assertions by the researcher are displayed using dashed lines.

CQ6 Location of each variant reading in the transcription of the manuscript witness.

CQ7 Causes for the textual variation.

CQ8 Description of each text revision act.

CQ9 Chronological sequence of the text revision acts.

CQ10 Provenance of the statements about textual variation in the notebook text.

5.3 Use Case 3: Influence of Other Authors on Giuseppe Raimondi's Production

Rossi and Wenzlawski examined Giuseppe Raimondi's archive to reconstruct his relationship with the works and ideas of Paul Valéry.[27] This use case describes some aspects of their reconstruction. In particular, they suggest that the newspaper article 'Mostro a due Teste' by Giuseppe Raimondi (in *Corriere della Sera*, 28 ottobre 1971), was influenced by Paul Valéry's 'Préface pour une nouvelle traduction de La Soiréè avec M. Teste'[28] (in *Commerce*, 1925, n. 4). Supporting evidence for this connection is provided by the presence of handwritten annotations by Giuseppe Raimondi on his copy of the 'Préface'.[29] Such annotations read "Mostri, prodotti dei pensieri" and "Idee mostri". The article title 'Mostro a due Teste' is a possible reworking of these notes.

5.3.1 Use Case Representation According to the FICLITDL Conceptual Model

Figure 5.10 illustrates the representation of the author's activity as a reader and how this relates to his activity as a writer. Using this conceptualisation, it is possible to query the archive for works which the author had read before writing a specific work and, if this information is made available by one or more subsequent scholars, which works have influenced the creation of a particular work and on which basis the scholar asserts that (in the case of this use case, for example, evidence of influence is provided by the handwritten annotations left by Raimondi on the margins of the 'Préface'. The same notes have likely been reworked to create the title of the newspaper article 'Mostro a due Teste', published some decades later in *Corriere della Sera* on 28 October 1971). Giuseppe Raimondi's act of reading 'Paul Valéry's Préface pour une nouvelle traduction de La Soiréè avec M. Teste' in the first months of 1925 is represented as an E7 Activity of type 'reading'; the time frame in which the reading activity takes place is represented as an E52 Time-Span; the activity is linked to a specific text and a specific reader using the properties *ficlitdlo:relatesToText* and *P14 carried out by*. Besides being connected to the reading activity, Giuseppe Raimondi

[27] See Rossi and Wenzlawski, 'Nello scrittoio di Giuseppe Raimondi'.
[28] Paul Valéry, 'Préface pour une nouvelle traduction de La Soiréè avec M. Teste', *Commerce* 4 (1925): 93–102.
[29] BIFICLIT, FR PER COMMER 1925 (p. 100).

is also attributed the role in time (pro:roleInTime) of 'reader' of the 'Préface'.

The left half of Figure 5.10 shows the representation of the creatorship of the newspaper article 'Mostro a due Teste'. The creation of the publication expression of the article is represented as an instance of F28 Expression Creation which *R17 created* the publication expression – that is, the text of the article as it appeared on the newspaper; Giuseppe Raimondi is attributed the role in time of author of such a text. The creation of the text is said to be *P15 influenced by* the text of the 'Préface'. A textual explanation of the circumstances of creation of the article text 'Mostro a due Teste' is linked to the creation activity as a scholarly note via the property *P3 has note*.

Figure 5.11 continues Figure 5.10. The figure depicts Giuseppe Raimondi annotating a copy of the 'Préface' which is still physically available in his library.[30] The creation of the annotation is represented at the top of the figure as an instance of E65 Creation that was *P15 carried out by* Giuseppe Raimondi. The time frame in which allegedly Raimondi added the annotations to the text is represented as an E52 Time-Span. The tool which was used to create the annotation – a pencil — is specified using the *P125 used object of type* property (this is useful for isolating all annotations which were created using a specific pen, as this often indicates the time when the annotation was taken as, as we know, annotations may be added in layers over time as the author reads and re-reads a text). There are two distinct annotations that were created as a result of the annotation act: "Mostri prodotti dei pensieri" and "Idee mostri". Each annotation has a body (the text of the annotation) and a target (the 'Préface'). The annotation has also a starting and an ending page associated (in this case page 100 of the 'Préface'). The target text as well as the two author's annotations are linked to the copy of Paul Valéry's 'Préface' that carries them and that is comprised within the author's personal library. This representation is accompanied by a description of the provenance of the reconstruction, as shown below in Listing 5.4. Because this augmentation of the finding aid is not directly provided by the scholars who proposed the interpretation but is derived from the article in which they express such an interpretation, the provenance is attributed to the agent manipulating the dataset and is stated to be derived from the article.

```
# Graph 1: Researcher's assertions.
grlod-np:assertion {
    # The researcher's reconstruction of the relationships between the
```

[30] Giuseppe Raimondi's copy is available at BIFICLIT, FR PER COMMER 1925.

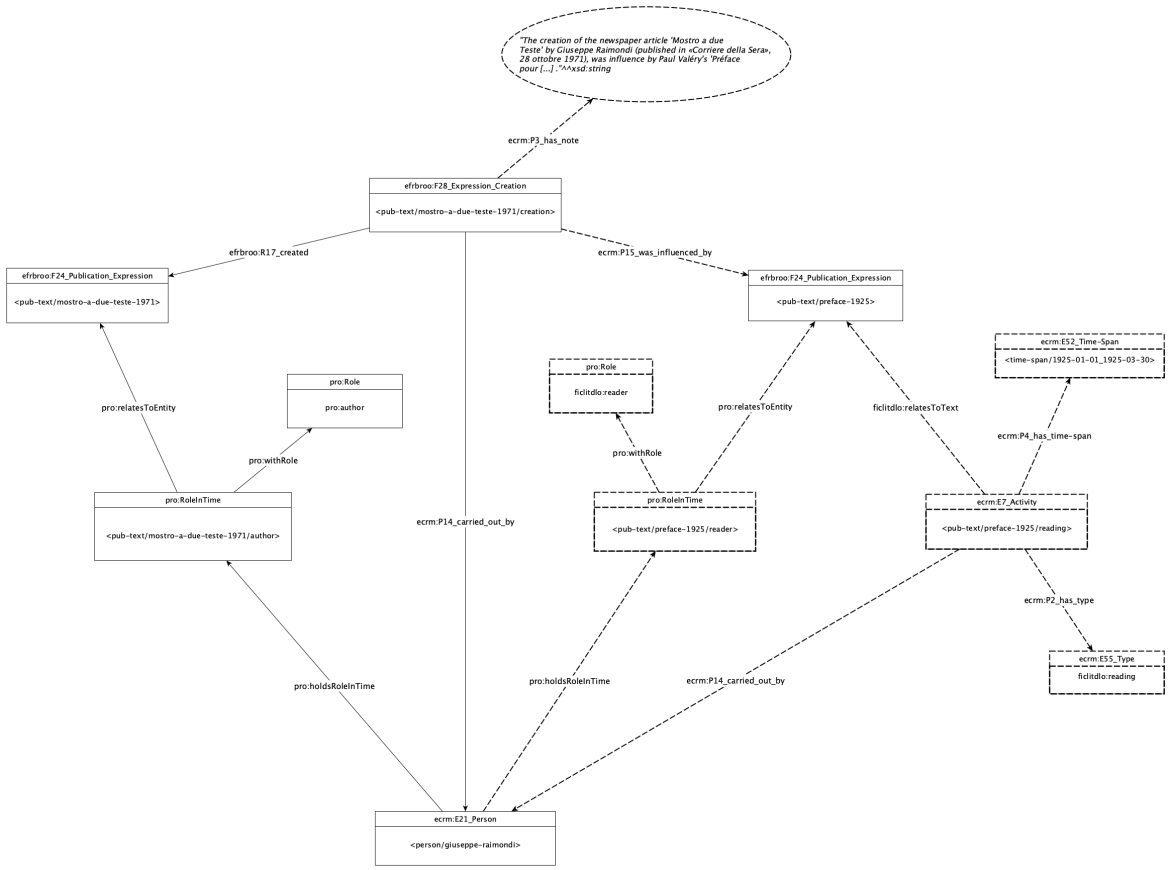

Figure 5.10: RDF representation of UC3, Part 1 (Raimondi reading Paul Valéry's 'Préface' in early 1925). Additional assertions by the researcher are displayed using dashed lines.

```
    texts (Figures 5.10-5.11).
}
# Graph 2: Provenance of the assertions.
grlod-np:provenance {
    grlod-np:assertion prov:generatedAtTime "2023-04-07"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> .
}
# Graph 3: Provenance of the nanopublication.
grlod-np:pubinfo {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub6>
        prov:generatedAtTime "2023-04-07"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> ;
        prov:wasDerivedFrom <https://doi.org/10.1400/276891> .
}
# Graph 4: the nanopublication and its components.
grlod-np:head {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub6>
        a np:Nanopublication ;
        np:hasAssertion grlod-np:assertion ;
```

