Alma Mater Studiorum · Università di Bologna

DOTTORATO DI RICERCA IN
Culture Letterarie e Filologiche
Ciclo XXXI

Settore Concorsuale: 11/A4 (primario); 01/B1 (secondario)
Settore Scientifico Disciplinare: M-STO/08

Mining Authoritativeness in Art Historical Photo Archives. Semantic Web Applications for Connoisseurship

Presentata da:
Marilena Daquino
Coordinatore Dottorato:
Prof. Luciano Formisano

Supervisore:
Prof.ssa Francesca Tomasi

Esame finale anno 2019

Abstract

The purpose of this work is threefold: (i) to facilitate knowledge discovery in art historical photo archives, (ii) to support users' decision-making process when evaluating contradictory artwork attributions, and (iii) to provide policies for information quality improvement in art historical photo archives. The approach is to leverage Semantic Web technologies in order to aggregate, assess, and recommend the most documented authorship attributions. In particular, findings of this work offer art historians an aid for retrieving relevant sources, assessing *textual authoritativeness* (i.e. internal grounds) of sources of attribution, and evaluating *cognitive authoritativeness* of cited scholars. At the same time, the retrieval process allows art historical data providers to define a low-cost data integration process to update and enrich their collection data. The contributions of this thesis are the following: (1) a methodology for representing questionable information by means of ontologies; (2) a conceptual framework of Information Quality measures addressing dimensions of textual and cognitive authoritativeness characterising art historical data, (3) a number of policies for metadata quality improvement in art historical photo archives as derived from the application of the framework, (4) a ranking model leveraging the conceptual framework, (5) a semantic crawler, called mAuth, that harvests authorship attributions in the Web of Data, and (6) an API and a Web Application to serve information to applications and final users for consuming data. Despite findings are limited to a restricted number of photo archives and datasets, the research impacts on a broader number of stakeholders, such as archives, museums, and libraries, which can reuse the conceptual framework for assessing questionable information, *mutatis mutandi*, to other near fields in the Humanities.

Keywords: Knowledge Discovery, Linked Data, Art History, Photography, Authoritativeness, Information Quality

Acknowledgements

This work comes at the end of hard times. Personal events and work experiences affected significantly the result - sometimes in unpleasant ways, but mostly in unexpected, joyful, and exciting ways!

First, I would like to express my gratitude to my supervisor **Francesca Tomasi** for the continuous support of my Ph.D study and related research. She supported me when I doubted of my skills, she made me understand what I am able to do, and to feel comfortable with my limits. She gave me all the strength I needed to focus on my objectives and free my creativity. She was with me when I was ill, when I felt alone, and she made me feel at home when I felt I had no home to go back. For such reasons, I want to thank her for being a friend, not an advisor only.

Secondly, I would like to thank **Francesca Mambelli** from the Federico Zeri Foundation. I cannot imagine this thesis without her doubts and insightful questions. She helped me to understand what an incredible work is under the hood of cultural heritage management and how fascinating it could be.

In the last few years I met an incredible amount of brilliant people thanks to my job. By attending events organised by the Semantic Web community and by discussing with PHAROS members I got precious feedback and I learnt, almost from scratch, all I needed to write this thesis. I want to thank all of them for supporting me in the evaluation of this work and for contributing to the knowledge I gained, which is priceless. Among them, a special thanks goes to my two reviewers, **Enrico Daga (KMI)** and **Costanza Caraffa (KHI)**, whose comments helped me to frame questions related to so different communities - which are (hopefully) highlighted in this work. In particular, thanks to Enrico for his patience when explaining me how *not* to write a PhD thesis and how not to be concerned (too much) of it.

Workshops, conferences, and summer schools have been fantastic places for networking. At SSSW2016 I met few of the people that keep positively affecting my career and that are now some of my best friends. There I met for the first time the **KMI crew**, who managed to make me move (twice) to the exotic Milton Keynes. A special thanks goes to Mathieu d'A., Alessandro Adamou, Enrico Daga (again), and Enrico Motta, for supporting and guiding me during my visiting period. The MK experience was unique, not only in working terms. The impressive number of good people that concentrate in such a *small* place is unspeakable. I cannot list all of them, but I can mention the ones that still walk with me despite the long distance: Ilaria, my beloved grumpy friend, that never makes me feel alone; Pinelopi, with whom I shared fears, ideas and lots of hugs during our cigarette breaks; Martin, who took care of me as only a brother would do; Tina, for pushing me all the time to challenge myself and make the best of situations; Angelo

and Andrea, for their friendship and kindness when helping me coding; Manu and Giorgio for their good advices, and the silly moments together (that are hard to distinguish sometimes); and (in random order) Carla, Davide, Paco, Anna, Anita, Francesco, Thivian, Vasa, Patrizia, Jaisha, Julian, Matteo, and all the others, for all the fun we had together.

Another group of people influenced my career, that is, the **AlmaDL** division at the University of Bologna, where I spent three years as research assistant. I want to thank Raffaele, Marialaura, Roberta, Piero, Jennifer, Fabrizio, and Antonio, with whom I moved my first steps in research. And of course I thank professors, colleagues and the admin team at the **Department of Classic Philology and Italian Studies** that continuously help me, including: Federico Condello, Paola Italia, Aldo Gangemi, Ivan Heibi, Alessandra Di Tella, Beatrice Nava, Morena, Laura, Virna, Valeria, and Antonella; Valentina Presutti, Mehewish Alam, Valentina Carriero, Andrea Nuzzolese (STLab - CNR); Fabio Vitali, and Angelo Di Iorio (DISI). Lastly, thanks to Silvio Peroni, my *panino-buddy* and a good advisor since I started this adventure.

I have been living in **Bologna** for almost fourteen years. Plenty of people crossed my path. Some moved away, some stayed for longer, and some keep bringing joy in my life every day. Among them are Giulia, my beautiful dreamer; Martina, as adventurous as cloudy; Maria Rosaria, the tiny crazy traveller; Andrea, the brilliant grumpy friend (who will raise his eyebrows when reading it); Daniel, the worst proofreader ever, but a thoughtful friend; Matteo and his family (Nadia, Vanni, Fazia, and Danilo), who were like a second family to me; Raffaele (again), for the litres of beer we poured while chit-chatting in all the pubs and restaurants of Bologna; Roberta, possibly the craziest flatmate I have ever had, with the biggest heart I could expect.

And of course, I must mention my hometown. **Domodossola** became famous thanks to *The Fortune wheel* TV program, where it was first used to spell words that start with D. Now everybody ask me "Does it exist for real?". Spoiler alert: it does, it is beautiful, and amazing people live there. My *pieces of heart* Simone and Federico; Gabriele (Teddy), smart and sensitive; Veronika, a friend and the best sister-in-law I could ask for. My siblings Miriam and Marco, my biggest treasure. My mother, who deserves all the best in this world. My cousins Lorella, Beppe, Alessia, my uncles and aunts, with whom I shared the best moments. And my beautiful niece Marika, who makes us smile again.

Lastly, I want to remember my grandparents, that keep looking down on us, and my father, who taught me to love science, photography, arts, and to be curious. I owe him everything. This work is dedicated to him.

To whom will not read this thesis,
but that gave me all the love I needed
to make it happen.

Introduction

Art historical photo archives have always been places for art historical research. Scholars, photographers, art dealers, antiquarians, and others, used to create, arrange, and study photographic collections depicting artworks in order to pursue their activities. Plenty of *connoisseurs*, i.e. experts that ascribe artworks to artists on the basis of their knowledge in the field of fine arts, used to attend photo archives and to record on the back of photographs their hypotheses. The latter include expertises on the authorship of artworks, provenance information of artworks, and bibliographic references.

The wealth preserved in photo archives bears witness to the diversity of scientific methods that characterised connoisseurship in the last two centuries. According to Carlo Ginzburg [Ginzburg, 1979] the epistemological model in the Humanities that emerged in the nineteenth century stems from connoisseurship. To this extent, characteristics of connoisseurs' methodologies are of great interest to other fields in the Humanities. However, most of connoisseurs' methodologies are not reproducible, and authoritativeness of scholars and sources is a key element when validating attributions.

This work is mainly devoted to understand how connoisseurs' methodologies sedimented in photo archives, to what extent we can formalize a definition of authoritativeness on the basis of second-hand knowledge providers, and to what extent such a definition can be leveraged in information technologies. In particular, a formal definition of authoritativeness would contribute to advancements of two research areas, namely: Connoisseurship, by developing more reliable recommending systems for scholars, and Library and Information Science, by defining new strategies for metadata quality improvement based on the aforementioned recommending systems.

In the last two decades, the development of information technologies has been affecting the way cultural heritage institutions provide services to their users. In fact, the role of photo archives has changed significantly. Image-based tools and online catalogues are scholars' main instruments for performing their studies remotely, and are the way data providers encourage art historians to keep relying on photographic collections for accomplishing their tasks. However, digital tools are not able to fulfill all of the sophisticated needs of scholars and archivists.

On one side, scholars have to gather significative amounts of sources in order to validate the veracity of their assumptions. While the aggregation of sources can be achieved by means of automatic methods, e.g. online aggregators, the evaluation of retrieved sources is still demanded to users. The latter have to collect relevant sources, select them on the basis of both provider's authoritativeness (cognitive authoritativeness)

and internal grounds of sources (textual authoritativeness), and draw conclusions on the basis of available data - whether this are sufficient or not for the task at hand. The task may be particularly challenging when information at hand is questionable, e.g. when contradictory authorship attributions are available, when sources are scarce, not well-documented, or not updated.

We argue that curated and automatic methods can effectively support users in their decision-making process when evaluating textual authoritativeness of sources. Tools tailored on domain-dependent features are fundamental to supply, say, the absence of archivists and domain experts, that used to advise users in art historical photo archives. In particular, this work aims at leveraging bespoke technologies for aggregating art historical data related to connoisseurship activities, identify features that characterise authoritative sources of information, and support scholars in evaluating their authoritativeness.

On the other side, to convey the necessary richness of information in a digital format is challenging. Cultural institutions are unanimously deemed high-quality metadata providers. However, the cataloguing process is a time-consuming activity and many factors can hinder resulting information quality, e.g. lack of time, human resources, and sources of information. Moreover, metadata standards, vocabularies, and ontologies are still debated in the cultural heritage domain, and the description of the same cultural object may significantly differ among data sources, affecting data integration processes. Secondly, updating information over time is an expensive task. Archival policies for information quality are not shared among providers, and strategies for metadata quality improvement are not available.

To this extent, this work has two objectives: first, to address a common kernel of descriptive elements to be shared among data providers when describing the heritage of art historical photo archives, with a specific focus on questionable information (attributions), so as to facilitate data integration. Secondly, we aim at providing effective means and policies for improving art historical data quality in a low-cost data integration process.

Semantic Web technologies and Linked Open Data are currently recognised as the *lingua franca* for sharing and integrating heterogeneous data sources, and are widely adopted by cultural heritage institutions to ensure a better experience to final users. The aim is to demonstrate that such technologies are (i) suitable for the development of tools and methods tailored on cataloguers' needs, and (ii) can effectively support scholars' and archivists' daily tasks.

Three specific research problems (RP) are tackled in this thesis, which can be summarised as follows:

- RP1. The formal representation of questionable information in the Photography and Arts domains

by leveraging well-grounded formal languages and technologies.

- RP2. The formalisation of the dimensions characterising the methodology of art historical data providers when publishing questionable information.
- RP3. Support users' decision-making process when assessing reliability of authorship attributions.

Each research problem corresponds to an objective we aim to achieve. In particular, this thesis has two research objectives (RO) and one technological objective (TO), namely:

- RO1. Define ontologies for representing the Photography and Arts domain, with a particular focus on questionable information.
- RO2. Define methods to assess the methodology undertaken by art historical photo archives when providing questionable information and the authoritativeness of the latter.
- TO3. Develop a system that implements the conceptual framework and supports the decision-making process of users.

The main contributions of this work and related chapters addressing their description are listed below:

- The analysis of features characterizing the Photography and Arts domain and connoisseurship activities, and the survey of cataloguing standards (Chapter I).
- The analysis of available ontologies and projects addressing the Photography and Arts domain in Cultural Heritage (Chapter II).
- The review of Information Quality dimensions that apply to art historical data (Chapter II).
- The transformation of the Zeri photo archive into a Linked Open Dataset so as to create a golden standard for representing questionable information in a machine-readable format (Chapter IV).
- The HiCO Ontology for representing questionable information and the interpretative process (Chapter V).
- The FEntry Ontology and OAEntry Ontology for representing the Photography and Arts domain respectively, derived from the mapping of the Italian cataloguing rules ICCD-OA and ICCD-F (Chapter V).

- A conceptual framework of Information Quality measures for defining textual authoritativeness of sources, derived from a comparative analysis of archival standards and catalogue data (Chapter VI).
- A set of dimensions identifying art historians' cognitive authority (Chapter VI).
- A ranking model for assessing textual authoritativeness of authorship attributions (Chapter VI).
- Policies for improving art historical data quality in a low-cost integration process (Chapter VI).
- A semantic crawler, called mAuth, for harvesting authorship attributions in a (extensible) number of data sources (Chapter VII).
- An API for integrating harvested data in online catalogues (Chapter VII).
- A web application for evaluating the conceptual framework and the ranking model (Chapter VII).

The thesis is structured in two parts.

The first part "Art Historical Photo Archives in the Age of Semantic Web" is dedicated to the background of this work, namely: characteristics of documentation in art historical photo archives, features of connoisseurship, basics of Semantic Web technologies, Knowledge Organization and Information Quality aspects. Chapter I provides the theoretical background on art historical photo archives, discussing contributions in Archival Science and Library and Information Science on the photograph of artworks, surveying cataloguing standards, and introducing peculiarities of connoisseurship activities. Chapter II provides an overview of the technologies leveraged in this study, a survey of existing ontologies and modelling approaches in the Cultural Heritage domain. Lastly, an overview of the dimensions commonly used for assessing data quality that apply to art historical photo archives are presented.

The second part "Semantic Web Applications for Connoisseurship" describes the research project, its outcomes, and the evaluation of results. In Chapter III are outlined research problems, hypotheses and assumptions, the methodology adopted to validate hypotheses, and the approach to the research. Chapter IV is dedicated to the use case that guided the development of the project, i.e. the Federico Zeri photo archive of the University of Bologna. Chapter V is dedicated to the description of the ontologies developed for representing questionable information in the Photography and Arts domain in the Cultural heritage. Chapter VI illustrates the data analysis performed on photo archives collection data for assessing their methodologies, so as to define a set of Information Quality dimensions, a ranking model, and strategies for metadata quality improvement. Chapter VII describes the artefact developed as a proof-of-concept of

the aforementioned conceptual framework. Chapter VIII describes the evaluation of the ontologies and the user-centered evaluation of the conceptual framework. Finally, in Chapter IX contributions, limitations and impact of research are summarised, and future works are addressed.

The writing style differs significantly between the two parts. The first part is a gentle introduction to problems related to the interdisciplinary scenario and presents both theoretical and technical aspects. The narrative starts from broader topics and narrows to the scope of specific research problems. In this case the writing style is closer to a humanistic vision. The second part presents the work done to tackle problems and achieve goals illustrated in Chapter III. Every chapter discusses a part of the work and the writing style is here technical.

Findings of this work, such as ontologies, datasets, data analysis results, user-study results, and the web application, are stored for the long-term preservation in bespoke repositories, are uniquely identified by means of DOIs, and are available online. All the URLs have been accessed in February 15, 2019.

Contents

List of Tables

List of Figures

Part I

ART HISTORICAL PHOTO ARCHIVES IN THE AGE OF SEMANTIC WEB

Chapter 1

Photography and Art Historical Research in Photo Archives

This chapter provides a gentle introduction to theoretical aspects related to art historical photo archives and connoisseurship activities. In detail, (1) features of the photograph of artworks and resources that are part of the ecosystem of art historical research are described, (2) cataloguing standards are surveyed, and (3) connoisseurship methodological issues are presented. Aspects here mentioned are not meant to provide a comprehensive account of photography in a diachronic perspective, nor an exhaustive history of its development and applications. The aim is to present a motivating scenario and the challenges arising from a knowledge representation standpoint.

1.1 An introduction to photography in the Cultural Heritage Domain

> [...] Photography was part of a systemic shift that had profound cultural repercussions. It revolutionized memory, changed the relationship between past and present, produced "a massive reorganization of knowledge and social practices", and occasioned a "major readjustment of the alphabet/image ratio in ordinary communication". [Schwartz, 1995]

Photography has been popular for research since its inception, due to the many applications and possibilities it offers to scholars and professionals in supporting their activities. Presently, photographs saturate museums, archives and libraries, and are integral to their management procedures [Lien and Edwards, 2014]. However, originality of photographs has always been deemed questionable, and photographic collections have been considered a problematic subject that challenged cultural institutions and curatorial practices. Writing the history of a photographic collection requires an incredible effort in terms of analysis, since it considered "historically vulnerable to the redrawing of collections boundaries and curatorial territories" [Edwards and Morton, 2015]. The methodology for tracing the history of collections of cultural objects is based on *social biography*, i.e. all the movements of cultural objects are tracked and plotted on a timeline [Gosden et al., 2007]. Being photographs serial objects that can appear in several forms during their life cycle, the model of social biography is only partially applicable, and new research methodologies had to be developed so as to expand the scope of traditional inquiries and address issues peculiar of photo collections.