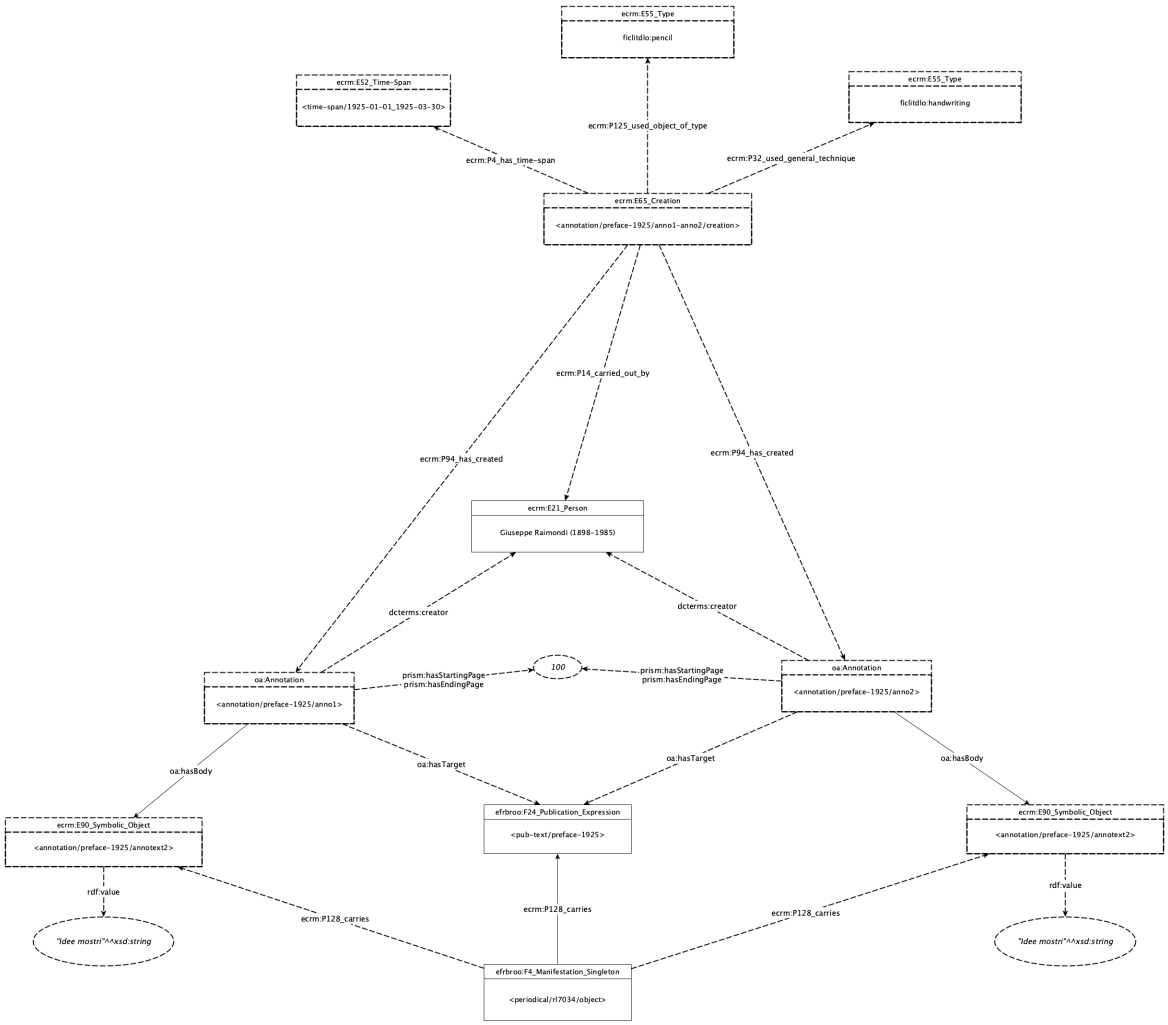

Figure 5.11: RDF representation of UC3, Part 2 (Raimondi annotating Paul Valéry's 'Préface', possibly in early 1925). Additional assertions by the researcher are displayed using dashed lines.

```
22          np:hasProvenance grlod-np:provenance ;
23          np:hasPublicationInfo grlod-np:pubinfo .
24  }
```

Listing 5.4: RDF representation of provenance information for the supplemental graph of statements according to the nanopublication data model (Turtle syntax).

5.3.2 Queries and Results

Below is a list of possible queries for the archival knowledge after it has been augmented with interpretations about the influence of Valéry on Raimondi. Each query has been formalised into SPARQL and run against the archival knowledge base. For each query, a tabular representation of the results is provided. [31]

[31] As with UC1 and UC2, the repository contains the formalisation of the queries in SPARQL for UC3 as well as the results. See https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/sparql and https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/

CQ1 What texts influenced Giuseppe Raimondi's work?

CQ2 What texts influenced a specific text by Giuseppe Raimondi?

CQ3 Which works by or about Paul Valéry did Raimondi read before writing a particular text?

CQ4 Which of these works has a copy in Giuseppe Raimondi's library?

CQ5 When did Raimondi most likely read a specific text?

CQ6 What annotations are present in a specific copy of a publication?

CQ7 Where are such annotations located?

CQ8 What type of annotations are they?

CQ9 When were the annotations most likely added to the text?

5.4 Use Case 4: Exchange of Letters and Other Items as Events

Analysing and understanding correspondence in personal archives is critical for reconstructing a complete picture of an author's life and activities. As was pointed out in Chapter 4, existing models are insufficient for representing the complexity of correspondence, which should be observed and represented as a complex event rather than a simple exchange of a single object between a sending and a receiving agent. Using the FICLITDL conceptual model, this use case illustrates how to represent correspondence as an event involving multiple entities. In Use case 3, I discussed the modelling of the influence of a text by Paul Valéry on the production of other texts by Giuseppe Raimondi. Use case 4 extends the reconstruction to include archival items from Giuseppe Raimondi's correspondence, which can aid in understanding the genesis and development of Raimondi's relationship with Valéry.

According to Rossi and Wezlaski, after reading the 'Préface' from 1925, Raimondi gains access to "La Soirée avec Monsieur Teste", Valéry's work that most fascinated the writer in the summer of 1925. It is Giuseppe Ungaretti who provides him with the book in August 1925, as evidenced by the handwritten note on the frontispiece: "Dono di Giuseppe Ungaretti a Gius. Raimondi. Bologna, 10.VIII.1925.". Raimondi, who had previously written two articles on Valéry for *Il Convegno*, will publish the

main/queries/results, respectively.

pamphlet *Il cartesiano signor Teste* in 1928.[32]

Raimondi sends all three works to Valéry, who responds with three letters written in Italian that are now available within Giuseppe Raimondi's archive.[33] The letters from Raimondi to Valéry are currently located in Valéry's personal archive at the Bibliothèque nationale de France (BNF).

5.4.1 Use Case Representation According to the FICLITDL Conceptual Model

Figure 5.12–5.14 depicts the letters Giuseppe Raimondi received from Paul Valéry. Because the archival description for correspondence and albums ends at the file level, the collection's base description does not include information about individual items. As a result, the entire representation is part of an additional contextualisation, which forms a separate graph.

Each letter representation is centred on a correspondence activity that allows multiple pieces of information, such as the agents involved, the date of dispatch, the places of origin and destination, and the items being sent, to be expressed in an interconnected manner. Furthermore, the representations include information about the language of the letters as well as links to transcriptions. The representations also state that each of the letters was published in 1955 as part of *La valigia delle Indie* and specify which text, either implicitly or explicitly, each letter refers to among those sent by Raimondi to Valéry. Listing 5.5 contains the representation of the provenance of the additional contextualisation described in Use case 4.

```
# Graph 1: Researcher's assertions.
grlod-np:assertion {
    # The researcher's reconstruction of the relationships between the
   texts (Figures 5.12-5.14).
}
# Graph 2: Provenance of the assertions.
grlod-np:provenance {
    grlod-np:assertion prov:generatedAtTime "2023-06-09"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> .
}
```

[32] Giuseppe Raimondi, 'Divagazioni intorno a Paul Valéry', *Il Convegno* 6, no. 2 (1925): 90–97; Giuseppe Raimondi, 'Ringraziamento a Commerce', *Il Convegno* 6, no. 9 (1925): 487–491; Giuseppe Raimondi, *Il cartesiano signor Teste*, Edizioni di Solaria (Firenze, 1928).

[33] The letters are also published in Giuseppe Raimondi, *La valigia delle Indie* (Firenze: Vallecchi, 1955), 350–57.