In the 1970s the interest in alternative historical narratives and new forms of representation drew the attention of museums and academic disciplines back to photographic collections. This expanding interest evolved over time and is now a relevant topic in the literature of museum collecting, such as: the history of the photo library of the Council for Industrial Design [Moriarty, 2000]; the implications of the historical arrangement of the photo archive at the Musée de l'homme in Paris [Barthe, 2000]; the Smithsonian's Schindler Catalogue of Native American photographs [Gidley, 2005]. Theoretical contributions also appear in essays on the nature of photographic collections [Edwards and Hart, 2004]; the formation of photographic collections [Franceschini et al., 2014], in archives [Belovari, 2013] and in art historical photo archives [Caraffa, 2011].

The nature of the photograph is further explored by archivists. According to Tim Schlak's review of archival scholarship on the photograph, "the literature from the 1970s and 1980s is replete with contradiction over the photograph's status as information, documentation, authentication and representation. Yet, these contradictions indicate varying and often competing notions of the photograph as truth and representation" [Schlak, 2008]. Photography was ambiguously perceived as an imitation of reality, and only later in time new questions came up on "who and what are being represented, and by whom, for what purposes (conscious or unconscious), and with what effect on which viewers" [Jussim, 1989]. In the way of a written document, the photograph was understood to be a valuable witness and source of documentation. This

modernist view of the photograph, akin to that of a written document, meant that cataloguing models and descriptions used for one could be adapted for the other. To this extent, the photograph became an evidence that could be described and stored, although the focus remained the "factual content [of the photograph] rather than the functional origins of visual images" [Schwartz, 2002].

Joan M. Schwartz's work is considered the intellectual foundation of postmodern archival writing on the photograph. She introduced into archival scholarship the key element that characterises the object photograph, i.e. the functional context of the document creation.

> By studying the photograph, not as a more or less accurate transcription of the material, but in terms of its relationships with the persons concurring in its formation, diplomatic principles and concepts may help to break the presumed link between the photographic image and visual 'truth' by revealing the photograph to be a mediated representation of reality: the product of a series of decisions; created by a will, for a purpose, to convey a message to an audience. [Schwartz, 1995]

Literature from the 1990s onwards "represents, however sporadic, a substantial attempt to give the photograph its archival due by stripping it of a reductionist insistence on an empiricist notion of truth" [Schlak, 2008].

In the following section we address the role of photographs in art historical research. We detail aspects related to the nature of the photograph of artworks and discuss its importance in art historical photo archives, wherein the functional context of photo collections is further investigated.

1.2 The photograph of artworks in art historical photo archives

Both photography and art history as academic disciplines first developed in the second half of the nineteenth century. They developed in parallel, but as art history became accepted as an academic discipline, photography continued to be seen as instrumental. In fact, photographic reproductions of artworks are the main documentary sources in art historical research, pillars of the methodology of art history. Although

they are privileged tools, their significant historical value took time to be appreciated by the historiography of art history. As aforementioned, neutrality and objectivity of photograph has been arisen several times in the literature. A photograph can be manipulated and interpreted differently according to the observer's cultural background and purposes. In this sense, photography is both "a visual language and a cluster of expectations and ideas" [Marien, 2006]. The "epistemological potential" of photographic collections [Caraffa, 2011] emerged when the introduction of archival methods shaped them into living entities, knowledge carriers, rather than impartial sources of information. In other terms, when photographs ceased to be a daily support for historians and sedimented into archives for their long-term preservation, archival science shed light on their historical value. The consequence of such an awareness is the widespread acknowledgement of the photograph as an *historicized thing*, that is, independent from the subject it represents and which worths a detailed analysis and description. The standardisation of cataloguing rules and the introduction of digital photography shed light once again on such topics, and demonstrated that photo collections have a key role in conveying non-neutral knowledge.

The photograph of artworks. As aforementioned, documentary photographs are the main tools of art historians. Historians can conduct comparative studies on different views of the same artwork, with different lights or perspectives, without moving from their desk. Moreover, photographs of artworks that are preserved in different places, can be collated and compared. Costanza Caraffa [Caraffa, 2011] summarises the ways photographs are useful as follows:

- As *aides-mémoires*, to remind details of artworks seen for real at least once.
- To compensate the lack of direct knowledge of works reproduced.
- To document artworks that do not exist anymore, or are inaccessible.
- To document changing conditions of the artwork in time.
- To support teaching with visual materials.
- To provide references (e.g. accession numbers) for the communication in the art-historical debate.

Such applications of photography in art historical photo archives strengthen the idea, embraced by positivism and by early archival scholarship, that the mechanical process underlying the creation of a pho-

tograph ensures objectivity and neutrality. However, photographs reflect the cultural and technological conditions of the time in which they are taken. Details can be hidden by photographer's techniques, or easily manipulated, and different observers can draw different conclusions. Therefore neutrality is not an intrinsic feature of the photograph, while the authoritativeness of the institution preserving the photograph guarantees the veracity and the quality of the document [Caraffa, 2011].

Photos bear more information than the ones related to their subject, which has to be properly processed, interpreted and returned into cataloguing records. Such an awareness is the result of a long debate between archives, museums, and libraries, on the models and standards to be used in their information systems. Photo archives questioned the use of bibliographic models - which include only the description to the features of the carrier and mainly focus on its subject - in favour of a hierarchical model. The archival model groups photos in series and fonds, according to their "instrumental participation in a function or process created [...] by a will, for a purpose, to convey a message, to an audience" [Schwartz, 2002].

Archives encouraged the shift from the *indexicality* of photographs (the reduction of the photograph to its visual content) to their *evidential* value (being part of a broader context that can not be de-contextualized). As such, the photograph becomes a monument itself, to be curated as an archival object provided of a functional context, a provenance, and a history of production and sedimentation [Mambelli, 2018]. The recognition of the functional context fostered the idea of photograph as historical source *tout-court*. Visual clues may be collated and compared with texts and with other artifacts to present a more complete picture on a topic, but it is not to be used for illustrative and supplementary purposes, ancillar of the traditional, textual ones. It is a primary evidence that can be used to support the historical inquiry.

To this extent, photo archives are other than libraries wherein images can be accessed by subject. They provide "a constellation of other data [...] that are, whether intentionally or not, registered in them" [Caraffa, 2011].

The features of the photograph of artworks. Several elements characterize photographic reproductions of artworks and must be addressed in archival descriptions in order to highlight their epistemological value. Some features of the photograph have already been introduced, e.g. originality, seriality, functional context. In this section we outline, in a glossary fashion, the dimensions that describe the multi-faceted nature of photographs of artworks. In particular, we include all the aspects that are deemed fundamental

by connoisseurs for pursuing their activities.

The following schema is based on, reviews, and extends Joan Schwartz's work [Schwartz, 1995]. She applied the diplomatic methodology to the appraisal of photographs, comparing dimensions characterizing a textual document to the ones deemed relevant for the description of photographs. The aim is to provide an overview of aspects related to the nature of photographs that may affect decisions and validity of claims made by art historians.

Visual authority. Authority can be defined as the demonstrated "truthfulness of facts" [Duranti, 1998] represented or conveyed by a medium. The visual authority of a photograph is derived from the verisimilitude, i.e. the realism and the accuracy of the depicted content. Nonetheless, visual authority can be undermined by photo editing technologies. By changing proportions, colors, light, and shade, important details can be hidden, and the observer's attention can be drawn to others. Thus, the message conveyed by the photograph does not equate with the content, and can lead to wrong assumptions when analysing the visual facts. For this reason, black and white photos have often been preferred by art historians for attributing the authorship of the depicted artwork. Federico Zeri, one of the most relevant art historians of the last century, claimed:

> Preciso che le fotografie debbono essere in bianco e nero: anche se puó sembrare un paradosso, non riesco a leggere correttamente le fotografie a colori dove ogni dato é affogato in una sorta di minestrone[.][1] [Zeri, 1995]

Relying on authoritative sources is a pillar for both art historians and photo archivists when supporting authorship attributions. Despite recording the physical description of photographs has often been seen as «an irrelevant exercise to "document the medium" [Schwartz, 1995], such an information is fundamental for evaluating the value of standpoints derived from the appraisal of photographs.

Validity of the photograph as a lens on the past. The authority of the photographic document is confirmed by the reputation and reliability of actors concurring to the creation of the document. Authority of photographs concurs to evaluate the validity of a photographic reproduction of an artwork. Validity can be

[1] "I specify that the photographs must be in black and white: even though it may seem a paradox, I can not correctly appraise color photographs where every data is drowned in a sort of minestrone" (author's translation)

defined as the extent to which physical form and features conform to requirements of a commissioning agent. For example, the photographer's stamp on the back of the photo, the address, and the date of the shot, are all elements that convey the compliance to quality standards.

Photographs identified in photographers' catalogues or part of photographic campaigns provide more context information to observers. For instance, campaigns may highlight art market interests in a specific artist or school. Moreover, parameters used to evaluate the validity of a photo shade light on how art history developed its research lines (providing evidences to the historiography of art history), and how art market developed business opportunities over time.

In art historical photo archives captions and annotations on the recto/verso of photographs are generally transcribed and provided to users so as to explore the archive as a lens on the past as addressed by the collections, and not only on the basis of its subjects.

Originality: unicity and seriality. The concept of original document has already been pointed out as one of the main challenges in the Photography and Arts domain. The reproducibility of images struggles curators in defining the unicity of the object.

A compromise is found by addressing the negative as the "truest record of the information captured by the camera" [Leary, 1985]. The negative number (i.e. the identifier of the negative included in a collection) recorded in cataloguing records is the primary source for tracing the trajectory of the manifestations of the photographic object.

Connoisseurs, dealers and archives exchanged photographs to carry out their own activities. Copies derived from the same negative were sold and preserved in several archives. To this extent the negative is the thread for reconstructing relations between archives, scholars, and trace the evolution of the interest in particular artists over time.

However, the negative is not sufficient to convey the context of the photograph. The ways the negative is printed and exposed provide information that the "truest" carrier could not store. For instance, photographs are often mounted on cards, and annotations are taken by photographers, scholars, and archivists. Pieces of information that are recorded by several actors, in different moments, and for diverse purposes and audiences, demonstrate that the context, rather than the lonely content, defines the photograph as a

part of a whole, i.e. the photographic collection, that is endowed of epistemological potential.

People involved in the life cycle of the photograph. Strictly related to the seriality of the photographic object is "the complexity of creative forces behind the photograph" [Schwartz, 1995]. Photographs are important because of the information the y provide about people. As stated by Luciana Duranti, "We identify, acquire, select, describe, communicate, and consult documents largely in relation to the persons they come from, are written by, directed to, concerned with or have effect on" [Duranti, 1998]. This idea, stated by a diplomatist but widely shared in the archival domain too, consists in evaluating the context of the cultural object as important as the object itself. By means of the photographic object several research scenarios can be explored in both the history of photography and art history, disclosing a dense network of human exchanges, namely:

- The creation of the photograph involves people in the conceptualization of the image (commissioners and photographers) and in the actual realization of the image (cameramen, scenographers). The history of photography is made by people, their relations, the situations that they create thanks to their innovative ideas and the way they communicate ideas.

- Photographic catalogues record identifiers of photographs original negatives. Printers provide high quality prints derived from the negative. Printing techniques adopted shade light on the technological development in the history of photography.

- The reproduction of the photograph to the wide public include people related to the publishing domain (publishers, distributors), and exhibitions (curators). By means of pictures disseminated through different media, scholars can reconstruct the how interests in specific artists and genres evolved. Moreover, history of restoration of artworks benefits of pictures taken at different times to reconstruct the physical evolution (or degradation) of the artworks.

- The acquisition of photographs include a broad network of people (collectors, owners, keepers, dealers). History of collecting and art market benefit of provenance information to define ratings of artists and evaluate trends in the market over time. Connoisseurs (scholars, art critics) obtain "trusty" evidences to rely on when attributing the authorship of a depicted artwork. The history of acquisitions is also the history of cultural institutions. The curation of such items in cultural

institutions requires the expertise of several figures (photo-archivists, revisors, supervisors), which take care of the acknowledgment of all the prior scenarios in accurate archival descriptions.

The four scenarios here outlined are addressed in cataloguing standards as four descriptive levels. According to The Functional Requirements for Bibliographic Records (FRBR), i.e., the conceptual model defined in Library and Information Science (LIS) for describing serial objects [Tillett, 2005], such levels represent (1) the conception of the work, (2) the realization of an expression, (3) the embodiment in a manifestation, and (4) the instantiation in a item.

Intrinsic and extrinsic features. "To date, detailed structural analyses of the intrinsic and extrinsic elements of photographic form produced by librarians for cataloguing purposes have been used by archivists as tools of description, not tools of appraisal" [Schwartz, 1995]. As pointed out by J. Schwartz, the physical description of photographs is mainly due to curatorial concerns, and to facilitate the discovery of photographs grouped under common labels. Thesauri like the *Getty Art and Architecture Thesaurus* [Petersen, 1990] aim at organising forms, functions, techniques and subject types of the photography in hierarchical lists of terms.

On the contrary, the evaluation of intrinsic features of photographs is differently perceived in art historical photo archives. A hermeneutic approach characterises the transcription of recto/verso of photographs, so as to highlight relations between the photograph and related entities (persons, organizations concurring in its creation), the connection with related cultural resources (bibliography), its provenance (prior archives or collections including the artwork), and scholars' assumptions (attributions).

Time(s) of the photograph. Time is captured by the photographer by fixing the image of an artwork in a precise time frame. However, the photograph carries other temporal information than the one related to its content, such as the time of the print, publication dates, and exhibitions dates.

The date of the shot is relevant to the analysis performed by historians of art restoration, who can benefit of the photographic documentation disposed in a timeline, and reconstruct the way the original artwork changed over time. Publications, as well as exhibitions, are evidences of the general interest in specific genres and artists whose works are represented in the photographs. The dates of the diverse publications contribute to build a chronology of bibliographic resources (e.g. handbooks, exhibitions catalogues, auc-

tion catalogues) related to the artwork.

Lastly, the history of cataloguing records reveals the history of institutional documentation - by whom changes are made, when and why, whether there is a cultural circumstance or a change in cataloguing standards, and so on.

Space and perspective. Space is determined by the photographer's sense of perspective. Such element determines the value of the document [Tunesi, 2014], and can tell stories about the creators' intentions (whether these are conscious or unconscious). Costanza Caraffa [Caraffa, 2011] provides a significant example of the usage of space to convey a conscious message to the observer. Analysing the photographic documentation preserved at the *Phototek of the Kunsthistorisches Institut in Florence* reproducing Vittore Carpaccio's *Sant'Orsola cycle*, she notices that some photographs, mounted on cards, are partially drawn. The drawing is meant to confer three-dimensionality to the represented scene, and emulate the perspective of an actual observer in the room. A similar situation is conveyed by collages, where the relation between single photos and the whole collection or album gives the observer an overall view of a work. The relation between the part and the whole is preserved in cataloguing records, so as to let the final user retrieve both particulars of the artwork and the complete series of photographs and reconstruct the way an artwork was originally displayed.

The ecosystem of art historical research. The art historical photo archive is a living entity, that grows organically according to several factors, like research interests in academy, acquisitions policies, and exchanges (whether photos are purchased from photographic agencies, received as part of an exchange with other institutions, cut out of auction house catalogues, or taken by private individuals). Among the objectives of photo archives, there is the gathering of massive amounts of photographs and other types of documents that would enable scholars to perform their research activities. It is worth to notice that by giving accessibility to visual representations of artworks, photo archives act as hubs in the art historical debate. Indeed, the visibility given to artists and genres through the documentation preserved, exhibitions, and other interventions, has amplified the value of some artworks [Schultz, 2015].

Nonetheless, not only photographs populate the bucket of art historical sources. Heterogeneous sources characterize the landscape of art historical research, which are addressed by existing cataloguing standards. We outline some of the resources that may be preserved or referenced in photo archives that are nonetheless

fundamental for connoisseurship activities.

The photographic collection. The arrangement of photographs in art historical photo archives may vary significantly. An arrangement based on provenance of photographic collections or the original arrangement tend to be preserved, e.g. fonds belonging to different providers may be merged into a single archive as a consequence of a bequest or an acquisition, but they are kept separated.

The classification of photographs can be based on the subject. For instance, folders and containers can be organised by artist, school, or geographical location. Series and subseries may be organised on the basis of the types of artworks depicted, e.g. architecture, sculpture, or painting, and divided by period. Both material objects (the photographs) and subjects (the artworks) are described in cataloguing records for the sake of accessibility to the collection. However, the semantic significance of photographs preserved may require bespoke standards to be applied, and to extend the description of the subject of photograph [Mambelli, 2014]. Archivists can record information about the artwork derived from the appraisal of the photographs (e.g. support, techniques, conservation status) and extracted from the analysis of the carrier (e.g. inscriptions, annotations, references to bibliography).

Along with the description of photos and artworks, authority files record information on people concurring to the creation of both the photograph and the depicted artwork. Photographers, firms, distributors, commissioners, and artists are disambiguated, uniquely identified, and briefly described in dedicated records.

In later phases of the cataloguing process, digitizations of photographs become a complement to online cataloguing records. This intervention requires an update of records, keeping track of the different manifestations the photo has, including the positive preserved, the original negative, the digital copies, and how to access them.

The Art history library. Art libraries have unique characteristics. Collections emphasize several arts forms, such as painting, sculpture, photography, graphic design, architecture. Various non-book formats, such as prints and slides, can be included in the collection, along with exhibition catalogs, auction guides and journals.