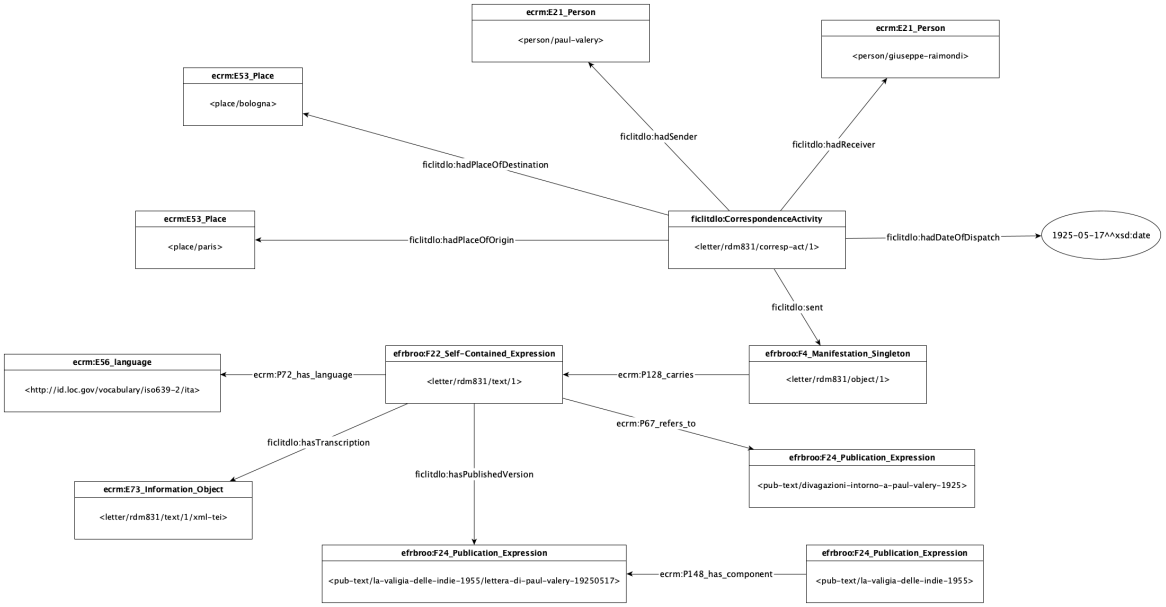

Figure 5.12: RDF representation of UC4, First letter by Paul Valéry.

```
# Graph 3: Provenance of the nanopublication.
grlod-np:pubinfo {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub7>
        prov:generatedAtTime "2023-06-09"^^xsd:date ;
        prov:wasAttributedTo <https://orcid.org/0000-0001-6007-9118> ;
        prov:wasDerivedFrom <https://doi.org/10.1400/276891> .
}
# Graph 4: the nanopublication and its components.
grlod-np:head {
    <https://w3id.org/giuseppe-raimondi-lod/nanopub/nanopub7>
        a np:Nanopublication ;
        np:hasAssertion grlod-np:assertion ;
        np:hasProvenance grlod-np:provenance ;
        np:hasPublicationInfo grlod-np:pubinfo .
}
```

Listing 5.5: RDF representation of provenance information for the supplemental graph of statements according to the nanopublication data model (Turtle syntax).

5.4.2 Queries and Results

A number of possible queries are listed below. The queries formalised into SPARQL and the corresponding results are available from https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/sparql and https://github.com/fgiovannetti/giuseppe-raimondi-lod/tree/main/queries/results, respectively.

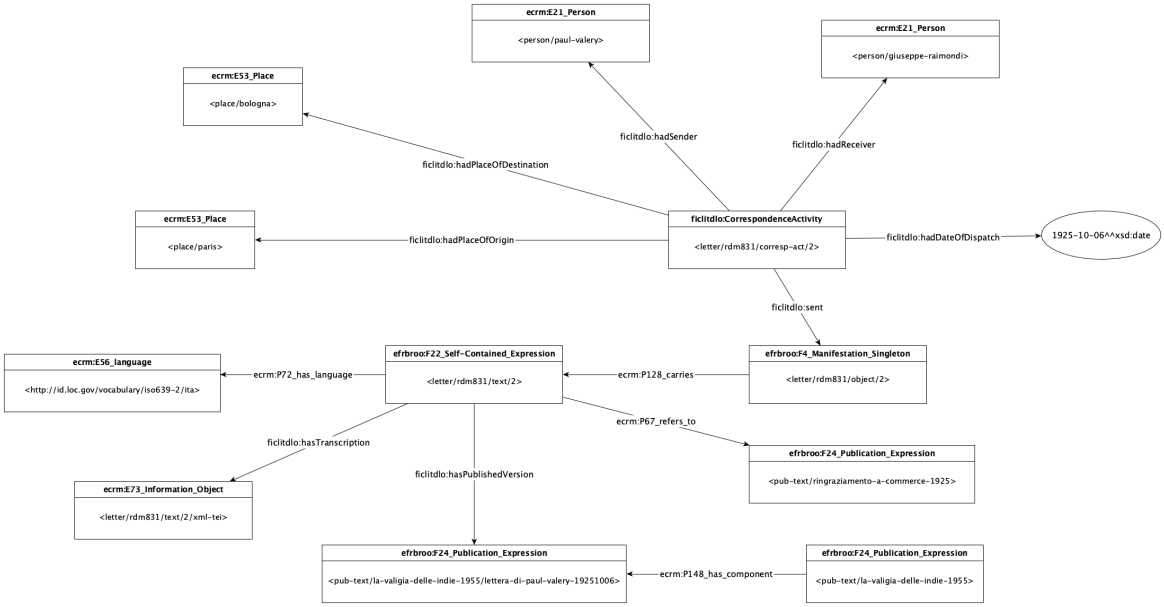

Figure 5.13: RDF representation of UC4, Second letter by Paul Valéry.

CQ1 Is there any correspondence between Paul Valéry and Giuseppe Raimondi in the archive?

CQ2 What is the chronological order of the letters?

CQ3 When were the letters mailed?

CQ4 What is the language of the letters?

CQ5 Do the letters mention any specific work by Raimondi?

CQ6 Where did the letters come from?

CQ7 Are the letters published anywhere?

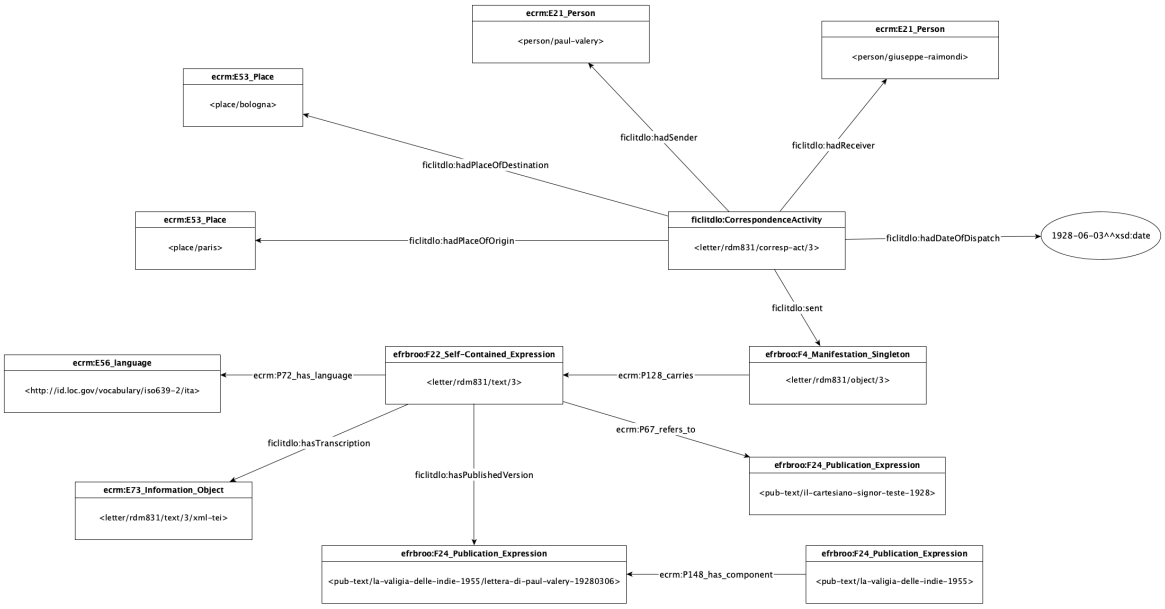

Figure 5.14: RDF representation of UC4, Third letter by Paul Valéry.

Chapter 6

Discussion and Conclusions

This study grew out of a larger goal of improving the descriptions of personal archives and libraries belonging to academic and literary writers held and maintained at the Ezio Raimondi Library of the FICLIT Department of the University of Bologna. The primary objective of the research was to establish a conceptual model for representing these types of personal archives, the applicability of which could be extended to similar personal archives held by other cultural institutions. Furthermore, this study provided an opportunity to consider the implications and potential of using knowledge graphs as the primary mode of archival representation for cultural institutions. It looked specifically at the possibility of subsequent researchers enriching the finding aid with additional interpretations after the archivist had completed the initial stages of arrangement and description.

This research adopted Giuseppe Raimondi's personal archive and library as a representative case study for developing the conceptual model. The study concentrated on writers' archives housed in a department of classical philology and Italian studies, which are primarily studied and consulted by textual and literary scholars. As a result, it was of particular interest to investigate how these specific researchers could effectively co-construct the archival description by representing their findings about the collections and authorship in general for the benefit of new readers and users of the archive. This study investigated the enhancement of archival descriptions with additional readings by subsequent researchers. It attempted to provide answers to a number of related questions, such as how to integrate 'a' reading of the archive with archival description. How should such a reading be represented so that new users of the archive can explore and evaluate it? How can conflicting, yet coexisting, readings of the archive be represented by ontologies?

To accomplish this, it was necessary to consider how to represent the many different kinds of objects found in such archives, such as notebooks, books, letters, postcards, and so on. It also entailed dealing with the challenge of describing these objects in a multidimensional manner, which necessitated the use of a multi-entity representation that separated the carrier (i.e. the physical object or document) from its content (i.e. the text in its broadest sense). Using this method, different perspectives on the same document could be provided within the same finding aid, effectively transforming it into an archival knowledge base. It also allowed for the description of objects at various levels of granularity, ranging from sets of documents to specific parts of a document, such as a particular section within a document, or even elements of content, such as a specific textual fragment in a notebook. The ability to associate transcriptions of texts with the representation, alongside facsimiles, was critical in representing authorial archives. This was deemed necessary to ensure the authority and reliability of the transcriptions, thereby improving the archival description. The research questions that guided this phase of the study were how semantic web technologies can be used to represent heterogeneous types of documents and relationships in a multidimensional way, particularly in complex scenarios involving multiple interconnected documents, and how records and their relationships can be described at various granularities, such as record sets, individual records, record parts, and content and text.

The concept of subsequent and continuous enrichment of the archival description — which, in practise, translates into the technical and methodological possibility of augmenting the finding aid over time while ensuring that provenance information, and thus the authoritativeness of the representation, is always explicit — has a strong connection with the literature and theory of digital scholarly editions. Both Shillingsburg and Gabler describe digital scholarly editions as 'knowledge sites',[1] relationally-structured research platforms that give readers the opportunity to interact with the material in a fluid and dynamic manner, where

> each user can choose an entry way, select a congenial set of enabling contextual materials, and emerge with a personalized interactive form of the work [...] always able to plug in for more information or different perspectives.[2]

[1] Hans Walter Gabler, 'Thoughts on Scholarly Editing', Publisher: Società editrice il Mulino, *Ecdotica* 7, no. 1 (2010): 15; Peter L. Shillingsburg, *From Gutenberg to Google: Electronic representations of literary texts* (Cambridge University Press, 2006), 88.

[2] Shillingsburg, *From Gutenberg to Google*, 88.

In light of this perspective, Rasmussen identifies and describes at least three different roles or functions for readers of a digital scholarly edition: the reader, the user, and the co-worker. She explains:

> The *reader* is mainly interested in scholarly editions as reliable academic versions of literary works [...]. The *user* also seeks an understanding of the work, but in a more intertextual context, where stress is placed either on the relation between the work's numerous texts and versions, or on the relation between the work's own texts and other texts that explain or relate to the work [...]. The *co-worker* seeks to go beyond the user and reader roles, and to contribute actively to the scholarly enterprise. This reader role could also have been called contributor or the like, but the term co-worker signals that, ideally speaking, the reader in this role is likely to take part in the editorial work at some level. This could consist of making annotations, reading proofs, adding encodings, or contributing in other ways to the site's total production of knowledge. The co-worker's contribution, in short, does not merely amount to additions or extensions, but forms a genuine part of the edition.[3]

These roles are characterised by an incremental involvement of the reader. Clearly, one role does not preclude the other: readers can also be users and co-workers (or, as I prefer, co-constructors) of the digital edition at the same time.

This perspective on digital scholarly editions as fluid, open, and constantly evolving environments, laboratories in which users can participate, can also be applied to digital finding aids that serve as digital surrogates for physical archives. However, in order to be able to do so, a rethinking of the concept of finding aid in terms of a network of variously interconnected entities was required.

To that end, the primary goal of this study was to develop a conceptual model for the representation of writers' archives that would allow for the multidimensional and integrated representation of heterogeneous types of documents and relationships within and across collections; the representation of archival records at various and multiple levels of granularity (i.e., from sets of documents to individual docu-

[3] Krista Stinne Greve Rasmussen, 'Reading or Using a Digital Edition? Reader Roles in Scholarly Editions', in *Digital Scholarly Editing : Theories and Practices*, ed. Matthew James Driscoll and Elena Pierazzo, Digital Humanities Series, Code: Digital Scholarly Editing : Theories and Practices (Cambridge: Open Book Publishers, 13 July 2017), 127, accessed 26 May 2023, http://books.openedition.org/obp/3406.

ments, document components, and document content); the incorporation of additional, possibly conflicting, contextualisations into the same finding aid.

The findings of this study support the hypothesis that existing ontologies widely used in the GLAM domain, such as CIDOC-CRM, FRBRoo, PROV-O, and others, can be easily extended with a small number of additional classes and properties to effectively represent heterogeneous types of writers' documents within archival collections, as well as the hierarchical relationships that typically characterise paper-based archives (i.e. forms part of/is composed of). The adoption of CIDOC-CRM as a base conceptual model allowed also a reconsideration of the modelling of the concept of collection versus fonds as suggested by Yeo.[4] While the former should be intended as an aggregate of physical archival documents, the latter can be seen as a conceptual entity that is the result of a process of scholarly interpretation and reconstruction by the archivist or researcher. As a result, in order to represent the concept of fonds (and its subconcepts of series and file), a new object property 'forms conceptually part of' was introduced, allowing the archival collection as it is physically organised as well as one or more conceptual fonds or reconstructions for such a collection to be represented.

Furthermore, the evaluation of the RDF representation of Giuseppe Raimondi's archive, which used competency questions, validates the hypothesis that data-centric representations far exceed document-centric approaches that belong to systems such as XML/EAD for representing the archival domain. RDF supports the representation of archival entities at multiple levels of granularity, allowing for more flexible and comprehensive descriptions. In addition, unlike hierarchical representations, RDF-based data models allow entities to be assigned to multiple classes at the same time, increasing the richness and expressiveness of the archival representation.

The use of established models in this study, such as nanopublications, supports the hypothesis that leveraging these models allows for the seamless integration of multiple interpretations into a unified archival knowledge base. This method improves data interoperability and accessibility. Furthermore, adding provenance information to RDF assertions improves the archival knowledge base's reliability and transparency, ensuring its authoritativeness for future use and analysis.

The study produced a model that allows users to create archival representations that remain fluid and provide multiple perspectives on the collection while evolving over time. This model goes beyond traditional approaches by explicitly accommod-

[4] Yeo, 'The Conceptual Fonds and the Physical Collection'.

ating various provenances, putting the archivist and other creators at the centre of the description. The subjectivity of the description is explicitly conveyed, distinguishing between those asserting something about the collection and those producing the representation.

Compared to the transition to XML that most archival institutions have already experienced, the transition to LOD entails a significant methodological shift in addition to a technological one. It necessitates a complete rethinking of archival representation as a process as well as a product. The concept of a finding aid is reimagined in this study as an open data-centric representation that can evolve over time while retaining versioning, rather than as a static document. This is especially important given the numerous transformations that a description undergoes as a result of changing technological devices over time.

Archival description is no longer constrained by the two-dimensionality of paper, nor by the limitations imposed by XML as a document-centric and hierarchical system. On the contrary, by using RDF, it liberates itself because RDF allows for the possibility — and this is the keystone — of treating anything as data and thus making anything describable, including metadata, in a chain of descriptions of descriptions, allowing for sophisticated representations. The possibility of adding descriptive layers on top of existing ones in the knowledge base can increase the complexity of the representation; however, the cultural institution determines the degree of complexity, as RDF is a straightforward mechanism based on direct relationships between subjects and objects.

The benefits of using linked open data for archival representations have been recognised by cultural institutions tasked with archival collection preservation. The International Council for Archives (ICA) has made significant contributions to this domain by creating a conceptual model and ontology for archival description that is intended to serve as a global standard. These efforts have been motivated by a desire to free archival description from the constraints of a hierarchical approach, in favour of a graph-based methodology. While the ICA's efforts to standardise are commendable, they missed opportunities to align their archival description standard with the most recent archival thinking on representation. Indeed, the archival literature on representation views the archivist as a central figure in the representation process, which is never neutral and is always the result of a subjective modelling process.