Special cases of art libraries are the personal libraries created by art historians, which may grow alongside their private photo archives, developed during their research activities or business. Among the others, a

significant example is the Federico Zeri's History of Art Library,[2] further discussed in Chapter IV. The organization of the library reflects the scholar's professional career, and the arrangement of volumes aims to preserve Zeri's mind map and cross-references among sections. It is worth to notice that references to volumes preserved in this and other art libraries appear on the back of photographs as evidences to support authorship attributions.

The classification used in art libraries conforms discipline standards, and produces subject-based bibliographic records. The cataloguing process differs significantly from the one adopted in photo archives, that is, provenance-based.

Specific bibliographic sources and archival sources for art history and history of photography. Diverse sources of information contribute to reconstruct the functional context of photographs. Cataloguers and scholars benefit of secondary sources such as exhibition catalogues, collection catalogues, museum catalogues, when seeking historical information.

Among the others, photographers' catalogues are precious resources. Photographers' catalogues mainly appear "in the form of numerical listings of a photographer's or firm's available work suitable for mail-ordering." [Sennett, 1986]. These documents record the holdings of galleries and museums, as well as architecture and landscapes, and many art objects, paintings, ruins, sites, and churches that may no longer exist. The study of such evidences in photo archives offers precious information on how to reconstruct the history of visual documentation, to make authorship attributions, and to date the objects. Likewise, auction catalogues and their illustrations help photo archivists and historians in identifying a work of art sold at auction, provide information on the costs of artworks and insights on market trends or interests. Moreover, they help to track the provenance of pieces of art and to reconstruct the history of collecting. Catalogues are described by using bibliographic standards, and may be referenced in the back of photographs to support attributions (authorship, dates, provenance, etc.).

Similarly, archival documents can support the cataloguing process and scholars' activities. Correspondences, expertises, and technical reports can be attached to photographs and provide insights on both photographs and artworks at hand. Such a documentation is generally inventoried and described at a high level, by using bespoke archival standards.

[2] See http://www.fondazionezeri.unibo.it/en/library/zeri-library

1.3 Cataloguing standards for describing the heritage of art historical photo archives

The debate around classification systems and cataloguing standards for the Photography and Arts domain focuses on one main, unsolved issue, i.e. "the tendency to reduce [photographs] to their visual content and therefore to identify them with the 'subject' illustrated - presupposing that it is always possible to say 'what is to be seen' in a photograph" [Caraffa, 2011]. Taxonomies proposed for classifying photographs are mainly based upon visuality, de-contextualising objects in favour of their subjects [Schwartz, 2002]. The poor expressivity of descriptive models narrows the boundaries of art historical research.

Despite systems of electronic classification and cataloguing emphasize such reduction of photographs to their material reality they represent (i.e. the artwork), few of current standards try to extend the set of fields to describe photographs "as autonomous objects and not just as the reproductions of something else" [Caraffa, 2011].

In the next sections we provide an overview of the most relevant standards for describing photographs, artworks, authorities for identifying people and places, bibliographic works, and archival documents. Such standards are among the ones adopted by art historical photo archives. For the sake of brevity, only extensive and representative domain-dependent standards are taken into account. Domain-independent standards, such as Dublin Core [Dublin Core Metadata et al., 2012], METS [Gartner, 2002], MODS [Gartner, 2003], and other lightweight standards are excluded from the survey since these address fields taken into account by domain-specific ones.

Standards here described can be grouped in four types, namely:

- *Content standards*: descriptive rules and field names, grouped in sections or themes, for the cataloguing of cultural objects.

- *Metadata content standards*: vocabularies of elements (or data element sets) for establishing a common understanding of the meaning of metadata (i.e. data about data). A standard exchange metadata format is the Extensible Markup Language (XML) [Bray et al., 1997].

- *Taxonomies and thesauri*: lists of terms, eventually disposed in a hierarchical structure, defining a consistent and normalized vocabulary for access points.
- *Authority files*: vocabularies of names for disambiguating and uniquely identifying names or subjects.

Content standards and metadata standards are here grouped according to their subject, i.e. photograph, work of art, bibliographic reference or archival document. Taxonomies, thesauri and authority files are described separately. Lastly, Italian cataloguing rules for describing photographs and artworks are thoroughly analysed in a separate section. The aim is to introduce topics addressed in the Zeri & LODE project - the use case discussed in Chapter IV - and to report on one of the most comprehensive element sets for the description of the Photography and Arts domain [Ronzino et al., 2011] that will be used to validate our hypotheses.

Standards for describing photographs. *Graphic Materials: Rules for Describing Original Items and Historical Collections* [Parker, 1982] is a set of guidelines, published in 1982 by the Library of Congress, for cataloging photographs, cartoons, popular and fine prints, architectural drawings, and other visual materials. These rules are a national standard supplement to Chapter 8 of the *Anglo-American Cataloguing Rules (AACR2)*, which focus on modern, published audiovisual materials. As widely discussed by J. Schwartz [Schwartz, 2002], bibliographic models pursue a subject-based classification of photographic documentation, and do not include sufficient indications for a context-based description. Its successor, *Descriptive Cataloging of Rare Materials (Graphics)* [Committee, 2013], published in 2013, expands on Graphic Materials by including instructions for born-digital materials, graphic material with formal title pages, and illustrations in books and serials, and considers collection-level records.

SEPIA Data Element Set (SEPIADES) [Klijn and de Lusenet, 2003] is a data element set that has been recently created to catalogue photographic collections, developed in the framework of the *SEPIA (Safeguarding European Photographic Images for Access)* project, that ran from 1999 until 2003. It includes elements for describing the multilevel hierarchy of the photo archive, the institute, the collection, including groupings and single items described as both visual and physical images. It offers a way to organise knowledge in hierarchical levels without reducing the organisation to a subject-based classification. Nonetheless, it only includes a restricted set of 21 mandatory elements, which does not allow to describe all the aspects related to the nature of the photograph.

Standards for describing artworks. *Categories for the Description of Works of Art (CDWA)* is a set of guidelines for the description of art, architecture, and other cultural works [Baca and Harpring, 2009]. CDWA is also a XML Schema, called *CDWA Lite*. The CDWA Lite schema has been enlarged and integrated into the *Lightweight Information Describing Objects (LIDO)* schema.

The *VRA Core 4.0* [Cowles, 2014] is a metadata standard (a XML schema) for the description of works of visual culture as well as the images that document them. It extends METS so as to describe cultural heritage resources. It includes 19 elements for describing three main entities, namely: collections, works, and images depicting artworks.

Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images [Harpring et al., 2006] is a content standard developed and maintained by the The Visual Resources Association Foundation (VRAF), mainly adopted in United States. It is designed "for members of the communities engaged in describing and documenting works of art, architecture, cultural artifacts, and images of these things" [Harpring et al., 2006]. Elements from *VRA Core 3.0* and from *Categories for the Description of Works of Art (CDWA)* are here included.

SPECTRUM [McKenna and Patsatzi, 2007] is the UK Museum Documentation Standard for Collections Management. It contains procedures for documenting objects and the processes they undergo, as well as identifying and describing information to support the procedures themselves.

LIDO [Coburn et al., 2010] is a XML schema launched in 2010. It meets requirements articulated by CDWA Lite, and is aligned to SPECTRUM. It's worth to notice that such standard emphasizes the description of the context of production of the object, allowing the description of relations between cultural objects and people concurring to the object creation.

MIDAS Heritage [Heritage, 2012] is a British cultural heritage standard for recording information on buildings, monuments, archaeological sites, shipwrecks and submerged landscapes, parks and gardens, battlefields, artefacts and ecofacts. The second edition, published in 2007, was published by English Heritage (today Historic England). It suggests the minimum level of information needed for recording heritage assets.

Taxonomies and thesauri for the arts and photography domain. The *Art and Architecture Thesaurus*

[Petersen, 1990] offers a controlled vocabulary of terms for describing items of art, architecture, photography and material culture. "It was designed to provide the 'hinge' between the object, its images, and related bibliographic material" [Petersen, 1990]. It includes hierarchical relationships, equivalences, and associative relationships between concepts. Concepts include abstract concepts (related to phenomena and human activities), physical attributes of objects, styles, periods, roles, professions, activities, materials, types of object, and brand names.

The *Getty Iconography Authority (IA)*[3] is a thesaurus, containing equivalence, hierarchical, and associative relationships between concepts representing proper names and other information for named events, themes and narratives from religion/mythology, legendary and fictional characters, themes from literature, works of literature and performing arts, and legendary and fictional places. It is based on the *Subject Authority* of the *Categories for the Description of Works of Art (CDWA)*.

The *Thesaurus for Graphic Materials* [Alexander and Meehleib, 2001] is an aid for indexing visual materials by subject and by genre/format. The thesaurus includes more than 7,000 subject terms and 650 genre/format terms to index types of photographs, prints, design drawings, ephemera, and other pictures. In 2007, the subject and genre/format vocabularies, previously maintained separately, were merged into a single list and migrated to new software, MultiTes.

Library of Congress Subject Headings (LCSH) [Chan, 1995] is a standard produced by the Library of Congress that contains around 5 million terms for both topics and names used as subjects. Given the broad range of topics addressed by the standard, it can be used or aligned to many other existing ones, even if it is not specifically designed for arts and photography domains.

Iconclass [Pałubicki et al., 1978] is a classification system designed for art and iconography. It is the most widely accepted scientific tool for the description of subjects represented in images (works of art, book illustrations, reproductions, photographs, etc.). It includes 28,000 hierarchically ordered definitions divided into ten main divisions, 14.000 keywords, and 40.000 references to books and articles of iconographical interest (not yet online).

The *UNESCO Thesaurus* [Aitchison, 1977] is a controlled and structured list of terms, including a multidisciplinary terminology, used in subject analysis and retrieval of documents and publications in the fields

[3] See http://www.getty.edu/research/tools/vocabularies/cona/about.html#ia

of education, culture, natural sciences, social and human sciences, communication and information.

The *UKAT UK Archival Thesaurus* [Carlisle, 2003] is a subject thesaurus that uses as its backbone the UNESCO Thesaurus, specifically designed for the archive sector.

Authority files of place and person names. The *Getty Thesaurus of Geographic Names (TGN)* is a structured, world-coverage vocabulary of around 2 million place names, including vernacular and historical names, coordinates, and place types, and descriptive notes, focusing on places important for the study of art and architecture [Harpring, 1997].

The *Getty Union List of Artist Names (ULAN)* is a structured vocabulary containing more than 600,000 names and biographical and bibliographic information about artists and architects [Harpring, 2010].

The *Library of Congress Name Authority File (LCNAF)*[4] is created by the Library of Congress with contributions from other libraries. It provides 8.2 million name authority records, including 6 million personal names, 1.4 million corporate names, 180,000 names for meetings, and 120,000 geographic names.

Finally, the *Cultural Objects Name Authority (CONA)* [Harpring, 2010] is part of the vocabularies published by the Getty Research Institute. It includes titles, attributions, depicted subjects, and other metadata about works of art, architecture, and cultural heritage. It contains links to artists and patrons, styles, dates, locations, studies and other related works, bibliography, images of the works, and subjects depicted in the works. It provides unique, persistent numeric identifiers for works and allows the disambiguation between similar works and authoritative identification of the work in a linked environment.

Standards for describing bibliographic entities. The *International Standard Bibliographic Description for Monographic Publications (ISBD)* [IFLA, 1974] is a standard issued by the *International Federation of Library Associations and Institutions (IFLA)* that specifies requirements for the description of published resources, including printed texts, maps, notated music, multimedia, and still images (e.g. photographs). Dated 1969, it was revised several times, and its evolution guided the creation of *Functional Requirements for Bibliographic Records (FRBR)*.

Functional Requirements for Bibliographic Records (FRBR) [Saur, 1998] is an entity-relationship model with four primary entities - work, expression, manifestation, and item - representing the products of intellectual

[4] http://authorities.loc.gov/

or artistic endeavour. A work is defined as the product of the intellectual or artistic activity by a person, a group, or a corporate body that is identified by a normalized title and/or name. An expression is the "realization of a work in the form of alphanumeric, musical, or choreographic notation, sound, image, object, movement, or any combinations of such forms" [Saur, 1998]. "[...] a manifestation represents all the physical objects that bear the same intellectual and physical characteristics" [Saur, 1998]. An item is a single example of one, single manifestation. Such a multi-layer definition is meant to be applied to bibliographic records and fits the representation of photographs as well, i.e. serial and multi-faceted objects. To this extent, FRBR is a good candidate for the high level representation of aspects characterising the nature of photographs.

Resource Description Access (RDA) [Coyle and Hillmann, 2007] is a set of data elements, guidelines, and instructions for describing library and cultural heritage resources that is going to replace *AACR2*. Among the underlying conceptual models for RDA is FRBR that is extended with a wide set of properties for describing relations between bibliographic records and people, organisations, and other bibliographic records.

Standards for describing archival keepings. *General International Standard Archival Description (ISAD (G))* [ICA, 1999] is the *International Council of Archives (ICA)* structural standard for archival description.

> Archival description represents a fonds, a complex body of materials, frequently in more than one form or medium, sharing a common provenance. The description involves a complex hierarchical and progressive analysis. It begins by describing the whole, then proceeds to identify and describe sub-components of the whole, and sub-components of sub-components, and so on. Frequently, but by no means always, the description terminates with a description of individual items. The description emphasises the intellectual structure and content of the material, rather than their physical characteristics. [Pitti, 1999]

As outlined by D. Pitti, the archival structure is mainly hierarchical and the description of the functional context is taken into account. Nonetheless, the archival description seldom addresses the content of single documents - the folder is often considered the smallest unit of description. Thus, considering the need to access information related to single documents in photo archives, such standard has to be integrated with others, addressing the specific nature of the photographic object.

Encoded archival description (EAD) [Pitti, 1999] is a XML schema for encoding archival finding aids, for representing the hierarchical structure of a fond, and to exchange information in digital format. It is based on ISAD(G), and it is used closely in association with Encoded Archival Context (EAC(CPF)) for describing agents concurring in the creation of the archive.

The *International Standard Archival Authority Record for Corporate Bodies, Persons and Families ISAAR(CPF)* [Thibodeau, 1995] was published by the International Council on Archives in 1996 and revised in 2000. It provides guidance for preparing archival authority records including descriptions of entities (corporate bodies, persons and families) associated with the creation and maintenance of archives. Archival authority records are similar to library authority records in as much as both forms of authority record need to support the creation of standardized access points in descriptions.

> Archival authority records, however, need to support a much wider set of requirements than is the case with library authority records. These additional requirements are associated with the importance of documenting information about records creators and the context of records creation in archival description systems. As such, archival authority records go much further and usually will contain much more information than library authority records.[5]

According to the official statement, such a description of the archive creator complies with requirements highlighted in photo archives with regard to the description of the functional context of cultural objects.

Similarly to EAD, the *Encoded Archival Context (EAC(CPF))* [Pitti, 2004] is the XML schema based on ISAAR(CPF) for sharing archival authority records in digital format.

In summary, the survey of existing cataloguing standards shows that the definition of photograph is often fudged into a broader labelling, such as "graphic materials". Existing rules do not provide extensive sets of elements for describing the archival nature of the photograph nor its features. Each type of surveyed standard may potentially contribute to provide a comprehensive description, namely: (1) archival standards allow a hierarchical representation of collections, preserving context information; (2) library standards allow a multilayer description of the nature of the catalogued object, and (3) museum standards focus on

[5] International Council on Archives, ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families. Second edition. 2004, p. 7.

visual aspects of objects. In the next section we finally introduce the Italian cataloguing rules for describing photographs and artworks, that present most of the aspects required to describe photographic collections and support connoisseurship activities.

Italian cataloguing rules for describing photographs and artworks. In Italy several cataloguing rules may apply to the description of photographs, according to the tradition of the cataloguing institute (i.e., archives, museums, libraries). For example, photographs may be catalogued according to ISBD in libraries, or according to ISAD(G) in archives.

The *Istituto Centrale per il Catalogo e la Documentazione (ICCD)* provides descriptive standards for the cataloguing of cultural heritage objects under the protection of the *Italian Ministry of Cultural Heritage and Activities and Tourism (MIBACT)*.[6] ICCD issued different cataloguing rules according to the different types of heritage, including archaeological sites, artworks, photographs, music instruments, numismatic, etc. The descriptive approach of such rules is close to the museum tradition.

All the cataloguing standards share a common kernel of descriptive elements, listed in the cataloguing standard called *ICCD-NTR*, and few sections addressing the description of the specific cultural object. Records are hierarchically structured in paragraphs, fields and subfields. Paragraphs are the main units of information, which address a specific theme related to the identification or description of the catalogued object. The amount of fields is comprehensive and potentially very detailed. Indeed, around two hundred fields for each normative document. The mandatory paragraphs are listed below, identified by a sigla in squared brackets and the english translation.

CD Codes. Identifiers of the cataloguing record and institution

OG Object definition. Type of the object, part/whole relation, and type of cataloguing process

LC Localization. Geographical location and repository of the artefact

DT Creation dates. Period, sources, and motivation for the attribution of the date

MT Technical data. Materials, medium and measures

[6] See `http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici` for the complete list of standards issued by ICCD.

CO Assessment of conservation status.

TU Juridical entity. Keeper, acquisition dates

DO Documentation. Objects (e.g. photo, audio, bibliography) documenting the artefact

AD Data access policies. User groups and motivation

CM Data management and update. Supervisors, cataloguers, and dates of review

Optional paragraphs include:

RV Relations between entries. Hierarchy of records and described artefacts

AC Other codes. Identifiers of records preserved elsewhere related to the catalogued object

LA Other localizations. Previous geographical locations and repositories of the artefact

UB Inventory. Inventorial numbers and classification

CS Cadastral data.