Finding aids have the potential to be transformed into cohesive and interconnected research platforms that greatly enhance discoverability and facilitate novel

modes of reading and exploration. This transformation can be achieved by harnessing the power of linked open data and semantic web technologies. Adopting LOD as the primary mode of archival representation provides several significant benefits, including but not limited to the following:

- A graph-based archival representation provides users with multiple access points to explore and navigate the archive rather than just the single named creator and the hierarchy that originates from it, allowing for a more versatile and adaptable discovery experience. In contrast to schemas, ontologies offer a mechanism for developing semantic vocabularies that can be expanded gradually. Likewise, the triple-based structure of RDF enables the seamless addition, removal, or modification of new statements in the knowledge base without disrupting the existing data within the dataset. This flexibility allows for the continuous evolution of the dataset while maintaining the integrity and coherence of the information it encompasses.

- Semantic web technologies enable the integration of diverse archival resources by facilitating semantic and syntactic interoperability. This capability allows archives to be conceptually interconnected and harmonised, ensuring information integration and exchange. By utilising these technologies, archives can overcome the limitations of disparate data formats and structures, creating a cohesive and interconnected ecosystem in which information can transcend the boundaries of a single collection or institution and be understood in a unified manner. This knowledge integration and exchange can help scholars and users better understand archival materials and improve their research experience.

- Moreover, these technologies offer mechanisms for representing parallel provenance, enabling the incorporation of multiple contextualisations within the same archival description. As a result, different aspects, perspectives, or historical contexts can be captured and associated with the archival resources, enriching their overall understanding and facilitating new research insights. The ability to represent parallel provenance enhances the richness and complexity of archival descriptions, allowing for a more comprehensive and nuanced portrayal of the materials held within the archive.

- The integration of textual scholarship with archival description is made pos-

sible by linked open data, which fosters collaboration and mutual enrichment between these domains. Scholars can connect archival materials to relevant textual resources such as critical editions, annotations, or commentaries enhancing the research experience and allowing for new insights.

This study has contributed to the development of a conceptual model for representing personal archives that can be collaboratively constructed by multiple scholars over time. However, two issues still remain unresolved. The first is related to the visualisation and effective navigation of the archival knowledge base by users who are not proficient in semantic web technologies. Thus, there is a need to address the challenge of defining a user-friendly and efficient interface. The second issue is linked to the incorporation of further interpretations into the knowledge base in the form of additional triples.

When it comes to interfaces, experiments conducted with Omeka S, described by its creators as a "web publishing platform for institutions [universities, galleries, libraries, archives, and museums] interested in connecting digital cultural heritage collections with other resources online", have led to the conclusion that this platform lacks the effectiveness required to display highly complex knowledge graphs, particularly those containing named graphs.[5] It is important to note that the underlying database of Omeka S does not rely on semantic web technologies; instead, it uses a SQL database for data storage. Even the provision of the option for users to download data in any of the existing RDF serialisations poses a complex task for interface programmers. Furthermore, the visualisation of multidimensional descriptions, which entails presenting different perspectives of an item from the physical standpoint, content perspective, etc., proves to be inefficient. Each perspective is displayed on separate pages, posing challenges in terms of navigation and integration between them. As of today, the most efficient approach for using the archival knowledge base of Giuseppe Raimondi's archive and library is accessing and querying the complete dataset using SPARQL. ResearchSpace, a free and open-source platform maintained by the British Museum, may represent an interesting alternative for future exploration as it leverages actual knowledge graphs as a method for storing data.[6]

The second unresolved issue with multiperspective archival knowledge bases is the enhancement of the base description with additional contextualisations. As pro-

[5] Omeka S homepage, https://omeka.org/s/.
[6] ResearchSpace —representing knowledge with context, https://researchspace.org.

posed by Murray and Tillet, these operations involve the creation of new graphs, where each graph represents a distinct perspective on the archive or a specific portion of it.[7] However, literary and textual researchers commonly lack proficiency in linked open data technologies, as well as the computational skills required to directly interact with a knowledge base via SPARQL. How could the addition of new scholarly interpretations in the form of graphs be managed and facilitated?

There are at least three potential solutions to address this problem. The first option involves engaging a specialist in semantic web technologies who can convert the scholarly interpretation provided by the domain expert into a knowledge graph. If they are trained in digital humanities, this specialist could be an archivist or librarian. However, it is important to note that this process adds an additional interpretative layer to the archival description, as the digital humanist's transformation may improperly change the intended meaning of the domain expert's interpretation. While the likelihood of distortion decreases when the domain expert and digital humanist collaborate closely on the creation of the knowledge graph, it remains necessary to explicitly express the role of the digital humanist in the dataset. This can be accomplished through the use of nanopublications, which provide metadata detailing the creation and dissemination of the nanopublication itself, providing a suitable method for clarifying the digital humanist's involvement.

The second option entails developing and using a user interface that allows researchers with limited expertise in semantic web technologies to easily contribute new triples to the dataset.[8] The proposal focuses on user annotation of existing knowledge bases, including archival ones. When domain experts propose new semantic relationships via the interface, they anticipate that the annotation will add to the knowledge base while adhering to the underlying conceptual model. In this context, the annotation application acts as an intermediary between the domain expert and the formal representation of the knowledge being inserted, thereby becoming a critical component of the knowledge creation and management chain.[9] It is ne-

[7] Murray and Tillett, 'Cataloging theory in search of graph theory and other ivory towers', 172, specifically defines resource description as graph creation.

[8] This solution is currently being researched and will be presented at DH2023, the Alliance of Digital Humanities Organisations (ADHO) annual conference, which will take place in Graz, Austria, from July 10 to 14. The paper is co-authored by Paolo Bonora, Martina Dello Buono, Francesca Giovannetti, and Francesca Tomasi. The description of the project is a realaboration of the long abstract presented for the conference.

[9] Paolo Bonora et al., 'Tell Me the Truth: Validating the Semantic Alignment between the Annotation User Interface and the Knowledge Base', in *DH2023: Conference Abstracts*, DH2023 Collaboration as Opportunity (2023), accessed 31 May 2023.

cessary to ensure semantic alignment between the annotation user interface and the resulting knowledge graph. On the one hand, the user should validate the triples representing the annotation, while the system should ensure that any subsequent representation of the newly added triples preserves the original meaning of the annotation.

To perform the initial validation, we propose using RDF-to-Text techniques, which provide the user with natural language translations of the triples in addition to the RDF representation.[10] Once the user confirms the correctness of the assertions, they will be added to the knowledge base.[11] The goal of this research is to develop a methodology for assessing the semantic transparency of user interfaces in annotation applications. Any layer that exists between the knowledge base and the user acts as a "semantic lens" that can introduce semantic distortions.[12] Our proposed approach employs selected use cases to investigate whether this phenomenon is detectable and, if so, how it can be measured using user feedback.[13]

A third solution, which is valid for textual interpretations, is the extraction of knowledge graphs from TEI-encoded documents. The text is marked up in TEI by the domain expert, and an automatic script converts TEI components into RDF triples. Experiments with this operation have been carried out as part of the development of LIFT (Linked Data from TEI), a tool made up of a set of adaptable scripts for extracting RDF triples from TEI encoded documents, available at https://projects.dharc.unibo.it/lift/.[14] This solution could be used to extract textual variation information from the TEI transcriptions of documents available in the archival knowledge base.

As stated in previous chapters, the current study only looked at the archives of twentieth-century, paper-based writers. The FICLIT Department will most likely acquire more writers' and scholars' archives in the near future, and it is important to

[10] On RDF-to-Text techniques see, for example, Axel-Cyrille Ngonga Ngomo, Diego Moussallem and Lorenz Bühmann, 'A Holistic Natural Language Generation Framework for the Semantic Web', Publisher: arXiv Version Number: 1, 2019, accessed 26 April 2023, https://doi.org/10.48550/ARXIV.1911.01248.

[11] A prototype of the annotation interface, produced by Martina Dello Buono, is accessible at https://tell-me-the-truth.github.io/web-app/.

[12] Silvio Peroni et al., 'Semantic Lenses as Exploration Method for Scholarly Articles', in *Bridging Between Cultural Heritage Institutions*, ed. Tiziana Catarci, Nicola Ferro and Antonella Poggi, vol. 385, Series Title: Communications in Computer and Information Science (Berlin, Heidelberg: Springer Berlin Heidelberg, 2014), The concept of 'semantic lens' is taken from, accessed 31 May 2023, https://doi.org/10.1007/978-3-642-54347-0_13.

[13] Bonora et al., 'Tell Me the Truth: Validating the Semantic Alignment between the Annotation User Interface and the Knowledge Base'.

[14] See Giovannetti and Tomasi, 'Linked data from TEI (LIFT)'.

note that these archives will almost certainly be in a hybrid or entirely digital format, necessitating significant changes to the conceptual model developed for this study.