LS Historical location. Names and dates of the historical name of a place

GE Georeferencing. Geographical coordinates

CT Cartographic references.

RE Archeological findings.

AU Authorship. Authors, sources, motivations, other attributions, cultural context

DA Analytical data. Iconography, inscriptions, stamps, etc.

UT Usage. Fields of application

MS Exhibitions. Titles, places and dates

AN Notes. Archivist's annotations

Several types of graphic materials are addressed in bespoke cataloguing rules, namely: photograph (ICCD-F), photo collection (ICCD-FF), drawings (ICCD-D), engravings (ICCD-MI), artworks (ICCD-OA), contemporary artworks (ICCD-OAC), prints (ICCD-S). Standards relevant to art historical photo archives are the ICCD-F, where F stands for fotografia, i.e. photograph, and ICCD-OA, where OA stands for opera d'arte, i.e. artwork. ICCD also provides guidelines for the realisation of authority files, which include - among the others - authors (ICCD-AUT) and bibliography (ICCD-BIB).

All schemas may contain a description of photographic documentation (see above *[DO] Documentation*). However, photographs are addressed as autonomous objects and are potentially described as cultural objects themselves by means of a dedicated record (ICCD-F). It's worth to notice that ICCD-F includes a paragraph dedicated to the archival description of the photograph, called *[UB] Archival classification*. It records the position of the photograph in the hierarchical organisation of the fond, a reference to the finding aid (where applicable), and a shelfmark.

Furthermore, all the paragraphs recording information obtained by means of the cataloguer's subjective interpretation include sub-fields for recording sources of information, criteria and motivations.

In particular, ICCD-F and ICCD-OA can be deemed the most representative standards for the formal representation of the wealth preserved by art historical photo archives. We assume that cataloguing records including all such information offer sufficient insights to tackle issues introduced by the functional context of objects, the nature of the photographic object, and questionable statements that can be derived from the appraisal of documents. In the second part of this work, we leverage cataloguing data compliant with such standards so as to extract information on cataloguers' hermeneutic approach and provide a formal definition of authoritativeness of records including authorship attributions.

1.4 Connoisseurship. Research and application fields in art historical photo archives

Art historical photo archives are places of many research activities. Few research fields have already been mentioned, such as history of photography, art history and its subfields, including history of restoration,

historiography of art history, history of collecting, history of cultural institutions, and connoisseurship. Among the others, the latter has a great impact in society, and receives more echo in public opinion, due to the many implications it has in academy and market.

Connoisseurship has been defined in many ways by scholarship, e.g. "the *grand dame* of the art history" [Maginnis, 1990], "the art of appreciation" [Eisner, 1996], the recognition of "the goodness of a picture" [Richardson, 1719], with the "express purpose to establish correct provenance, or more specifically to find out who painted what, in particular to authenticate paintings" [Gombrich, 1985]. The vagueness of some of these definitions shows the questionable nature of the practise. Nevertheless, it is of great financial importance in Art world, and because of this bias it has been criticised by scholars.

The social and financial pressure on connoisseurs' attributions is among the arguments arisen against connoisseurship as a doubtful practise. "Markets - galleries, dealers, auction houses - exercise undue pressures. They induce experts to make judgements - both of quality and of attribution - that suit their own pockets rather than the purposes of disinterested scholarship" [Freedberg, 2006].

Secondly, it is hard to define a shared and reproducible methodology characterising the process of ascribing an artwork to an artist. Many authors tried to define the scientific approach of connoisseurship. Scholarship often refers to the Morellian method of making attributions [Morelli and Richter, 1883] as the touchstone for defining - or contradicting - criteria of connoisseurship. The Morellian method is based on the recognition (or better the intuition) of features of a painting that are unconscious rather than conscious. Carlo Ginzburg [Ginzburg, 1979] shows how the epistemological model in the Humanities that emerged in nineteenth century owes its methodology to connoisseurship, and to the Morellian method in particular. He explains that such method relies, rather than on intuition, on the expertise in many disciplines, and points out its interdisciplinary nature and its "cognitive richness".

Another element that affects reliability of connoisseurship is the lack of evidences capable to support a subjective statement. In fact, pitfalls derived from the subjectivity of the method persist even when technology and scientific data are applied to the analysis, e.g. e-ray photography, spectrography, and infrared reflectography, since "the eye" is still a key variable in the process [Freedberg, 2006].

Quantitative approaches, i.e. the comparison of as many visual evidences as possible, are not applicable in most cases, or they may induce to misleading assumptions anyway. Gary Schwartz explains the flaws

of such an approach as follows:

> Essentially, what the connoisseur does is to define a relationship between an existing work and an historical category. Dealing with works of uncertain status, the connoisseur treats the other two elements in the equation as givens: the categories are formed by works whose authorship is firmly documented, and defining the relation is an analytic technique whose ins and outs can be explained, although they mysteriously continue to resist codification. A closer look reveals that the two 'givens' the categories and the techniques by which unknown works are matched to them-are actually quite dubious. The connoisseur's comparative material consists, in theory, of existing works whose authorship is documented. This sample, historically precious as it is, is however insufficient for the stated purposes of connoisseurship. The disappearance from sight of the entire oeuvre of many documented masters distorts the historical record, so that the connoisseur's categories do not correspond to historical reality. It is as if the ordered contents of a number of containers were to be dumped on the ground in a heap, and half the containers then broken, and one then tried to sort the same material into half the original number of containers. A valuable, perhaps necessary exercise, but one should not entertain any illusions concerning its historical truth. [Schwartz, 1988]

Lastly, relying on shared and authoritative opinions becomes a fundamental criterion used by connoisseurs to support their own claims. By quoting the judgement provided by an authority in the field, connoisseurs back up their assumptions thanks to the trustworthiness granted to somebody else. Nonetheless, this approach may also lead to mis-ascriptions, since the definition of authoritativeness applied to connoisseurs is questionable too. As Freedberg wonders,

> We can all agree that in making attributions we rely not only on our own judgement, but also, to a greater or lesser degree, on the opinion of the best possible authority in the field. But on what bases do we decide to trust an authority? How do we determine the criteria for defence? The easy answer is to say that we rely on the most impartial judge of paintings, the one who is least likely to be swayed by market or social pressures, the one that ha that indefinable quality, 'the best eye', as we so often like to say. [Freedberg, 2006]

In Freedberg's analysis trust is defined as a social product, in which the shared acknowledgment of a person as an expert implies the general acceptance of her opinions. Although, the general acknowledgment may not coincide with an individual one. Somebody can question the reliability of the acknowledged person because there are evidences of a bias (e.g. the aforementioned pressure of market on her opinions). Hence the questions: how to define authoritativeness of people we trust? What are the evidences that support such a judgement? Following Freedberg's reasoning, a criterion is to rely on experts that published their peer-reviewed results. Even though, publications may be tainted by external pressures too, and trust is still the key to evaluate authoritativeness.

Indeed Freedberg's last guess is crucial: "trust plays a defining if not imperative role in the constitution of [e]very kind of knowledge" [Freedberg, 2006]. Trust is also defined as a matter of cooperation, i.e. the hope that moves people to believe in others' opinions, the benevolence of the trustee in believing, and the acknowledged impartiality of the knowledge provider. That's where photo archives come into play. Archives are deeply involved in the creation of such a trustworthy network, made of heterogeneous evidences (more or less biased), opinions (more or less evaluated), and trustee's benevolence (the users they are devoted to). Archives offer primary sources required for comparative studies, including both textual and visual evidences, and make a first evaluation of those.

However archives are far from being neutral. [Brilliant, 1988] describes the path that leads a historian to get close to her subject, that starts from the retrieval of contextual information provided by bibliography and archival documents.

> Art historians may act like art critics in grasping the visual properties of objects, but they act like historians in surrounding the artifact with causes, effects, and circumstances - the ingredients of significance. The historical dimension of art history then requires the kind of information found in books, in periodicals, in old records, and in the varied forms of data collection and control which depend on texts and on writing. Learning about an art object diffuses the scholar's effort since context is a generalized abstraction; only gradually, as the connections become clear, can the historian close in on the subject of research. If the art library incorporates the discipline's mine of historical information, then the enterprising scholar must know where and how to dig up the bibliographical lore, always hoping to find

> a few unexpected treasures. [Brilliant, 1988]

Secondly, in photo archives archivists offer their own point of view, by helping users to get the sources they need, and by providing them with additional information, i.e. cataloguing records. As pointed out in the previous section, records include significant knowledge about the subject of photographs, which let alone is the result of archivists' attributions. The historical approach that underpins archivists' choices is condensed in these secondary sources of information. But are archivists - and archives - authoritative connoisseurs?

The methodology used by photo archivists is not always clearly stated in cataloguing records, assuming that the authoritativeness of the cultural institution issuing the information is enough to support the reliability of the statement - especially when circumstances (e.g. time, resources, background information) prevent them to detail such aspects in cataloguing forms.

Lastly, archives have been perceived for a long time as impartial and passive resources to be exploited for various historical and cultural purposes, or even as "neutral repositories of facts" [Schwartz and Cook, 2002]. Despite this misconception, archives are actually powerful knowledge providers. Rather than impartial witnesses and curators of history, they shape it continuously, and contribute to the storytelling of our past. J. M. Schartz and T. Cook describe the power of archives and their record-keeping systems as follows:

> Archivists have long been viewed from outside the profession as "hewers of wood and drawers of water", as those who received records from their creators and passed them on to researchers. Inside the profession, archivists have perceived themselves as neutral, objective, impartial. From both perspectives, archivists and their materials seem to be the very antithesis of power. [...] Nevertheless, various postmodern reflections in the past two decades have made it manifestly clear that archives - as institutions - wield power over the administrative, legal, and fiscal accountability of governments, corporations, and individuals, and engage in powerful public policy debates around the fight to know, freedom of information, protection of privacy, copyright and intellectual property, and protocols for electronic commerce. Archives - as records - wield power over the shape and direction of historical scholarship, collective memory, and national identity, over how we know ourselves as individuals, groups,

> and societies. And ultimately, in the pursuit of their professional responsibilities, archivists - as keepers of archives - wield power over those very records central to memory and identity formation through active management of records before they come to archives, their appraisal and selection as archives, and afterwards their constantly evolving description, preservation, and use. [Schwartz and Cook, 2002]

The authors point out how "[c]ertain stories are privileged and others marginalized" in archives. Authors refer to archives in general terms, but similar conclusions can be drawn for art historical photo archives too. In fact, the boundaries between trust in archives and their role of history rewriters is blurring. Acquisitions policies, accuracy of classification, and publication of selected records are all instruments that archives have in order to exercise their power on art history.

In summary, in this chapter we first highlighted the groundbreaking role of photography in the Cultural Heritage domain. In particular, photographic reproductions of artworks are acknowledged as fundamental tools for art historians, and art historical photo archives are addressed as hubs of art historical research. However, the nature of the photographic object is controversial. It is not a neutral objective evidence, but a historicized source.

Art historical photo archives would ensure authoritative sources are served to final users, giving value to the wealth of information there sedimented by means of descriptions complying with accurate cataloguing standards. As a result of the literature review, we realize that only the Italian cataloguing rules ICCD-F and ICCD-OA are applied to describe features related to cataloguers' hermeneutic approach and can effectively support historians in validating questionable information such as attributions. For this reason we restrict the study to photo archives that either use such standards or record all the aspects required to connoisseurship activities.

Chapter 2

Semantic Web Technologies and Digital Humanities Approaches to Art historical Research

In this chapter we describe the technical background of the research presented in the next chapters. First, basic concepts of Semantic Web technologies are introduced, including architecture components, vocabularies and ontologies, and principles for publishing Linked Open Data (LOD). Secondly, the state of the art of Linked Datasets and ontologies for describing the Cultural Heritage domain is introduced. Different approaches to knowledge organisation in *Digital Humanities (DH)* and *Library and Information Science (LIS)* fields are introduced so as to address problems related to the formal representation of metadata created by art historical photo archives. Lastly, we review existing methodologies for assessing Information Quality and address the dimensions that characterise the judgement of sources recording authorship attributions.

2.1 Semantic Web and Linked Open Data

Nowadays, the World Wide Web plays a key role in the dissemination and retrieval of information [Berners-Lee et al., 1992]. Information is mainly published on the web in the form of hypertext docu-

ments (1) annotated with Hypertext Markup Language (HTML) [Pemberton et al., 2000], (2) identified by Uniform Resource Identifiers (URI) [Berners-Lee et al., 2004], and (3) accessible through specific protocols such as Hypertext Transfer Protocol (HTTP) [Berners-Lee et al., 1996]. Hyperlinks interconnect documents with each others and allow users to access them, by means of web browsers. Web pages are mainly designed for human consumption and underlying data is not easy to be interpreted and reused by machines for intelligent tasks.

The *Semantic Web* [Berners-Lee et al., 2001] is an extension of the current web that is mainly focused on the exchange and the interoperability of data rather than the linking of documents. For this reason, the Semantic Web is also called Web of Data or Web 3.0. The rationale of such an approach is explained by Bizer, Heath and Berners-Lee [Bizer et al., 2011] as follows:

> Traditionally, data published on the Web has been made available as raw dumps in formats such as CSV or XML, or marked up as HTML tables, sacrificing much of its structure and semantics. In the conventional hypertext Web, the nature of the relationship between two linked documents is implicit, as the data format, i.e. HTML, is not sufficiently expressive to enable individual entities described in a particular document to be connected by typed links to related entities. [Bizer et al., 2011]

The aim is to create a space where both documents and data are published according to open standards and good practices (such as URI, HTTP, and the aforementioned XML format), are interlinked according to rules defined in data models, and are processed by machines in expressive ways. This collection of interrelated data is called *Linked Open Data*. The architecture of semantic web is illustrated in Figure 2.1.

Unicode[1] is the standard for encoding international character sets. A *URI* is a string that uniquely identifies a resource. *XML documents* [Bray et al., 1997] contain semi-structured information organised in a hierarchy of elements and attributes. *XML Namespaces* [Bray et al., 1999] allow to refeence vocabularies used in the same XML document, and *XML Schema* [Fallside and Walmsley, 2004] defines elements, attributes and rules to be applied to a set of XML documents.

[1] See `http://unicode.org/`

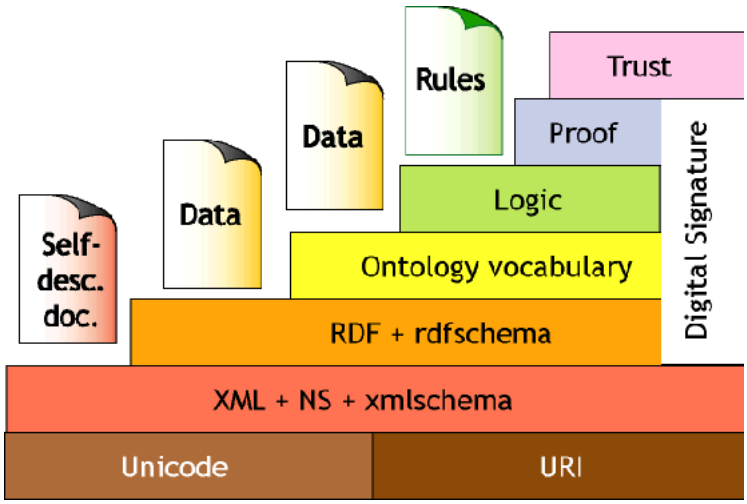


Figure 2.1: Semantic Web Architecture. Image from Berners-Lee, Tim. 2000. *Semantic web-xml2000*. `https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`.

The *Resource Description Framework (RDF)* [Hayes, 2004] is a framework for representing information in graph form. It is based on triple patterns in the form `<subject><predicate><object>`. XML is the standard serialization format for RDF, but several others are interchangeable and widely adopted.[2] *RDF Schema (RDFS)* [Brickley, 2000] describes taxonomies of classes and properties for defining terms and relations between terms referenced in triples. Ontologies, mainly formalized by means of the *Web Ontology Language (OWL)* [Antoniou and Van Harmelen, 2004], offer more constructs than RDFS in order to define semantics and reason within a knowledge base (i.e., a dataset including both the ontology and data). Ontologies are based on *Descriptive Logics*, and rules can be added to constructs provided by ontologies by means of bespoke languages, e.g. *SWRL* [Horrocks et al., 2004]. For querying RDF data, RDFS, and OWL ontologies, the *Simple Protocol and RDF Query Language (SPARQL)* [Prud et al., 2006] is available.

The *Proof layer* is supposed to demonstrate why agents (i.e. machines) should trust provided information, by creating a trusted network of information and data providers. Currently technologies for these layers do not exist. The security layer is not part of the Semantic Web stack, but developed as a separate Security Architecture that interfaces with that.

In order to frame the scope of this work, we briefly detail the RDF data model, the usage of ontologies, and Linked Data principles.

[2] See for example [Alexander, 2008]

Resource Description Framework (RDF). The data model of RDF is pretty simple. The basic construct is a RDF statement, represented by a triple including three terms, namely: a subject, a predicate (or property), and an object. The subject is described by means of predicates, that are attributes of the subject. The object is the target value of the triple. The three terms of a triple are all identified by permanent, dereferenceable URIs.

Data represented as RDF statements may refer to online documents, but also to real entities (e.g. people, artworks) and abstract concepts (e.g. time, activities, techniques), which are linked together by meaningful relations. Things described in RDF are also called *resources*. Every triple can be graphically represented as node and arc diagrams, as exemplified in Figure 2.2.

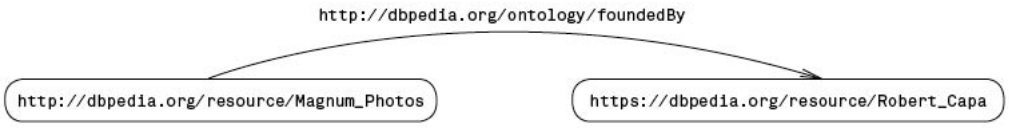


Figure 2.2: A triple statement example

In the above example, subject and object are nodes, and the predicate is the arc connecting the two nodes. All of the three terms are identified by URIs, namely:

- the URI `<http://dbpedia.org/resource/Magnum_Photos>` represents the well known photographic cooperative Magnum Photos
- the URI `<https://dbpedia.org/resource/Robert_Capa>` represents the photographer Robert Capa
- the URI `<http://dbpedia.org/ontology/foundedBy>` represents the relation "is founded by" between the subject (Magnum Photos) and the object (Robert Capa).

A triple statement can link resources to other resources, like in the above example, or to a literal value, like a text, a number, or a date.

The RDF model can be represented as a directed graph, based on the arcs connecting several nodes. The benefits of the RDF model over existing ones, e.g. the tree model (XML-alike), are several [Bergman, 2009], namely:

- *Everything is identified by HTTP URIs.* URIs are web-compatible and scalable.
- *It is easy to implement.* Triples can be efficiently implemented and stored than other models, which require variable-length fields and a more cumbersome implementation.
- *It is graph-based.* RDF is the canonization of a (directed) graph, and so as such has all the advantages (and generality) of structuring information using graphs. Graphs are modular and can be both readily combined and broken apart, allowing exploration of large networks.
- *It is interoperable.* RDF can both capture and convey metadata in unstructured (e.g. text), semi-structured (e.g. HTML documents) or structured sources (e.g. databases).
- *Mapping flexibility and data integration.* The separation of concerns regarding instance data (called ABox) and schema structure (TBox) makes easy and cost-effective the incorporation of new datasets wherein only new attributes require a structure update.

Vocabularies and ontologies. In the context of Semantic Web, vocabularies and ontologies are used to define terms that describe a domain. Vocabularies can be very simple, and describe only few classes and relations among individuals belonging to classes. Ontologies place more constraints on sets of individuals and the relationships [Cimiano, 2006], and can represent a domain in many expressive ways. However, the distinction between the two terms is not clear, which are often used at the same way.

Plenty of definitions of formal ontologies exist in literature. Thomas Gruber defined ontology as "an explicit specification of conceptualization" [Gruber, 1993]; Borst as a "formal specification of a shared conceptualization" [Borst, 1997], and Studer et al. as "a formal, explicit specification of a shared conceptualization" [Studer et al., 1998]. Nicola Guarino et al. [Guarino et al., 2009] summarise the activities in ontology engineering as follows:

> Computational ontologies are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation [...]. The ontology engineer analyzes relevant entities and organizes them into concepts and relations, being represented, respectively, by unary and binary predicates. The backbone of an ontology consists of a generalization/specialization hierarchy of concepts, i.e., a taxonomy. [Guarino et al., 2009]

An ontology includes a formal description of (1) concepts in a domain of discourse (called classes), (2) properties of each concept describing various features and attributes of the concept (called slots, roles or properties), and (3) restrictions on slots (facets or role restrictions). An ontology, defined in a terminological section (called TBox) together with a set of individual instances (called ABox) of classes constitutes a knowledge base. [Noy et al., 2001].

Classes are the focus of most ontologies, which describe concepts in the domain. For example, we can describe images with the class "Image". A class can include subclasses that represent more specific concepts than the superclass and may form a taxonomic hierarchy of concepts. For example "Moving image" and "Still image" are both subclasses of the class "Image". Properties describe attributes of both classes and their instances (i.e. individuals, or unary relations, belonging to classes) and are inherited by individuals of subclasses. For example "creator", "date", "subject" are all properties, also called binary relations, of an image. Allowed values of these properties may be individuals of classes such as "Person", "Time span", "Subject".

Ontologies are pillars of Semantic Web: they are crucial in data integration activities [Wache et al., 2001], they help to disambiguate terms used in the different datasets, and the knowledge they carry can be used to discover (say, infer) new latent relations by means or reasoners. Guidelines for ontology design and methodologies for ontology development have been created[3] so as to ensure comprehensiveness, representativeness, and consistency of resulting models. Good practices, such as documenting the ontology and reusing existing models, rather than creating new ones from scratch, facilitate developers to reuse ontologies in new contexts [Poveda Villalón et al., 2012].

Linked Open Data. Publishing Linked Data requires to comply with a set of best practises, also called Linked Data principles [Berners-Lee, 2006], namely:

- Use URIs as names for things. As said, URIs can identify both documents and real entities, that are both subjects and objects of triples
- Use HTTP URIs so that people can look up those names. Machines dereference HTTP URIs so as to return a human-readable description of the resource when users' request it in browsers, and RDF structured data to application that ask for data

[3] See [Fernández-López and Gómez-Pérez, 2002] for an overview

- Provide useful information about what a name identifies when it is looked up, using open standards such as RDF, SPARQL, etc.

- Refer to other things using their HTTP URI-based names when publishing data on the Web. Rather than the untyped hyperlinks used in the Web of Documents, here links are meaningful relations, i.e. RDF predicates identified by HTTP URIs as well.

Linked Data allow to connect different data sources and let data be easily discovered and used by applications. Different data providers can contribute portions of their data as statements. The integration of data sources can offer a bigger picture of a domain, also including contradictory statements on the same topic. Moreover, the integration of data sources belonging to different domains offers a powerful tool for data-driven applications, that can leverage heterogeneous information to build expressive services on top of those sources. For example, by georeferencing artworks (i.e. mixing geographical data and data about artworks), maps of cultural sites can be created to recommend touristic routes.

2.2 Knowledge Organization in the Cultural Heritage domain

Nowadays cultural heritage institutions like galleries, libraries, archives and museums (GLAMs, or simply LAMs) are looking for new ways to engage and educate patrons. Linked Open Data (LOD) are deemed good candidates to support institutions in disseminating their heritage. According to [Marden et al., 2013],

> LOD gives LAMs the opportunity to set their collections free from silos and place them in multiple contexts by pairing them with different LOD sets from around the world. Essentially, LOD allows users to interrelate communication artifacts without needing the interpretation of an archivist, curator or librarian. This ability for users to create their own relationships between artifacts is an important aspect of communication design. [Marden et al., 2013]

Leveraging LOD offer cultural institutions a variety of benefits, including: (1) to enrich their own collection data integrating information they miss, (2) to provide context information to data they already have, (3) to expose data in expressive and easy-to-access ways for reuse. As a consequence, they can (1) promote

their siloed data, (2) increase their visibility, and (3) establish themselves as trusty and authoritative sources of high quality data in the Semantic Web. Benefits for final users, whether these are researchers, developers, or citizens, include (1) availability of higher data quality, (2) improved accessibility of data, and (3) better user experience in new services.

[Van Hooland and Verborgh, 2014] highlight the importance and the political standing of linked open data publishing in the cultural heritage domain. They point out how competitors in the market of knowledge bases (e.g. Google knowledge graph[4], Facebook [Ugander et al., 2011]) and metadata schemes (e.g. Schema.org [Guha et al., 2016], developed by Google, Bing,[5] and Yahoo![6]) play a significant role in imposing the way semantics has to be expressed, and how the linked open data cloud is going to evolve. To hinder this uncontrolled growth of monopolies, the authors emphasise the disruptive potential of Library and Information Science (LIS) and Digital Humanities (DH) fields.

> [...] the LIS and DH communities should use their unique potential to stand up and launch a debate on these matters. Though their historical interest in exceptional values and outliers, the humanities can and should collaborate with engineers on how to facilitate and safeguard the access to the long tail of values which traditionally are disregarded in a probabilistic approach. [Van Hooland and Verborgh, 2014]

Linked Open Data in GLAM. In recent years, many cultural institutions published their data as linked datasets. The Library of Congress (LOC) published its subject headings and authority names into the *LOC Open Data Portal* [Summers et al., 2008]. Likewise, the *Swedish Union Catalog - LIBRIS* [Sanders, 2017], began sharing linked data in 2008, and to date is the only library that completely moved their data management workflow to the Linked Data paradigm. Other library projects using linked data include the *British National Bibliography* [Deliot, 2014], and the *French National Bibliography* [Simon et al., 2013].

Similar efforts in archival domain include: the *Linked Open Copac and Archives Hub (LOCAH)*, evolved in *Linking Lives Project* [Ruddock, 2011], which publishes biographical data from UK and Irish archives; the *Linked Jazz Project* [Pattuelli et al., 2013], including archival records and photographs of jazz artists; the

[4] `https://www.google.com/intl/bn/search/about/`
[5] `https://www.bing.com/`
[6] `https://www.yahoo.com/`

Social Networks and Archival Context (SNAC) initiative [Larson and Janakiraman, 2011], which publishes archival records of archive creators; the *ReLOAD Linked Open Archival Data* [Ricci, 2014], publishing archival data of Italian cultural institutes.

Museums related projects include: the *American Art Collaborative (AAC)*, a consortium of 14 art museums in the United States [Knoblock et al., 2017]. the *British Museum*[7], and the *Amsterdam Museum* [De Boer et al., 2012].

In the spirit of the Linked Open Data movement, opportunities for cooperation came up in the Cultural Heritage domain. Data belonging to different types of institute are gathered and served by means of aggregators. Among such experiences are the *Dutch Heritage Innovators Network project Open Cultuur Data* [Grob et al., 2011], *Europeana* [Haslhofer and Isaac, 2011], and *Worldcat*, the OCLC Online Union Catalog [Bennett et al., 2003].

Other resources published as LOD contribute to facilitate data integration, such as the already mentioned Getty thesauri *AAT, TGN*, and *ULAN* [Harpring, 2013], the *VIAF* authority file,[8] *DBPedia* [Auer et al., 2007], *Wikidata* [Erxleben et al., 2014], and *geoNames* [Wick and Vatant, 2012]. However, to date there is no authority file addressing pieces of art and photographs available as LOD.

Ontologies for the Cultural Heritage domain. Ontologies are widely used in the Cultural Heritage domain for knowledge organization and information management tasks. As aforementioned, the Cultural Heritage domain is not a single, homogeneous, domain. Different traditions in knowledge organization characterize the landscape of GLAMs from which originated different conceptual models, cataloguing rules, and record-keeping systems.

To this extent, plenty of ontologies were developed to represent subsets of the Cultural Heritage domain, namely: libraries, archives, and museums. Ontologies and vocabularies are mainly derived from existing metadata standards so as to represent contents of existing cataloguing records in a machine-readable format. The shift from a document-centric view (typical of the previous standards) to a data-centric view lies down in the adoption of LOD for metadata modelling, encoding, representation, and sharing [Di Noia et al., 2016].

[7] `http://bnb.data.bl.uk`
[8] `http://viaf.org`

Table 2.1 includes a survey of existing ontologies, grouped by field of application (LAM), and a brief description [Daquino et al., 2019]. For the sake of completeness, the last section of the table includes cross-domain ontologies and task ontologies. The latter provide a representation of concepts and relations related to single topics characterizing multiple domains (e.g. provenance of information, people). Lastly, Ontology Design Patterns [Gangemi and Presutti, 2009], i.e. basic building blocks to be used in ontology design, are here referenced to include other scenarios not addressed in prior models. The aim of the survey is to investigate and review the wealth of solutions that can be used for representing the Arts and Photography domain.

Table 2.1: Survey of ontologies and vocabularies

Libraries Ontologies	
Resource Description and Access RDA Ontology[9]	RDA formalizes terms defined in the omonimous cataloguing standard. It uses the entity-relation (ER) model of FRBR. Some of the relationships are entirely new to library cataloging, such as relationships between Works and relationships between persons and Works and Expressions, which have traditionally been expressed in library metadata as attributes of the person heading [Hillmann et al., 2010]
BIBO Ontology[10]	BIBO is widely used in the bibliographic community. The ontology can be used as a citation ontology, as a document classification ontology, or simply as a way to describe any kind of document in RDF. *bibo:Document* is the core class of this model. It includes DC, PRISM and FOAF terms, and adds other classes and properties such as *bibo:AcademicArticle*, *bibo:Journal*, *bibo:Collection*, *bibo:Book*, *bibo:Chapter* and *bibo:Issue* to describe the publishing domain.
BIBFRAME[11]	BIBFRAME is developed by the Library of Congress, and it is considered the successor of MARC. BIBFRAME 2.0 organizes information into three core levels of abstraction: Work, Instance, and Item. It lacks of the definition of Manifestation, which is included in the broader definition of Item [Kroeger, 2013].

[9] `http://www.rdaregistry.info/rgAbout/rdaont/`

[10] `http://bibliontology.com/`

[11] `https://www.loc.gov/bibframe/`

Semantic Publishing and Referencing (SPAR) Ontologies[12]	The SPAR ontologies are a suite of modular ontologies that together describe aspects of semantic publishing and referencing, including resources, workflows and involved agents. Among the others, it includes a FRBR OWL2-DL model, which improves the FRBR Core ontology; FaBiO, the FRBR-aligned Bibliographic Ontology; CiTO, the Citation Typing Ontology, for the characterization of bibliographic citations; and the Publishing Roles Ontology (PRO) for defining agents' roles (e.g. author, publisher, editor, reviewer) [Peroni and Shotton, 2018b].
Archives Ontologies	
EAC-CPF Ontology[13]	The EAC-CPF Ontology formalizes terms of ISAAR(CPF) for describing archive creators. It records authority control information and the description of relations between the subject and other entities. [Mazzini and Ricci, 2011]
OAD Ontology[14]	The OAD Ontology formalizes terms of ISAD(G) archival standard. It represents hierarchical levels of description of archival materials, and integrates the description of authority records by reusing EAC-CPF Ontology.
Sistema Archivistico Nazionale (SAN) Ontology[15]	The SAN Ontology is the ontology developed by the Italian Central Institute for Archives (ICAR) for representing the catalogue of archival resources (CAT). It includes the description of keepers, archive creators, fonds, and their subdivisions. It is aligned to EAC-CPF and OAD Ontology.
Museums Ontologies	

[12] http://sparontologies.net

[13] http://labs.regesta.com/progettoReload/lontologia-eac-cpf/

[14] http://labs.regesta.com/progettoReload/oad-ontology/

[15] http://www.maas.ccr.it/SAN-LOD/lode/

CIDOC-Conceptual Reference Model[16]	CIDOC CRM contains classes and logical groups of properties for describing cultural objects and aspects characterising the Cultural Heritage domain, including participation, parthood and structure, location, assessment and identification, purpose, motivation, use, and so on. Properties have put temporal entities and, with it, events in a central place. It models the explicit modeling of events, and the participation of objects and people [Doerr et al., 2003].
FRBRoo	FRBRoo model is closely related to IFLA's FRBR family of conceptual models. It is the object-oriented version of these models. FRBRoo version 1 was based on FRBR alone, while FRBRoo version 2 is based on three models, namely: FRBR, FRAD and FRSAD. It makes explicit some concepts that were implicit in the original models, thus making it easier to move from conceptual model to real-life applications. [Doerr et al., 2008]
Cross-domain and Task Ontologies	
Getty thesauri: Art and Architecture Thesaurus (AAT), Union List of Artists Names (ULAN) and Getty Thesaurus of Geographic Names (TGN)[17]	Inter-linked vocabularies mainly devoted to represent terms used in visual arts. Vocabularies share the same data structure, are multilingual, and identify respectively features of cultural objects (AAT), people (ULAN), and places (TGN) [Harpring, 2013].

[16] http://www.cidoc-crm.org/

[17] http://www.getty.edu/research/tools/vocabularies/lod/index.html

Europeana Data Model (EDM)[18]	EDM is the model adopted by Europeana partners. It incorporates community standards such as LIDO for museum, EAD for archives and METS for digital libraries. The Europeana approach, on the one side ensures great consistency and interoperability between providers. Unfortunately, on the other side it may lose the richness of the original data [De Boer et al., 2012].
PROV Ontology[19]	The PROV Ontology is developed by the World Wide Web Consortium (W3C). It is defined as a general, high-level standard for provenance, whereas provenance descriptions are defined for preservation of information resources, referring to representation, interchange, query, access, and validation of provenance [Moreau et al., 2015].
FOAF[20]	FOAF models data about authors of resources and relations between people and digital resources.
Ontology Design Patterns	ODPs are small (or cleverly modularized) ontologies with explicit documentation of design rationales and reengineering practices. Under the assumption that there exist classes of problems that can be solved by applying common solutions (as it has been experienced in software engineering), ODPs propose to support reusability on the design side specifically [Gangemi and Presutti, 2009].

Photography and Arts related projects. The American Art Collaborative (AAC) adopts CIDOC-CRM to map partners' data to RDF [Knoblock et al., 2017]. Likewise, the British Museum published its collection according to the CIDOC-CRM. The Rijksmuseum and many others published their extensive art collection data using the Europeana Data Model [Dijkshoorn et al., 2018]. Projects in the library domain, such as Linked Data for Production (LD4P) that includes Columbia, Cornell, Harvard, Library of Congress, Princeton, and Stanford [Schreur, 2018], use BIBFRAME to describe, among the others, photographic resources.