It is also important to consider the complexity and variability of archives when attempting to generalise the findings of this study. Archives, like texts, are complex systems with unique features. The transformation scripts developed in the framework of this study may need to be modified to accommodate document types other than those considered for this research, such as notebooks, albums, loose manuscripts, postcards, letters, and annotated printed volumes. This is due to the fact that no two writers' archives are the same.

This study contributes significantly to the investigation of creating and using finding aids for personal archives in the form of knowledge graphs by focusing on a specific category of personal archives, namely the archives of twentieth-century writers and intellectuals, and on selected types of archival documents. By providing a foundation for the use of knowledge graphs to represent personal archives, the conceptual model holds potential for application to a wide range of archival collections, albeit with necessary adjustments to account for the unique characteristics of each collection.

Conclusions

> Le parole chiave per il prossimo futuro della descrizione archivistica, se siamo davvero interessati a rinforzarne il valore pubblico, devono essere multidimensionalità, condivisione, apertura e connessione semantica con le informazioni attendibili in rete, anche non di fonte archivistica, che rappresentano l'evidenza creativa, amministrativa, politica e dialogica della vita individuale e sociale nel tempo.[15]

Even in the digital age, the finding aid remains an essential tool for discovery and access. Its reconfiguration as an archival knowledge base has the potential to transform the finding aid into an expanding research platform that can explicitly represent complex interrelationships between heterogeneous entities, as recommended by Feliciati.[16]

The primary goal of this research was to develop a conceptual model for representing writers' and scholars' archives that allows for the expression of such inter-

[15] Feliciati, 'Archives in a Graph', 101.

[16] This section is a re-elaboration of the concluding remarks presented in Giovannetti and Tomasi, 'The linked finding aid as a platform for textual research'.

relationships, with a particular emphasis on connecting texts. The model, through the use of named graphs and, in particular, nanopublications, enables the continuous enrichment of the digital finding aid with subsequent scholarly reconstructions of the context(s) characterising the records. Indeed, when reconceived as a knowledge base, the finding aid not only serves as the starting point for the research cycle within the archive it represents, but it also serves as the end point of such a cycle, as new knowledge gained about the archive along the way contributes to the growth and enrichment of the archival description itself.

Using the Giuseppe Raimondi Fonds as a case study, this contribution sought to demonstrate how the use of linked open data in conjunction with event-based and provenance-centric archival descriptions can broaden the role of the digital finding aid, transforming it into an ever-evolving research platform, open to dialogue with digital scholarly editions and other cultural resources on the web.

Bibliography

Allemang, Dean, and James Hendler. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier, 2011.

Archivrat, Internationaler, and International Council of Archives, eds. *ISAD(G): General international standard archival description; adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 Sept. 1999*. 2. Ed. ICA Standards. Ottawa: International Council of Archives, 2000.

Bailey, Jefferson. 'Disrespect des Fonds: Rethinking Arrangement and Description in Born-Digital Archives'. *Archive Journal*, 3rd ser., 2013. Accessed 16 March 2021. http://dev.archivejournal.net/?p=4722.

Bearman, David A. 'Documenting Documentation'. *Archivaria* 34 (1992).

Bekiari, Chryssoula, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian Emil Ore, Stephen Stead and Athanasios Velios. *Definition of the CIDOC Conceptual Reference Model (Version 7.1.2)*. 2022.

Bekiari, Chryssoula, Martin Doerr, Patrick Le Boeuf and Pat Riva. *Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism*. 2.4. IFLA, 2015.

———, eds. *LRMoo (formerly FRBRoo) object-oriented definition and mapping from IFLA LRM (version 0.7)*, June 2021.

Berners-Lee, Tim. 'Design issues: Linked data'. *http://www.w3.org/DesignIssues/LinkedData*, 2006.

———. 'Semantic Web on XML'. XML 2000 Washington DC, 6 December 2000. Accessed 24 April 2023. http://www.w3.org/2000/Talks/1206-xml2k-tbl.

Bizer, Christian, Tom Heath and Tim Berners-Lee. 'Linked data: The story so far'. In *Semantic services, interoperability and web applications: emerging concepts*, 205–227. IGI global, 2011.

Boles, Frank. 'Disrespecting Original Order'. Publisher: Society of American Archivists, *The American Archivist* 45, no. 1 (1982): 26–32. Accessed 26 March 2021. http://www.jstor.org/stable/40292457.

Bonora, Paolo, Martina Dello Buono, Francesca Giovannetti and Francesca Tomasi. 'Tell Me the Truth: Validating the Semantic Alignment between the Annotation User Interface and the Knowledge Base'. In *DH2023: Conference Abstracts*. DH2023 Collaboration as Opportunity. 2023. Accessed 31 May 2023.

Bountouri, Lina, and Manolis Gergatsoulis. 'Mapping encoded archival description to CIDOC CRM'. In *First Workshop on Digital Information Management*, 8–25. 2011.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT press, 1999.

Caproni, Attilio. 'Biblioteca privata: ipotesi di definizione'. *Bibliotheca* 5, no. 1 (2006): 22–28.

———. 'Le biblioteche d'autore: definizione, caratteristiche e specificità: alcuni appunti'. *Bibliotheca* 3, no. 1 (2004): 15–22.

Ciula, Arianna, and Øyvind Eide. 'Modelling in digital humanities: Signs in context'. Publisher: Oxford University Press, *Digital Scholarship in the Humanities* 32 (2017): 33–46.

Comments on "Records in Context". InterPARES Trust, 10 December 2016.

Conti, Eleonora, ed. *Giuseppe Ungaretti. Lettere a Giuseppe Raimondi, 1918-1966*. Vol. 53. Pàtron, 2004.

Cook, Terry. 'The Concept of the Archival Fonds in the Post-Custodial Era: Theory, Problems and Solutions'. *Archivaria* 35 (1992).

Cremante, Renzo. 'Introduzione'. In *Conservare il Novecento: convegno nazionale, Ferrara, Salone internazionale dell'arte del restauro e della conservazione dei beni culturali e ambientali, 25-26 marzo 2000: atti*, edited by Maurizio Messina and Giuliana Zagra, 53–57. Roma: Associazione italiana biblioteche, 2001.

Crocetti, Luigi. 'Indicizzare la libertà'. *Biblioteche oggi* 20, no. 2 (2002): 8–11.

———. 'La tradizione culturale italiana del Novecento'. *AIAA Journal* 17, no. 9 (1999).

———. 'Parole introduttive a Conservare il Novecento'. In *Conservare il Novecento: convegno nazionale, Ferrara, Salone internazionale dell'arte del restauro e della conservazione dei beni culturali e ambientali, 25-26 marzo 2000: atti*, edited by Maurizio Messina and Giuliana Zagra, 23–26. Roma: Associazione italiana biblioteche, 2001.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead and Matthew Stiff, eds. 'CIDOC CRM: CIDOC Conceptual Reference Model', November 2011. http://www.cidoc-crm.org/cidoc-crm/.

Daquino, Marilena, Francesca Giovannetti and Francesca Tomasi. 'Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini'. *Umanistica Digitale*, no. 7 (2019): 49–75. Accessed 19 December 2019.

Daquino, Marilena, Francesca Mambelli, Silvio Peroni, Francesca Tomasi and Fabio Vitali. 'Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data'. Publisher: ACM New York, NY, USA, *Journal on Computing and Cultural Heritage (JOCCH)* 10, no. 4 (2017): 1–21.

Daquino, Marilena, and Francesca Tomasi. 'Linked Cultural Objects: dagli standard di catalogazione ai modelli per il web of data. Spunti di riflessione dalla Fototeca Zeri'. Number: 1, *Umanistica Digitale*, no. 1 (1 October 2017). Accessed 10 February 2021. https://doi.org/10.6092/issn.2532-8816/7195.

De Pasquale, Andrea. 'Private archives in the library. Types, acquisition, treatment and description'. 10, no. 3 (2019): 34–46.

Dekker, Ronald H., and Gregor Middell. *CollateX, Version 1.7.1*. In collaboration with The Interedition Development Group, 2015. https://collatex.net/.

Desideri, Laura. 'Le raccomandazioni di Luigi Crocetti'. In *Conservare il Novecento: gli archivi culturali. Ferrara, Salone internazionale dell'arte del restauro, 27 marzo 2009: atti del convegno*, edited by Laura Desideri and Giuliana Zagra. Roma: Associazione italiana biblioteche, 2010. https://doi.org/10.1400/285768.

Doerr, Martin, and Nicholas Crofts. 'Electronic Communication on Diverse Data - The Role of the oo CIDOC Reference Model', 1998.

Douglas, Jennifer. 'A call to rethink archival creation: exploring types of creation in personal archives'. Publisher: Springer, *Archival Science* 18, no. 1 (2018): 29–49.

Douglas, Jennifer Lynn. 'Archiving Authors: Rethinking the Analysis and Representation of Personal Archives'. PhD diss., University of Toronto, 2013.

Duff, Wendy M., and Verne Harris. 'Stories and names: archival description as narrating records and constructing meanings'. *Archival Science* 2, no. 3 (2002): 263–285.

Edelstein, Jeff, Lola Galla, Carolyn Li-Madeo, Julia Marden, Alison Rhonemus and Noreen Y. Whysel. 'Linked Open Data for Cultural Heritage', 2013.

Eggert, Paul. 'The Archival Impulse and the Editorial Impulse'. *Variants*, no. 14 (20 March 2019): 3–22. Accessed 14 June 2021. https://doi.org/10.4000/variants.570. http://journals.openedition.org/variants/570.

———. 'The reader-oriented scholarly edition'. *Digital Scholarship in the Humanities* 31, no. 4 (December 2016): 797–810. Accessed 6 May 2021. https://doi.org/10.1093/llc/fqw043.

Falco, Riccardo, Aldo Gangemi, Silvio Peroni, David Shotton and Fabio Vitali. 'Modelling OWL ontologies with Graffoo'. In *European Semantic Web Conference*, 320–325. Springer, 2014.

Feliciati, Pierluigi. 'Archives in a Graph : the Records in Contexts Ontology within the framework of standards and practices of Archival Description'. *JLIS*, no. 1 (2021). Accessed 2 May 2023. https://doi.org/10.4403/jlis.it-12675. https://doi.org/10.4403/jlis.it-12675.

Fernández-López, Mariano, María Poveda-Villalón, Mari Carmen Suárez-Figueroa and Asunción Gómez-Pérez. 'Why are ontologies not reused across the same domain?' *Journal of Web Semantics* 57 (2019). https://doi.org/10.1016/j.websem.2018.12.010.

Flanders, Julia, and Fotis Jannidis. *Knowledge organization and data modeling in the humanities*. 2015. https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-111270.

Gabler, Hans Walter. 'Thoughts on Scholarly Editing'. Publisher: Società editrice il Mulino, *Ecdotica* 7, no. 1 (2010): 105–127.

Giovannetti, Francesca, and Francesca Tomasi. 'Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation'. *Digital Humanities Quarterly* 16, no. 2 (2022).

———. 'The linked finding aid as a platform for textual research: The case study of the Giuseppe Raimondi Archive'. In *Proceedings of the Linked Archives International Workshop*, 104–113. 2021.

Greetham, David C. *Textual Scholarship: An Introduction*. Google-Books-ID: RDJ4nTA_3Z8C. Routledge, 1994.

Groth, Paul, Andrew Gibson and Jan Velterop. 'The anatomy of a nanopublication'. *Information Services & Use* 30, no. 1 (2010): 51–56.

Gruber, Thomas R. 'Toward principles for the design of ontologies used for knowledge sharing?' *International Journal of Human-Computer Studies* 43, no. 5 (1 November 1995): 907–928. Accessed 29 March 2023. https://doi.org/10.1006/ijhc.1995.1081. https://www.sciencedirect.com/science/article/pii/S1071581985710816.

Grüninger, Michael, and Mark S. Fox. 'Methodology for the design and evaluation of ontologies'. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*. Montreal. 1995.

Hooland, Seth van, and Ruben Verborgh. *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. London: Facet Publishing, 2014.

Horsman, Peter. 'Taming the Elephant: An Orthodox Approach to the Principle of Provenance'. In *The Principle of Provenance: Report from the First Stockholm Conference on the Archival Principle of Provenance, 2-3 September 1993*, 51–63. Stockholm: Swedish National Archives, 1994.

Hurley, Chris. 'Parallel provenance [Series of parts]: Part 1: What, if anything, is archival description?' An earlier version of this article was presented at the Archives and Collective Memory: Challenges and Issues in a Pluralised Archival Role seminar (2004: Melbourne). *Archives and Manuscripts* 33, no. 1 (May 2005): 110.

InterPARES Trust. *The InterPARES 2 Project Glossary: archival bond*. 2024. Accessed 17 February 2023. http://interparestrust.org/terminology/term/archival%20bond.

Ishida, Richard. *Who uses Unicode?* 17 March 2023. https://www.w3.org/International/questions/qa-who-uses-unicode.

Keet, C. Maria. 'Open World Assumption'. In *Encyclopedia of Systems Biology*, edited by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho and Hiroki Yokota, 1567–1567. New York, NY: Springer New York, 2013. https://doi.org/10.1007/978-1-4419-9863-7_734.

Klein, Michel, Dieter Fensel, Frank Van Harmelen and Ian Horrocks. *The relation between ontologies and XML schemas*, 2001.

Koch, Inês, Nuno Freitas, Cristina Ribeiro, Carla Teixeira Lopes and João Rocha da Silva. 'Knowledge graph implementation of archival descriptions through CIDOC-CRM'. In *International Conference on Theory and Practice of Digital Libraries*, 99–106. Springer, 2019.

Le Boeuf, Patrick. 'Modeling rare and unique documents: using FRBROO/CIDOC CRM'. Publisher: Taylor & Francis, *Journal of Archival Organization* 10, no. 2 (2012): 96–106.

Light, Michelle, and Tom Hyry. 'Colophons and Annotations: New Directions for the Finding Aid'. Publisher: Society of American Archivists, *The American Archivist* 65, no. 2 (2002): 216–230.

Lonsdale, Deryle, David W. Embley, Yihong Ding, Li Xu and Martin Hepp. 'Reusing ontologies and language components for ontology generation'. *Data & Knowledge Engineering* 69, no. 4 (2010): 318–330.

MacNeil, Heather. 'Archivalterity: Rethinking Original Order'. *Archivaria*, 2008, 1–24.

Moreau, Luc, and Paolo Missier. *PROV-DM: The PROV Data Model*. W3C Recommendation 30 April 2013. 2013. Accessed 3 May 2023. https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Murray, Ronald J., and Barbara B. Tillett. 'Cataloging theory in search of graph theory and other ivory towers'. *Information Technology and Libraries* 30, no. 4 (2011): 170–184.

Ngomo, Axel-Cyrille Ngonga, Diego Moussallem and Lorenz Bühmann. 'A Holistic Natural Language Generation Framework for the Semantic Web'. Publisher: arXiv Version Number: 1, 2019. Accessed 26 April 2023. https://doi.org/10.48550/ARXIV.1911.01248.

Noy, Natasha, and Alan Rector. *Defining N-ary Relations on the Semantic Web*. W3C Working Group Note 12 April 2006. 2006. Accessed 3 May 2023. https://www.w3.org/TR/swbp-n-aryRelations/.

Oldman, Dominic, Martin Doerr and Stefan Gradmann. 'Zen and the art of Linked Data: new strategies for a Semantic Web of humanist knowledge'. In *A new companion to digital humanities*, edited by Susan Schreibman, Ray Siemens and John Unsworth, 251–273. Publisher: Wiley Online Library. Wiley, 2015.

Ore, Christian-Emil, and Øyvind Eide. 'TEI and cultural heritage ontologies: Exchange of information?' *Literary and Linguistic Computing* 24, no. 2 (1 June 2009): 161–172. Accessed 28 March 2023. https://doi.org/10.1093/llc/fqp010.

Oury, Clément, Vincent Boulet, Gordon Dunsire, Louise Howlett and Regina Reynolds. 'Definition of PRESSoo: A conceptual model for Bibliographic Information Pertaining to Serials and Other Continuing Resources (Version 1.3)'. Publisher: International Federation of Library Associations and Institutions (IFLA), 2017.

Peroni, Silvio. 'A Simplified Agile Methodology for Ontology Development'. In *OWL: Experiences and Directions – Reasoner Evaluation*, edited by Mauro Dragoni, María Poveda-Villalón and Ernesto Jimenez-Ruiz, 55–69. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017.

Peroni, Silvio, Aldo Gangemi and Fabio Vitali. 'Dealing with markup semantics'. In *Proceedings of the 7th International Conference on Semantic Systems*, 111–118. 2011.

Peroni, Silvio, and David Shotton. 'The SPAR Ontologies'. In *The Semantic Web – ISWC 2018*, edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee and Elena Simperl, 119–136. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018.

Peroni, Silvio, David Shotton and Fabio Vitali. 'Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents'. In *Proceedings of the 8th International Conference on Semantic Systems*, 9–16. ACM, 2012.

Peroni, Silvio, Francesca Tomasi, Fabio Vitali and Jacopo Zingoni. 'Semantic Lenses as Exploration Method for Scholarly Articles'. In *Bridging Between Cultural Heritage Institutions*, edited by Tiziana Catarci, Nicola Ferro and Antonella Poggi, 385:118–129. Series Title: Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. Accessed 31 May 2023. https://doi.org/10.1007/978-3-642-54347-0_13.

Pinto, H. Sofia. 'Towards Ontology Reuse'. *Proceedings of AAAI 99's Workshop on Ontology Management* 13 (1999): 67–73.