[18] `https://pro.europeana.eu/resources/standardization-tools/edm-documentation`

[19] `https://www.w3.org/TR/prov-o/`

[20] `http://xmlns.com/foaf/spec/`

Likewise, photo archives set up international projects aiming to make interoperable their heritage. *Europeana photography* is a EU-funded digitization project aimed at enriching Europeana with masterpieces of early photography. The project developed the *Europeana Photography multilingual vocabulary*, which includes about 561 concepts for describing photographic techniques, photographic practices and keywords [Van Steen, 2014]. Data provided by members of the consortium *PHOTOCONSORTIUM* [Fresa, 2014] is transformed in RDF according to the Europeana Data Model (EDM). The *ArCo project*[21] is an ongoing Italian national project for transforming cataloguing records gathered by the *Sistema Informativo Generale del Catalogo (SIGECweb)* into RDF according to a bespoke set of ontologies. This includes data belonging to photo archives and art historical photo archives.

Specifically devoted to art historical photo archives is the *PHAROS consortium* [Reist et al., 2015], which includes fourteen European and North American art historical photo archives committed to creating a digital research platform. The aim is to enable a comprehensive consolidated access to photo archive images and their associated scholarly documentation. The chosen ontology for mapping data sources is CIDOC-CRM. The project is ongoing, and the definition of a final version of the CIDOC-CRM application profile is in progress. Among the partners that have already published Linked Open Data there are: Yale Center for British Art [Delmas-Glass, 2016], Bildarchiv Foto Marburg (which published its collections through Europeana according to EDM), Villa I Tatti [Klic et al., 2018], and the Federico Zeri Foundation [Daquino et al., 2017].

In conclusion, we can summarise aspects related to the usage of ontologies for representing cultural heritage metadata as follows:

- The multifaceted nature of cultural objects can be addressed by means of the FRBR four-layers structure, which facilitates the integration of heterogeneous information. The FRBR Ontology and the FaBiO Ontology are good candidates to represent the *object-driven* description of the photographic object.

- CIDOC-CRM pursues an *event-driven* representation, that is, a representation of objects as participants to events, which allows to define the milestones of the "world-line" of a cultural object (i.e., its

[21] `http://wit.istc.cnr.it/arco`

social biography). To this extent, CIDOC-CRM is a good candidate for the description of museum objects, while FRBR is used to represent serial objects, including books, articles, and photographs.

- The SPAR Ontologies provide plenty of constructs for characterizing the publishing domain, such as defining citations and people's role in situations, which are not fully represented in CIDOC-CRM or FRBR-alike models.

- The PROV-O Ontology is the W3C recommended standard for describing data provenance, which includes concepts of activities producing and consuming data.

Such ontologies can be reused as they areto describe most of the peculiarities of art historical photo archives. However, an important gap emerged from the survey of ontologies and projects, that is: domain and cross-domain ontologies do not take into account hermeneutic aspects and the representation of questionable information. These include information about sources, motivations, and criteria adopted when supporting a subjective claim. Unfortunately, institutions do not always share their methodology and an *interpretation-driven* approach is lacking in the aforementioned models. Nonetheless, several art historical photo archives do include such information, and it is of high value for scholars. We argue that a task ontology is necessary to fill the gap in knowledge representation so as to describe the degree of subjectivity of statements and enable systems for semiautomatic validation of their reliability.

2.3 Information quality and authoritativeness assessment

In this section we survey relevant works on Information Quality in the web, outline information quality dimensions, and discuss related measures and assessment methods. We conclude with some consideration on how art historical data quality can be assessed. The aim of the literature review is to address which dimensions apply to authorship attributions published in catalogues of art historical photo archives as Linked Open Data, and envision ways to assess their authoritativeness.

Dimensions of IQ. Information Quality (IQ)[22] is the fitness for purpose of information. It encompasses both domain-dependent and domain-independent dimensions and it can be assessed by means of quali-

[22]Information Quality and Data Quality are here interchangeable, since also in the literature these are not always clearly distinguished.

tative and quantitative measures. In the Web making judgments of IQ is a difficult task since there is no quality control mechanism. The factors influencing judgment of quality are several, and research fields face differently such factors.

In Library and Information Science, scholars and librarians developed guidelines and checklists and mainly focused on functional aspects of metadata [Cooke, 1999] [Park, 2009]. However, supporting stakeholders in assessing reliability of questionable information is not taken into account in existing frameworks. Indeed, so far methods for modelling and reasoning on argumentation [Walton, 2013] and reliability of statements, have not been considered neither in cataloguing practices, nor in the Arts field. Computer Scientists developed a number of frameworks and methodologies for data quality assessment that also take into account content quality [Lee et al., 2002, Batini et al., 2009]. Knight and Burn [Knight and Burn, 2005] reviewed the most common dimensions available in a number of IQ frameworks, namely:

- Accuracy. The extent to which data are correct, reliable and certified free of error.
- Consistency. The extent to which information is presented in the same format and compatible with previous data.
- Security. The extent to which access to information is restricted to maintain its security.
- Timeliness. The extent to which the information is sufficiently up-to-date.
- Completeness. The extent to which information is not missing and is of sufficient breadth.
- Concise. The extent to which information is compactly represented without being overwhelming.
- Reliability. The extent to which information is correct and reliable.
- Accessibility. The extent to which information is available, or easily and quickly retrievable.
- Availability. The extent to which information is physically accessible.
- Objectivity. The extent to which information is unbiased, unprejudiced and impartial.
- Relevancy. The extent to which information is helpful for the task at hand.
- Usability. The extent to which information is clear and easily used.

- Understandability. The extent to which data are clear and easily comprehended.
- Amount of data. The extent to which the quantity of available data is appropriate.
- Believability. The extent to which information is regarded as true and credible.
- Navigation. The extent to which data are easily found and linked.
- Reputation. The extent to which information is highly regarded in terms of source or content.
- Useful. The extent to which information is applicable and helpful for the task at hand.
- Efficiency. The extent to which data quickly meet the information needs.
- Value-Added. The extent to which information provides advantages from its usage.

Rieh [Rieh, 2002] focused on the way users make judgements on IQ. According to Rieh, the judgement of information quality can be identified in terms of (1) characteristics of information objects, (2) characteristics of sources, (3) ranking in search output, and (4) general assumption. Specifically, the judgement involves the extent to which users think that the information is useful, good, current, and accurate, and the extent to which users think that they can trust the information. We can distinguish the former group of features as the *textual authority* of an information source and the latter group as the *cognitive authority* of the source [Wilson, 1983]. Rieh [Rieh, 2002] defines five dimensions characterising textual authority, namely: goodness, accuracy, currency, usefulness, and importance. Likewise, six facets define cognitive authority, namely: trustworthiness, reliability, scholarliness, credibility, officialness, and authoritativeness.

Likewise, according to Farahat et al. [Farahat et al., 2007] authoritativeness in information retrieval can be interpreted in two ways. The first idea relies on a graph-theoretical notion, and is grounded in social networks. For instance, in the sentence "An authoritative source states that *La Schiavona* is a Titian's painting", the term "authoritative" can be interpreted in terms of cognitive authority, that is, "authoritative" is the source relatively close to the artwork, such as the museum or the scholar that ascribed the artwork to an artist first, and that has authority on the matter. Such an idea of authoritativeness is at the basis of citation indexes - where an "authoritative" source is relatively central in the network of citations in scholarly literature - and link-analysis approaches implemented by search engines - where "authoritative" pages are

generally those that are linked to a high number of other pages. A second concept of authoritativeness is broadly defined as "textual". For example, the statement "The Zeri photo archive issued an authoritative cataloguing record on the painting *La Schiavona*" does not necessarily imply that the photo archive had any close relation to the scholars who had first-hand knowledge of the artwork, or for that matter that scholars are generally disposed to cite the Zeri cataloguing record, although that may very well be the case. Rather, the record is authoritative on internal grounds, i.e. the cataloguing record reads as if it is well-researched, documented, and contains numerous references which contribute to validate its reliability.

Measures and IQ assessment methods. Naumann and Rolker [Naumann and Rolker, 2005] defined a set of IQ dimensions and a three-fold assessment approach. In particular, dimensions are grouped in three assessment classes, namely: (1) the subject, i.e., the user (2) the object, i.e., the source, and (3) the information retrieval process. Table 2.2 lists dimensions and assessment methods grouped according to three classes.

Table 2.2: Classification of IQ dimensions and metrics

Assessment Class	IQ Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevance	Continuous user assessment
	Reputation	User experience
	Understandability	User sampling
	Value-added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer support	Parsing, contract
	Documentation	Parsing
	Objectivity	Expert input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

Subject criteria describe general aspects deemed relevant by the user when judging a source of information.

Such criteria highly depend on the user perception, hence user studies are the main tools for evaluating the goodness of a source of information. Object criteria refer to the intrinsic and extrinsic features of the document, which can be generally addressed by relying on automatic methods and domain experts' consultancy. Process criteria regard the aspects that affect the user's judgment when comparing a number of sources and choosing the most authoritative one. The latter can be suppoerted by automatic methods iteratively improved and continuously assessed.

The definition of a framework for IQ assessment is strictly related to trust. The Semantic Web aims at creating a trusted network of datasets and data providers so as to let applications automatically deduce and recommend the best candidate information. However, adopting existing approaches for assessing IQ in Linked Data is not straightforward [Zaveri et al., 2016]. When assessing trustworthiness of datasets, provenance is a crucial aspect [Lei et al., 2007]. Bizer and Cyganiak [Bizer and Cyganiak, 2009] classified Linked Data quality dimensions into three categories: (1) Content Based, i.e., the information content itself, (2) Context Based, i.e., information about the context in which information was claimed, (3) Rating Based, i.e., based on the ratings about the data or the rating related to the information provider.

Zaveri et al. [Zaveri et al., 2016] classify dimensions of Linked Data Quality (LDQ) into four groups, namely: (1) Accessibility, (2) Intrinsic, (3) Contextual, and (4) Representational group. They define a comprehensive set of 18 dimensions and 69 related metrics for the assessment, further classified in quantitative or qualitative measures. Among such dimensions, trustworthiness is defined "as the degree to which the information is accepted to be correct, true, real and credible". Authors suggest several methods to assess it, including (1) scores associated to absolute beliefs and disbeliefs, (2) opinion-based methods applied asking users to annotate data, (3) trust annotations in Semantic web-based social networks. Reasoning can be used to encode blacklists of harmful datasets, or to define authorities as sources that adopt consistently Linked Data principles. Trust ontologies can be applied to unknown data, using either content-based methods or metadata-based methods (e.g. reputation assignment, user rating, and provenance).

According to Zaveri et al. [Zaveri et al., 2016] trustworthiness of information can be based on the association between the author and the dataset, that transfers trust from content to resources. Such an assumption is particularly relevant since we want to demonstrate to what extent cultural heritage institutions are authoritative information providers. The assessment of information providers' trustworthiness

could be defined by (1) constructing decision networks informed by provenance graphs, (2) by relying on lists of trusted providers (which can be further annotated with a level of trust), or by (3) applying trust ratings assigned by users.

According to Farahat et al. [Farahat et al., 2007] authoritativeness of a document can be estimated by combining textual, non-topical cues and link analysis. Results of their work demonstrated the importance of textual authority combined to social authority of a web document - that is the networked structure obtained by using algorithms such as Google PageRank [Brin and Page, 1998] and Hyperlink-Induced Topic Search (HITS) algorithms [Kleinberg, 1998] - when ranking search results.

IQ assessment in art historical photo archives. No specific studies or methodologies are available for assessing art historical data quality. Cataloguing rules do not provide precise guidelines on how to rate subjective, questionable attributions. Only CDWA online guidelines[23] dedicate a section to the choice of the most authoritative sources (emphasis added).

> **Disagreement among sources** Know your sources. When two sources disagree, prefer the information obtained from *the most scholarly, authoritative, recent source*.
>
> [...] Sources It is critical for the cataloger to cite sources of information. In order for the information to be considered reliable, it must be derived from authoritative sources. Online sites to which any member of the public may contribute are not considered reliable. In general, authoritative sources are compiled or researched by verified, known scholars and experts, and published (online or in hardcopy) by reliable authoritative publishers. Scholarly catalogs, text books, monographs, encyclopedia, dictionaries, and journal articles authored by an expert are reliable sources. A scholar's spoken opinion or email may be a source, if the person is a known expert on the topic (such sources must also be cited). Information may be derived from unpublished documents such as inventories, letters, bills of sale, photo mounts, and inscriptions on the work itself, if proven to be authentic by experts. Repository records are considered the preferred reliable source of information about a given object; if such records are reflected on the museum's Web site, the site may be considered authoritative. Specific reliable sources are listed elsewhere in CDWA, in context for various subcategories.

[23] `http://www.getty.edu/research/publications/electronic_publications/cdwa/introduction.html`

Despite such a statement in natural language is not sufficient to define a shared methodology for IQ assessment, we assume a set of IQ criteria may be derived from the comparison of definitions provided by Naumann and Rolker [Naumann and Rolker, 2005] and such a statement. In detail:

- *The most scholarly source.* According to the guidelines, provenance is the first criterion in the evaluation of the trustworthiness of a resource. The reputation of the information provider allows reliability to be transferred to the information source. Published sources imply a peer-review process had been undertaken, hence the source is supposed to be verifiable. Oral sources are evaluated only on the basis of the reputation of the provider. Unpublished sources that are produced by cultural institutions, are deemed relevant because of the trustworthiness of the cataloguing process that originated them. In summary, being a scholarly source implies relying on criteria like relevance, reputation, and reliability. However, how to assess providers' authoritativeness is not clear, especially when contradictory and equally authoritative sources are available.

- *The most authoritative source.* Authoritativeness is not included in the dimensions surveyed in [Naumann and Rolker, 2005]. However, Rieh [Rieh, 2002] includes authoritativeness in the dimensions for describing cognitive authority, i.e., trustworthiness, reliability, scholarliness, credibility, officialness, and authoritativeness.

- *The most recent source.* It refers to the timeliness of the cited source. This ensures that an accurate literature review has been performed. However, it is not clear to what extent this dimension affects the choice of the source.

In conclusion, we can summarise contributions and assumptions illustrated in this chapter as follows:

- The survey of ontologies shows that existing ontologies naturally cover a number of aspects of the Photography and Arts domain. Therefore they can be reused in this research to represent a significant amount of metadata produced by art historical photo archives. However, none of the existing ontologies covers aspects related to the hermeneutic approach and the representation of questionable information that can result from connoisseurship activities and the cataloguing process.

- Existing IQ measures cover a number of features that apply to the Cultural Heritage domain. Measures and metrics that apply to the information retrieval process related to art historical data can be

reused. However, there are no studies and IQ frameworks dedicated to the assessment of questionable information produced during the cataloguing process in art historical photo archives, such as authorship attributions. The empirical study shows that bespoke IQ measures to assess textual and cognitive authoritativeness of sources must be created so as to take into account alle the features that characterize the domain. According to our preliminary study, we assume authoritativeness can be addressed by comparison in a data integration process, evaluating dimensions such as reputation, reliability, relevance, and timeliness.

- Semantic Web technologies are deemed suitable to accomplish a number of tasks that are common in the Cultural Heritage domain, such as knowledge organization, data integration, and aggregation. Such technologies can be used for representing in a machine-readable format the knowledge produced by photo archives and explore benefits of its usage when aggregating, analysing, and comparing data sources.

Part II

SEMANTIC WEB APPLICATIONS FOR CONNOISSEURSHIP

Chapter 3

Methodology and Approach to the Research

In this chapter the research design and the procedures used in conducting the study are presented. Leveraging Linked Open Data of art historical photo archives in connoisseurship activities is a twofold issue. Knowledge representation aspects and information quality issues must be tackled in order to efficiently support the decision-making process of users evaluating authorship attributions.

The description of the art historical ecosystem presented in Chapter I, the analysis of the state of art in ontology development, and the literature on Information Quality assessment presented in Chapter II showed a number of gaps in the state of art. In the following sections we outline our research questions, hypotheses, assumptions, and limits of the research. Lastly, the research methodology and approach to the research are described.

3.1 Research problems

This work aims at providing theoretical foundations and technical solutions for publishing and assessing authoritativeness of authorship attributions recorded in secondary sources by means of Semantic Web technologies. According to the state of art presented in Chapter I and Chapter II, a number of challenges can be identified.

Challenges can be summarised in three research problems, described as follows:

- **RP1.** The formal representation of questionable information in the Photography and Arts domains by leveraging well-grounded formal languages and technologies.
- **RP2.** The formalisation of the dimensions characterising the methodology of art historical data providers when publishing questionable information.
- **RP3.** Support users' decision-making process when assessing reliability of authorship attributions.

For the first research problem (**RP1**), i.e. the formal representation of questionable information in the Photography and Arts domains, a number of research questions are tackled in this thesis, namely:

- **How can we represent cultural heritage data belonging to art historical photo archives?** The aim is to map existing models and develop bespoke ontologies so as to facilitate the data integration process and address all the information required in connoisseurship related activities.
- **How can we represent the interpretative process that underpins the creation of questionable information, such as authorship attributions?** The aim is to develop a model for representing provenance of information and all the aspects relevant to assess the veracity of a statement.