Posner, Ernst. 'Some Aspects of Archival Development since the French Revolution'. Society of American Archivists, *The American Archivist* 3, no. 3 (1940): 159–172.

Price, Kenneth M. 'Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?' *Digital Humanities Quarterly* 3, no. 3 (2009): 11.

Raboni, Giulia, and Paola Italia. *Che cos' è la filologia d'autore*. Publisher: Carocci. Roma: Carocci, 2010.

Raimondi, Giuseppe. 'Album De Pisis'. Album, BFICLIT FR.A ALBUM DEPISI, RDm 835, 1916.

———. 'Così nacque La Raccolta'. *Il Resto del Carlino*, 20 August 1970.

———. 'Divagazioni intorno a Paul Valéry'. *Il Convegno* 6, no. 2 (1925): 90–97.

———. *I tetti sulla città*. Milano: Mondadori, 1977.

———. *Il cartesiano signor Teste*. Edizioni di Solaria. Firenze, 1928.

———. *La valigia delle Indie*. Firenze: Vallecchi, 1955.

———. ‘Le nevi di Villon’. Articolo ms. BFICLIT FR.A ARTICOLI1 1972 12.22, RDa 349, 22 December 1972.

———. ‘Le nevi di Villon’. *Il Resto del Carlino*, 2 January 1973.

———. ‘Le nevi di Villon’. *La Nazione*, 4 January 1973.

———. ‘Nelle ballate di François Villon. Gelo invernale e nostalgia di legna accesa’. *Il Giorno*, 7 June 1978.

———. ‘Ringraziamento a Commerce’. *Il Convegno* 6, no. 9 (1925): 487–491.

———. ‘Una forca per il poeta François Villon’. Quaderno ms. BFICLIT FR.A QUADERNI.1 1976 03, RDq 303, 1976.

Raimondi, Giuseppe, and Clemente Mazzotta. *Giuseppe Raimondi fra poeti e pittori: mostra di carteggi*. Edizioni Alfa, 1977.

Rasmussen, Krista Stinne Greve. ‘Reading or Using a Digital Edition? Reader Roles in Scholarly Editions’. In *Digital Scholarly Editing : Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 119–133. Digital Humanities Series. Code: Digital Scholarly Editing : Theories and Practices. Cambridge: Open Book Publishers, 13 July 2017. Accessed 26 May 2023. http://books.openedition.org/obp/3406.

‘RDF 1.1 Primer’. Accessed 27 April 2023. https://dvcs.w3.org/hg/rdf/raw-file/tip/rdf-primer/Overview.html.

Records in Contexts Conceptual Model. Consultation Draft v0.2. International Council on Archives Expert Group on Archival Description, 2021.

Riva, Pat, Patrick Le Boeuf and Maya Žumer. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. 2017.

Riva, Pat, and Maja Žumer. ‘FRBRoo, the IFLA Library Reference Model, and now LRMoo: a circle of development’, 2018, 7. https://library.ifla.org/id/eprint/2130/1/074-riva-en.pdf.

Robinson, Peter. ‘Why digital humanists should get out of textual scholarship’. Social, Digital, Scholarly Editing Conference. University of Saskatchewan, 11 July 2013.

Rossi, Federica, and Alina Wenzlawski. 'Nello scrittoio di Giuseppe Raimondi: carte e libri di un letterato bolognese su Paul Valéry'. In *Il privilegio della parola scritta: gestione, conservazione e valorizzazione di carte e libri di persona*, edited by Giovanni Di Domenico and Fiammetta Sabba, 177–194. Roma: Associazione italiana biblioteche, 2020.

Sabba, Fiammetta. 'Biblioteche e carte d'autore: tra questioni cruciali e modelli di studio e gestione'. *AIB studi* 56, no. 3 (2016).

Sanderson, Robert, Paolo Ciccarese and Benjamin Young. *Web annotation data model. W3C Recommendation*. 2017-02-23. Http://web.archive.org/save/https://www.w3.org/TR/annotation-model/. 2017.

Sauermann, Leo, and Richard Cyganiak. *Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008*. 3 December 2008. Accessed 9 January 2020. https://www.w3.org/TR/cooluris/.

Schreibman, Susan, Ray Siemens and John Unsworth, eds. *A companion to digital humanities*. John Wiley & Sons, 2008. https://companions.digitalhumanities.org/DH.

———, eds. *A new companion to digital humanities*. John Wiley & Sons, 2015.

Shillingsburg, Peter L. *From Gutenberg to Google: Electronic representations of literary texts*. Cambridge University Press, 2006.

Simperl, Elena. 'Guidelines for reusing ontologies on the semantic web'. Publisher: World Scientific, *International Journal of Semantic Computing* 4, no. 2 (2010): 239–283.

Sowinski, Piotr, Katarzyna Wasielewska-Michniewska, Maria Ganzha, Marcin Paprzycki and Costin Badica. 'Ontology Reuse: the Real Test of Ontological Design'. 14 September 2022. https://doi.org/10.3233/FAIA220292.

Sperberg-McQueen, C., David Dubin, Claus Huitfeldt and Allen Renear. 'Drawing inferences on the basis of markup', 1 January 2002.

Stadler, Peter, Marcel Illetschko and Sabine Seifert. 'Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <corresp Desc>'. Number: Issue 9 Publisher: Text Encoding Initiative Consortium, *Journal of the Text Encoding Initiative*, Issue 9 2016. Accessed 24 June 2021. https://doi.org/10.4000/jtei.1433. http://journals.openedition.org/jtei/1433.

TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.5.0. Https://web.archive.org/save/https://tei-c.org/guidelines/. TEI Consortium, 2019. http://www.tei-c.org/Guidelines/P5/.

Thomassen, Theo. 'A first introduction to archival science'. Publisher: Springer, *Archival science* 1, no. 4 (2001): 373–385.

Toccafondi, Diana. 'Gli archivi letterari del Novecento: un laboratorio per la collaborazione tra professionisti'. In *Conservare il Novecento: gli archivi culturali. Ferrara, Salone internazionale dell'arte del restauro, 27 marzo 2009: atti del convegno*, 39–46. Associazione italiana biblioteche, 2010. Accessed 27 January 2022. https://www.torrossa.com/it/resources/an/5088757.

Tomasi, Francesca. 'Archivi di persona in linked open data: il modello concettuale come strumento di integrazione nei GLAMs'. *AIB studi* 57, no. 2 (20 November 2017). Accessed 13 September 2019. https://doi.org/10.2426/aibstudi-11647.

———. 'L'edizione delle carte degli scrittori: l'archivio letterario incontra il (nativo) digitale'. *Biblioteche oggi Trends* 7, no. 1 (2021): 81–91.

Tomasi, Francesca, and Marilena Daquino. 'The archival domain in a disciplinary-integrated ontological perspective'. Number: 3, *JLIS.it* 6, no. 3 (15 September 2015): 13–38. Accessed 1 May 2023. https://doi.org/10.4403/jlis.it-11133.

Valéry, Paul. 'Préface pour une nouvelle traduction de La Soiréè avec M. Teste'. *Commerce* 4 (1925): 93–102.

Van Hulle, Dirk, and Mark Nixon. *Samuel Beckett Digital Manuscript Project: Editorial Principles and Practice*. December 2021. Accessed 24 June 2021. https://www.beckettarchive.org/editorial.jsp.

W3C. 'RDF Syntax', 31 January 2023. Accessed 24 April 2023. https://www.w3.org/wiki/RdfSyntax.

Yakel, Elizabeth. 'Archival representation'. *Archival Science* 3, no. 1 (2003): 1–25.

Yeo, Geoffrey. 'Bringing Things Together: Aggregate Records in a Digital Age'. *Archivaria*, 2012, 43–91.

———. 'Contexts, Original Orders, and Item-Level Orientation: Responding Creatively to Users' Needs and Technological Change'. *Journal of Archival Organization* 12, no. 3 (2015): 170–185.

———. 'The Conceptual Fonds and the Physical Collection'. *Archivaria* 73 (2012): 43–80.

Acknowledgements

I would like to thank Prof. Francesca Tomasi for her supervision, mentorship, and friendship, as well as for always being supportive and encouraging during and before my PhD journey. I consider myself fortunate to have met her and, through her, the Digital Humanities. Our years of study together have left an indelible mark on my academic and personal life, and I will be forever indebted to her for her profound influence on my growth as a scholar and as an individual.
I would also like to thank all of the researchers at the FICLIT Department, whose collaboration and insights have enriched my research experience. Their knowledge and contributions have been influential in the development of my work. I am also grateful to my PhD colleagues for creating a nurturing and intellectually stimulating environment. Their friendship and shared experiences have served as a source of inspiration and motivation.
Finally, I would like to thank all of my family, friends, and colleagues for their encouragement and support throughout this PhD journey.

Francesca Giovannetti, Bologna, June 2023