For the second research problem (**RP2**), i.e. the analysis of the methodology of art historical data providers, the following questions are tackled:

- **What are the criteria characterizing the methodology of art historical data providers when reviewing authorship attributions?** The aim is to identify and rate criteria that are deemed relevant by art historical photo archives when choosing an authorship attribution among the others. The set of identified criteria contribute to build a conceptual framework characterizing the methodology in art historical research.
- **To what extent we can address and measure, either qualitatively or quantitatively, the dimensions characterizing textual authoritativeness?** The aim is to (1) identify internal grounds of an information source that affect the decision-making process when validating an authorship attribution, (2) describe how such dimensions interact as part of a conceptual framework, and (3) define

methods for measuring authoritativeness. The objective is to support users' judgment and suggest policies for data quality improvement in art historical photo archives.

- **To what extent we can address and measure features characterizing cognitive authoritativeness?** The aim is to address, though in a very early stage, which metrics fit for the purpose of measuring cited scholars' cognitive authoritativeness. The objective is to provide additional information to users when reviewing scholars' contradictory attributions.

For the third research problem (**RP3**), i.e. the assessment of reliability of statements in connoisseurship related activities, the following research questions are addressed:

- **Can we assess textual authoritativeness of authorship attributions by leveraging Linked Open Data and Semantic Web technologies?** The aim is to define an ontology-based ranking model to sort results of a research performed against several art historical Linked datasets having different degrees of data quality. The objective is to measure users' satisfaction with regard to the ranking model, including different types of users and a number of usability measures.
- **Can technical solutions be developed to support users' decision-making process when choosing between competing authorship attributions?** By building a system that implements models, and methods developed in this thesis as a proof of concept, the aim is to highlight well-documented authorship attributions and suggest the most authoritative one first.

3.2 Hypotheses and assumptions

The following hypotheses follow from the research questions presented in the previous section.

3.2.1 Hypotheses

For the first research problem (**RP1**), that concerns the formal representation of questionable information in the Photography and Arts domains, the following hypotheses can be identified.

- **H1. We can reuse existing ontologies for representing information included in cataloguing records produced by art historical photo archives.** Golden standards and well-known ontologies in the Cultural Heritage domain provide terms for describing the heterogeneous Photography and Arts domain.

- **H2. The interpretative process that generates questionable information can be effectively represented by using Semantic Web technologies, such as ontologies.** By extending existing ontologies with new terms it is possible to describe features characterising textual authoritativeness of questionable information, such as authorship attributions.

For the second research problem (**RP2**), related to the definition of dimensions characterizing the methodology of art historical providers and the dimensions of textual authoritativeness, the evaluation of the following hypothesis is here proposed.

- **H3. Analytical data and domain experts' feedback can be used to formalize the criteria underpinning the methodology of art historical data providers when publishing authorship attributions.** (1) The analysis of cataloguing rules, (2) their implementation in photo archives catalogues, and (3) a comparative analysis of archival policies when choosing a favourite authorship attribution, allow to define a rating of criteria that concur to assess textual authoritativeness of sources recording authorship attributions.

- **H4. The evaluation of textual authoritativeness of sources recording authorship attributions can be based on a documentary, evidence-based approach.** Authoritativeness of secondary sources (such as cataloguing records) differs from the veracity of primary sources (e.g. historians' expertises). The former can be addressed by means of a set of dimensions grouped into a conceptual framework, that is shared among providers. Such dimensions contribute to define features of IQ users' judgement and archival policies for data quality improvement in art historical photo archives.

- **H5. Measuring scholars' authoritativeness in the arts field can be achieved by developing bespoke metrics.** Bespoke metrics measuring the perception of scholars' authoritativeness in a community can be extracted from a set of relevant sources at hand. Users' perception of the importance

of having such metrics can be evaluated. However, such metrics are in a very early stage, due to the lack of extensive citation indexes in the Humanities, and are not meat to affect the ranking model.

Finally, for the third research problem (**RP3**), i.e. the assessment of reliability of statements in connoisseurship related activities, the following hypotheses are tested.

- **H6. Linked Open Data and Semantic Web technologies can support and satisfy retrieval and aggregation requirements of research activities in the Arts and Humanities.** Common activities include the retrieval and the aggregation of relevant sources of information recording authorship attributions. LOD can be leveraged in user-centric application that take into account features of the art historical domain and return tailored solutions to their needs.
- **H7. Automatic and curated methods can support the decision-making process in connoisseurship activities.** By leveraging the aforementioned conceptual framework in a ranking model, it is possible to recommend well-documented secondary sources recording authorship attributions.

3.2.2 Assumptions

The aforementioned hypotheses are evaluated considering the following assumptions.

- **A1. Art historical data are served as secondary sources.** We rely on peer-reviewed secondary sources that have already assessed the goodness of primary sources, e.g. scholars' opinions, bibliography.
- **A2. Art historical photo archives include detailed information on their methodology, and are therefore the subject of this study.** Sources such as museum records do not include detailed information on the hermeneutic approach and, at this stage, do not contribute to address features of textual authoritativeness. Therefore the latter are excluded from the data analysis.
- **A3. Art historical data is provided as RDF data according to one or more existing ontologies.** To leverage Semantic Web technologies we either rely on existing RDF datasets or we transform data sources into RDF. Schemas of existing RDF datasets, although heterogeneous, can be consistently mapped and allow a comparison between data sources.

- **A4. Data providers' authoritativeness can be deduced by relying on third party opinions.** A list of trusted data providers in the Arts field, labelled as domain experts or not, is sufficient for assessing their reputation.

- **A5. Measuring cognitive authoritativeness in the Arts field can be achieved on an artist basis.** Since citation data lack for art historians, we can define bespoke metrics for scholars that worked on a specific set of artists. So doing we can create flexible, relative metrics for defining cognitive authoritativeness.

3.3 Methodology

This research is based on the *design-science method* proposed by [Hevner et al., 2004]. The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts. Artifacts are broadly defined as *constructs* (vocabularies and symbols), *models* (abstractions and representations), *methods* (algorithms and practices), and *instantiations* (implemented and prototype systems).

Constructs provide the language in which problems and solutions are defined and communicated. Models use constructs to represent a real world situation, i.e. a specific domain. Methods, i.e. algorithms and best practices, define processes and provide guidance on how to solve problems in the specific domain. Instantiations show that constructs, models, or methods can be implemented in a working system. They demonstrate feasibility, enabling concrete assessment of an artifact suitability to its intended purpose. They also enable researchers to learn about the real world, how the artifact affects it, and how users appropriate it.

In this research, an artefact for harvesting, consuming, and comparing art historical data is developed by leveraging Semantic Web technologies. The artefact is called *mAuth - mining authoritativeness in art history*. It encompasses a set of constructs, models, methods, and instantiations, that are designed and implemented to evaluate the hypotheses addressed in the previous section. In particular, it includes the following components:

- Constructs and models are designed by developing ontologies so as to represent respectively (1) metadata produced by art historical photo archives and (2) questionable information.
- Methods are developed to (1) harvest information sources recording authorship attributions, and to (2) rank results of the research.
- The instantiation is composed of few software components that correspond to the different methods, constructs, and models.

3.4 Research objectives and contributions

Based on the aforementioned hypotheses and research methodology, this research is organised so as to pursue two research objectives (**RO**) and one technological objective (**TO**). A number of outcomes are associated to the description of the objectives.

- **RO1. Define ontologies for representing the Photography and Arts domain, with a particular focus on questionable information.** An in-depth analysis on how knowledge sediments in art historical photo archives is performed by means of (1) the mapping to RDF of information addressed in ICCD-OA and ICCD-F cataloguing rules, and (2) the transformation of the Zeri photo archive catalogue into a RDF dataset. Such actions contribute to validate H1 and H2. **Outcomes**: mapping documents of ICCD-OA and ICCD-F to RDF; OA-Entry Ontology; F-Entry Ontology; HiCO Ontology; the Zeri Photo Archive RDF dataset.
- **RO2. Define methods to assess the methodology undertaken by art historical photo archives, and textual authoritativeness of sources recording authorship attributions.** The specific objective is to design a conceptual framework that addresses dimensions of textual authoritativeness of information sources used in connoisseurship activities. The framework includes: (1) a rating of criteria and motivations supporting the choice of an authorship attribution; (2) a number of dimensions and measures for evaluating textual authoritativeness of sources recording attributions; (3) bespoke metrics for assessing cognitive authoritativeness of scholars cited as source of information. Such a framework is meant to validate H3, H4, and H5. **Outcomes**: a rating of criteria; a

conceptual framework of Information Quality measures; mAuth ranking model; policies for data quality improvement.

- **TO1. Develop a system that implements the conceptual framework and supports the users' decision-making process.** The objective is to develop a semantic crawler that harvests art historical Linked datasets, stores information in a bespoke knowledge base, implements an ontology-based ranking model, and serves sorted lists of authorship attributions to both applications and users. The outcomes are meant to support the validation of H6 and H7. **Outcomes**: mAuth knowledge base; mAuth framework, including a semantic crawler, an API, and a Web application.

3.5 Approach to the research

The research approach is designed as a three-step process, corresponding to the three main challenges, namely: knowledge representation, information quality assessment, and decision-making support. Figure 3.1 illustrates the main actions undertaken during the three phases and the resulting main contributions of this research project.

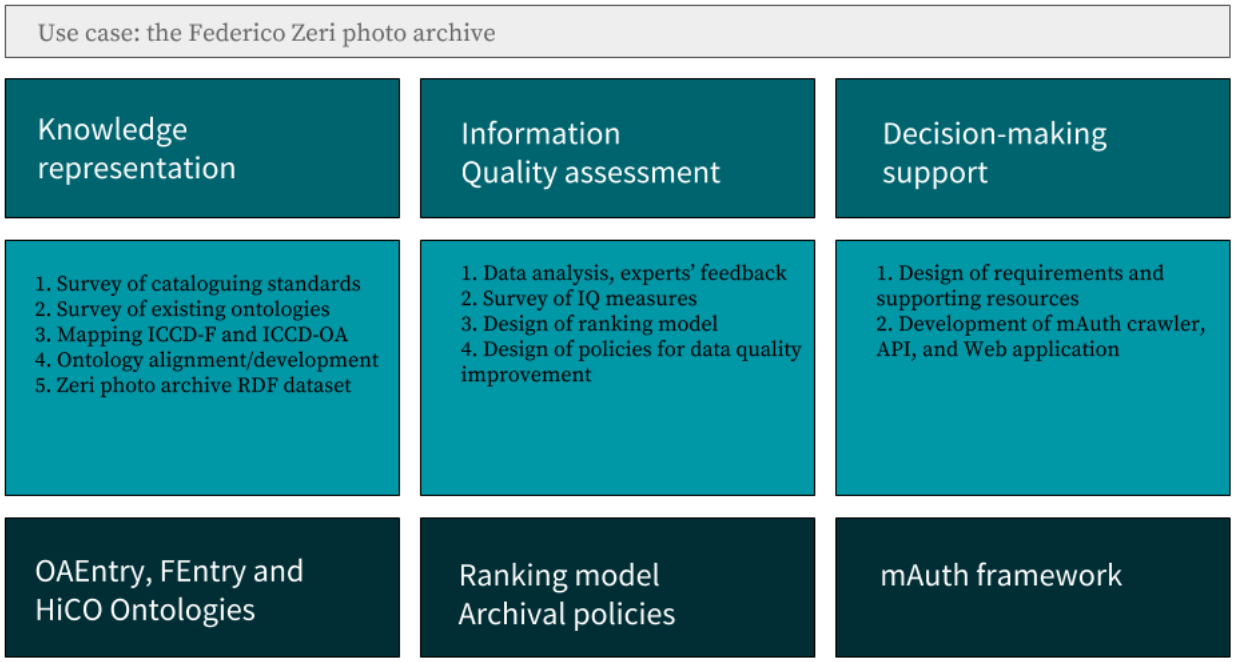


Figure 3.1: Approach to the research

The first research problem concerns knowledge representation issues in the art historical research field. To tackle this problem the following actions were undertaken.

- **Survey of cataloguing standards.** To have an overall view of the domain, the main cataloguing standards adopted in libraries, museums, and archives were surveyed. Results of the survey are outlined in Chapter I.
- **Mapping of ICCD-F and ICCD-OA to RDF.** The Italian cataloguing standards ICCD-F and ICCD-OA resulted the best candidate for the mapping to RDF of the entities relations characterizing the art historical research field.
- **Survey of ontologies in the Cultural Heritage domain.** The survey is presented in Chapter II. Results highlighted CIDOC-CRM as the golden standard for describing artworks, the SPAR Ontologies for representing photographs and bibliographic resources, and the PROV Ontology for representing aspects related to the subjectivity of statements. Such ontologies are the target of the mapping of ICCD-F and ICCD-OA to RDF.
- **Ontology alignment and development.** Existing predicates are collected from the aforementioned ontologies, and missing classes and properties are created. All terms are gathered in bespoke ontologies. The OAEntry Ontology for describing artworks is created, and the existing FEntry Ontology for describing photographs is revised and extended. The HiCO Ontology is developed for representing questionable information and the hermeneutic process. The ontologies are designed by adopting a data-centric methodology for ontology development.
- **RDF transformation of the Zeri photo archive catalogue.** We transformed a relevant subset of the Zeri photo archive catalogue into RDF so as to assess consistency and comprehensiveness of the developed models.

The second research problem regards the assessment of the methodology of photo archives and the formal definition of textual authoritativeness of sources of attribution. The problems tackled are described below.

- **Data analysis and domain experts' feedback on criteria used in art historical photo archives.** A survey on existing photo archives cataloguing data is performed to extract dimensions characteriz-

ing their methodology. Three archives are selected among PHAROS members, which present (1) a homogeneity of topics addressed in their data, and (2) similar cataloguing standards and resulting datasets. The result of the analysis is a controlled vocabulary of terms used for classifying the motivation underpinning a questionable statement (e.g. "bibliography", "art historian's attribution", "museum attribution"). Secondly, we define a rating of these according to their degree of reliability. The rating is obtained by means of the quantitative analysis performed over cataloguing records of the aforementioned archives and by consulting domain experts for double-check.

- **Survey of Information Quality measures to address features of textual authoritativeness and cognitive authoritativeness.** Existing metrics were surveyed (outlined in Chapter II) and pruned so as to define a representative subset of measures that apply to art historical data. The subset is defined by taking into account best practises in the art historical community, e.g. the aforementioned guidelines provided by the Getty Research Institute, and aspects highlighted from data analysis. Metrics for defining cognitive authoritativeness are developed by tuning existing indexes. The rating of criteria and the other measures are gathered into a conceptual framework.

- **Design of a ranking model.** The rating of criteria and the measures are interconnected and weighted as part of a ranking model for recommending authoritative authorship attributions.

- **Design of policies for data quality improvement in art historical photo archives.** The data analysis, the rating of criteria, and the measures addressed in prior steps contribute to define a number of policies for data quality improvement in art historical photo archives. A low-cost process for data integration that leverages findings of this research is also designed.

The third research problem concerns the application of the aforementioned findings into a recommender system capable to support users' decision-making process. In particular, the problems to be tackled are the following.

- **Design requirements and develop resources for leveraging the potential of Linked Open Data.** A toolkit of resources was realized for supporting the mAuth framework. In particular, a number of linksets, i.e. datasets including only links about equivalences between individuals belonging to different datasets, was developed to speed up the harvesting. Linksets include links between same

artworks, same artists, same organizations, and same art historians cited as sources of attributions. Secondly, a dataset describing art historians, their attributions, and statistics on the acceptance of their attributions is created.

- **Development of a framework for harvesting, ranking and consuming data on authorship attributions.** mAuth is developed as a Python framework. It works as an aggregator of attributions harvested from a number of linked datasets and serves structured information to both applications and users, by means of an API and a Web interface. An ontology-based ranking model is implemented for recommending the most documented attributions.

To conduct the research and cope with the aforementioned problems, a use case is set up. The notable Federico Zeri photo archive is a representative case study for tackling issues related to RP1 and RP2 and provides a golden standard for the publication of art historical data.

All the phases of the research are detailed in the following chapters.

- Chapter IV. Describes the design of a use case for leading ontology development, data analysis and the definition of a golden standard for art historical data.
- Chapter V. Presents phases of the development of ontologies for representing the Photography and Arts domain and the subjectivity of their information.
- Chapter VI. Illustrates how we obtained a conceptual framework for defining textual authoritativeness and cognitive authoritativeness.
- Chapter VII. Provides an overview of the design of a proof-of-concept recommender system.
- Chapter VIII. Presents the evaluation of artefacts, including ontologies, conceptual framework and mAuth framework.

Chapter 4

The Federico Zeri's Photo Archive Use Case

In this chapter the Federico Zeri photo archive and the Zeri & LODE use case are introduced. The Federico Zeri photo archive offers a comprehensive application profile of the Italian cataloguing rules ICCD-OA and ICCD-F and it includes plenty of information related to connoisseurship activities. The transformation of the Zeri's catalogue into Linked Open Data contributes to validate our initial hypotheses H1 (*We can reuse existing ontologies for representing information included in cataloguing records produced by art historical photo archives*) and H2 (*The interpretative process that generates questionable information can be effectively represented by using Semantic Web technologies, such as ontologies*). Domain experts from the photo archive provided feedbacks on the mapping and supported the analysis of data related to authorship attributions. In particular, they provided insights on the policies, the approach, the methodology, and the preferences of photo archivists. Such a preliminary analysis contributed to validate hypothesis H3 (*Analytical data and domain experts' feedback can be used to formalize the criteria underpinning the methodology of art historical data providers when publishing authorship attributions*).

4.1 The Federico Zeri's collections

The Federico Zeri photo archive was created by one of the most important art historians of the twentieth century, Federico Zeri (1921-1998). His extensive library of art books, auction catalogues and individual photos of monuments and artworks were the main tools of his work. He began to collect them in the

1940s and, over time, made them into "the world's largest private archive of Italian paintings", which became an essential reference work for the historical sequencing of out-of-context works. At the time of his death, the archive included more than 46.000 volumes, 37.000 auction catalogues, 60 periodicals and 290.000 individual photographs.

In order to preserve his bequest and put it to best use, the Zeri Foundation was set up in his name at the University of Bologna, and it has come to be recognized as one of the most important research and training centres for art historians in the world. Among its activities, the Zeri Foundation undertook the cataloguing of Zeri's repository, employing to that end two Italian metadata content standards, ICCD-F and ICCD-OA. The work of cataloguing Zeri's collection in compliance with the two Italian standards has resulted in the Zeri Photo Archive catalogue, which is accessible through a web interface.[1] In this thesis, the original names ICCD-F and ICCD-OA are used to refer to aforementioned content standards, while the English translation F Entry and OA Entry will be used to refer to metadata documents recorded compliant with the aforementioned content standards, and the resulting ontologies, named F Entry and OA Entry as well.

As already mentioned, the heritage preserved in Zeri's collections includes heterogeneous sources of information. In the art historical inquiry, relying on diverse types of resources help to tackle a problem from several perspectives, and to support questionable statements that came up during the cataloguing process, and may give authoritativeness to archivists' statements. The variety of resources held by the Zeri photo archive includes (i) photographs and attached documentation, annotated by plenty of scholars, (ii) books and journals part of the art library, (iii) photographic catalogues, and (iv) auction catalogues. Five main research areas can be explored at the Zeri photo archive, namely: art history, history of photography, history of restoration, connoisseurship, and history of collecting.

The adoption of the Zeri photo archive as leading use case for pursuing goals defined in this thesis is motivated by the following reasons.

- **Comprehensiveness of sources for art historical inquiry.** The archive includes both a photographic collection and an art library, highly interconnected with the archival and photographic materials. It presents the optimal scenario for the study of the Photography and Arts domains.

[1] `http://catalogo.fondazionezeri.unibo.it`

- **Application profile of ICCD-OA and ICCD-F.** Cataloguing records comply with ICCD-OA and ICCD-F standards. Specifically, around 225 fields are included. It provides an optimal use case for the mapping of the two standards to RDF.

- **Hermeneutic approach to the cataloguing.** Attributions are stated by archivists by either recording Federico Zeri's will or by replacing those with more recent attributions. The catalogue is enriched with information related to the history of attributions, including discarded attributions found by cataloguers. The methodology adopted by archivists when ascribing an artwork to an artist is well-documented and described in cataloguing data by means of controlled vocabularies. Hence, subjectivity of statements published by photo archives can be properly analysed.

- **Golden standard for art historical data publication.** According to aforementioned requirements, the publication of the Zeri photo archive as a Linked Open Dataset would provide a golden standard for the publication of data belonging to art historical photo archives.

4.2 The Zeri & LODE project

The objectives of the *Zeri & LODE* project are (1) the development of ontologies for representing information included in the catalogue of the Zeri Photo Archive and (2) the publication of data as Linked Open Data (LOD).

The Zeri & LODE project has been set in the context of a broader project, called *PHAROS: An International Consortium of Photo Archives*. PHAROS is one of the first steps towards the creation of a digital infrastructure of the notable photographic archives of works of art in Europe and the United States of America. The Consortium enables the active collaboration between institutions responsible for fourteen photo archives, so as to create a common platform for research on images and metadata of Western and non-Western works of art in all media. The Zeri Foundation is part of the consortium, and its photo archive is planned to be one of the first assets to be included in PHAROS. The PHAROS Consortium suggested Linked Open Data and the CIDOC-CRM conceptual reference model as the required representational framework for data exchange within the consortium.

In Chapter II, we mentioned the best practices for publishing Linked Data [Berners-Lee, 2006]. The W3C Government Linked Data Working Group created official guidelines for publishing and accessing open (government) data using the Linked Data principles.[2] Three existing methodologies are suggested, proposed by Hyland et al. [Hyland and Wood, 2011], Hausenblas et al. [Hausenblas and Cygankiak, 2016] and Villazón-Terrazas et al. [Villazón-Terrazas et al., 2011].

In the Zeri & LODE project we adopted a methodology based on the aforementioned ones. The approach consists of six steps corresponding to the following tasks: (1) Study of the domain, (2) Data modelling, (3) Data transformation, (4) Data reconciliation, (5) Data publication, and (6) Data exploration. Reusable components in the form of tools, vocabularies, datasets, and services, are adopted whenever applicable.

Study of the domain. In Chapter I we presented a study of the domain focused on cataloguing rules and content standards widely adopted by cultural institutions for describing artefacts, such as photographs, artworks, archival documents and bibliographic entities. These provide an overview of types of information characterising the domain. The ICCD-OA and ICCD-F cataloguing rules resulted the most complete standards for representing information related to connoisseurship activities, including information on the methodology adopted by cataloguers to record questionable information. We chose to rely on data represented by means of such standard so as to define the scope of the connoisseurship domain to be analysed.

Along with the analysis of standards, domain experts were consulted. A photo archivist with a background in art history employed at the Zeri Foundation provided (1) consultancy on the definition of competency questions so as to support the ontology development; (2) explanatory notes on ICCD-OA and ICCD-F cataloguing rules; (3) explanatory notes on the usage of ICCD-OA and ICCD-F standards at the Zeri photo archive, which may slightly differ for internal purposes; and (4) feedback on aspects related to knowledge organisation.

Lastly, we explored existing Linked Datasets and ontologies. In Chapter II are listed similar projects that deal with Photography and Arts domain in Cultural Heritage. However, at the beginning of this project no datasets belonging to art historical photo archives were available. The literature review shows that ontologies and vocabularies currently used do not fulfil requirements of connoisseurship inquiry.

[2] Best Practices for Publishing Linked Data. `http://www.w3.org/TR/ld-bp/`

Data modelling. To develop the ontologies for representing the Zeri photo archive, we adopted the *Simplified Agile Methodology for Ontology Development (SAMOD)* [Peroni, 2016], which leverages guidelines proposed in well-known ontology development methodologies, such as [Uschold and King, 1995] and [Fernández-López et al., 1997]. It allowed us to develop the ontologies by means of several small and iterative steps and to create documentation by using examples of data at the same time. In particular, this methodology required us to consider small issues defined by competency questions, and to test our under-development model immediately on existing data. We evaluated the logical consistency of the model by means of a reasoner, and we checked data consistency by testing the model on the Zeri dataset. Coherency of models is assessed by the domain expert who checked the correctness of vocabularies, thesauri, and relations. The documentation of the F Entry and OA Entry ontology development processes is available online.[3]

Good practices in ontology development were respected, so as to ensure semantic interoperability and to facilitate the reuse of ontologies. When applicable, classes and properties were either refactored or aligned to terms belonging to existing ontologies. In particular, several ontologies contributed to the ontology refactoring process. Each chosen ontology covers one aspect of the heterogeneous scenario.

CIDOC-CRM [Le Boeuf et al., 2015] was used to describe artworks and photographs. The SPAR ontologies [Peroni and Shotton, 2018b] were used to describe bibliographic entities, the cataloguing process, and some aspects of photographic documentation. In particular, FaBiO was chosen for managing bibliographies according to the FRBR conceptual model, CiTO was reused for describing citations that cataloguers included [Peroni and Shotton, 2012], and PRO was fundamental for documenting people's role with regard to photography, arts, publishing and cataloguing domains [Peroni et al., 2012]. Finally, the HiCO Ontology [Daquino and Tomasi, 2015], was created by extending PROV-O [Lebo et al., 2013], in order to describe hermeneutic aspects related to subjective attributions. Classes and predicates belonging to the aforementioned models that are effectively reused were gathered into two specular ontologies, i.e. the F Entry Ontology [Gonano et al., 2014] and the OA Entry Ontology [Daquino et al., 2017].

Data transformation. The data transformation to RDF can be performed in different ways, and with

[3]Documentation of models is available respectively at `http://www.essepuntato.it/2014/03/fentry/samod` for the F Entry Ontology and at `http://oaentryontology.sourceforge.net/samod/OAdevelopment.zip` for the OA Entry Ontology

various software tools. Assuming the cataloguing process of the subset of records in scope is over, and only few changes may occur over time, the transformation was performed as a one-time task.

We gathered about 31.000 F Entries and 19.000 OA Entries stored as XML documents (compliant with no particular schema) that contain metadata prescribed by the ICCD-F and ICCD-OA content standards. XML contents were organized in elements called paragraphs, corresponding to ICCD-F/OA descriptive sections. Data cleansing was needed in order to extract structured data from discursive text fields. The resulting RDF dataset was created by means of a XSL transformation. A RDF/XML file has been created for each of record. Several versions of the transformed dataset, that do not differ in terms of content, were necessary as backup and are currently stored in a long-term preservation repository [Daquino et al., 2016]. Metadata about the dataset itself is included, described by using VoID vocabulary [Alexander et al., 2009], so as to help users to determine the characteristics of the dataset.

Data reconciliation. Once transformed into RDF, data are interlinked with other data already published in external datasets. Links are created between datasets in order to accomplish two tasks, namely: (1) uniquely identify features characterising cultural objects, such as materials, supports, and techniques, by using terms belonging to thesauri and authority files acknowledged as golden standards; and (2) reconcile entities with those that are described in other datasets and are identified as being the same, such as people, places, and artworks. The aim is to allow data to be integrated in new user-centric applications and services for supporting connoisseurship and other research activities.

In details, several open vocabularies created by cataloguers in original data were aligned to existing controlled vocabularies. The latter include The Getty Art and Architecture Thesaurus[4] - for aligning terms referring to object types, materials, support, and techniques - and ICONCLASS[5] to uniquely identify subjects of the artworks. Moreover, VIAF,[6] the Getty Union List of Artist Names (ULAN),[7] Wikidata,[8] and DBpedia[9] were chosen to identify artists, photographers, art historians, and artworks; geoNames[10] and, again, Wikidata and Dbpedia, were considered for identifying places. Links to online available doc-

[4] http://www.getty.edu/research/tools/vocabularies/aat/
[5] http://www.iconclass.nl/home
[6] http://viaf.org
[7] http://www.getty.edu/research/tools/vocabularies/ulan
[8] https://www.wikidata.org
[9] http://dbpedia.org
[10] http://www.geonames.org

uments (not published as Linked Data) have been created for artists, photographers, and art historians to Wikipedia web pages; art historians are linked to the Dictionary of Art Historians[11] web pages; artworks are matched with the image collection of Villa I Tatti - Berenson Library (University of Harvard), Frick Art Reference Collections, DBpedia, Wikidata and VIAF.

The reconciliation was achieved by adopting several concurrent approaches with a different degree of precision. Such methods are listed below:

- Alignments to terms belonging to the Getty Art and Architecture Thesaurus are achieved by first manually translating terms to english and by querying the SPARQL endpoint of Getty Vocabularies.
- Subjects and corresponding ICONCLASS identifiers were previously listed by cataloguers in a spreadsheet. Since URIs minted by ICONCLASS do not adopt opaque URIS but include the aforementioned identifiers, the link to corresponding ICONCLASS URIs was created straightforward with a bespoke Python script.
- Links to VIAF and geoNames were created semiautomatically by using bespoke plugins implemented for Open Refine.[12]
- A PERL script[13] was useful for accessing the Wikidata Query Service, so as to directly link entities to Wikidata entities and also to Wikipedia pages, Dbpedia and Getty ULAN entities.
- Links to the Dictionary of Art Historians were obtained by means of web scraping techniques. A bespoke Python script extracts information from such web pages, looks for matches in VIAF, that are in turn linked to Zeri's entities, and the transitive link is created.
- Artworks are matched by means of a computer vision tool called Pastec[14] for image similarity matching. Only links between artworks that present a matching with a similarity score over a defined threshold (i.e. 30.0) are included.

Data publication. The Zeri photo archive RDF dataset is publicly available at `https://w3id.org/zericatalog`. Data to be published was selected on a thematic basis, including a collection of 19.000

[11] `http://arthistorians.info/`
[12] `http://openrefine.org/`
[13] `http://search.cpan.org/dist/App-wdq/lib/App/wdq.pm`
[14] `http://pastec.io/`

OA Entries describing artworks of the fifteenth and sixteenth centuries, and 31.000 F Entries describing photographs portraying the aforementioned works. RDF statements mainly refer to photographs and works of art and include information about 4,500 bibliographic entities, 6,000 artists, 2,000 photographers and 2,000 auction and photographers' catalogues. Such additional information was provided by the Zeri Foundation by means of other XML documents that comply with ICCD guidelines for creating people authority files[15] and bibliographic references authority files.[16]

All the RDF resources were labelled in Italian and, where possible, in English in order to facilitate their understanding for a larger audience. In addition, IRIs of these resources were created in English in order to ensure their easy reuse in other non-Italian datasets.

There is plenty of tools and software platforms which allow Linked Data publishing. We chose Apache Jena Fuseki2 triplestore[17] because it is easy to deploy and manage even for non-expert users. Data stored in the triplestore are distributed under the license CC-BY, Creative Commons Attribution 4.0 International.[18] To date, the dataset includes about 11,400,000 RDF statements linking 1,600,000 unique typed entities. Among these, about 3,000 are linked with external resources already available in the Linked Open Data. In particular, we created links to 2,200 different VIAF resources, 1,200 to Getty ULAN resources, 1,500 different GeoNames resources, and 2,260 different Dbpedia and Wikidata resources. RDF data can be queried in SPARQL by making appropriate REST requests to the related SPARQL endpoint made available by the triplestore at `https://w3id.org/zericatalog/sparql`. A web interface allows users to query the triplestore directly on the Web by means of SPARQL queries (available at `http://data.fondazionezeri.unibo.it/query/`).

Data exploration. We defined some use-case scenarios and we reused applications and services to showcase the (re)usability of the dataset by means of links to other Linked Data datasets. The use cases include text-based scenarios and specific SPARQL queries describing the ways in which the data can be browsed, retrieved, and used. A specific focus is given on how to leverage links to other Linked Data and reach information not available in the original data source, so as to extend its context. The aim is to present the potential of the linked dataset to future interested parties.

[15] `http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici/Standard/55`
[16] `http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici/Standard/58`
[17] `https://jena.apache.org/documentation/fuseki2`
[18] `http://creativecommons.org/licenses/by/4.0/`

We used the RDF browser LODView[19] to allow direct browsing of all the RDF data included in the triplestore in a user-friendly way. All F Entries and OA Entries defined in the dataset include links to the current catalogue entries of the Zeri Foundation (available at http://catalogo.fondazionezeri.unibo.it), enabling users to go from the LOD-based view of the traditional catalogue web pages. The inverse link from the catalogue pages to the related RDF resources has been implemented as well. Here pieces of information that were reconciled to external datasets, e.g. artists, are directly linked to VIAF, Wikidata, and DBpedia, allowing users to browse additional biographical resources.

[19] http://lodview.it/

Chapter 5

Knowledge Representation of Questionable Information in the Photography and Arts Domain

In this chapter aspects related to knowledge representation of the Photography and Arts domains are introduced. The ontology for describing questionable information is presented first, i.e. the HiCO ontology. The HiCO ontology contributes to validate our hypothesis H2 (*The interpretative process that generates questionable information can be effectively represented by using Semantic Web technologies, such as ontologies*). Secondly, two complementary ontologies, namely the F Entry Ontology and the OA Entry Ontology, are described to showcase what types of information characterize the domain and which may include questionable information. Lastly, the mapping between CIDOC-CRM, the two aforementioned ontologies and the ICCD-F/OA content standards is described. The mapping of ICCD-OA and ICCD-F cataloguing rules to RDF allowed us to validate our initial hypothesis H1 (*We can reuse existing ontologies for representing information included in cataloguing records produced by art historical photo archives*).

5.1 The Historical Context Ontology (HiCO)

The cataloguing of cultural objects is the result of a hermeneutical activity made by one or more cataloguers. Catalogue records can be seen as complex assertions on intrinsic and extrinsic aspects of objects they describe and that can be deemed questionable. Information may change over time because of new findings. The validity of pieces of information that may change over time is bound to a number of contexts. Such contexts can be defined as follows:

- The *context of an object* includes any statement on the relations between a cultural object (e.g. a photo, an artwork, a document) and entities involved in the object lifecycle (e.g. people, places, dates). Statements on the context of a cultural object answer questions such as: Who is the author of the artwork? When was the artwork created? Where was it created? How was it created?

- The *context of a statement* includes the provenance of the aforementioned statements, which answers the question: Who claims it? When was it claimed? It represents the activities performed by an actor in making such a statement, i.e. the hermeneutical activity. The latter includes sources, criteria, and motivations supporting the statement, which also answer questions such as: What type of statement is it? How is the conclusion reached? Is it authoritative?

Moreover, when catalogue information comes in a machine-readable format, a third context layer applies:

- The *meta-context of a statement* includes provenance information of the formal representation of the aforementioned information. This context provides the extent of the validity of the machine-readable version of a statement, whether a human or a machine extracted such statements from a digital source, the original source, and when it was extracted. It answers the questions such as: Who is responsible for the machine-readable version of the statement? Where is it extracted from? When was it extracted?

Such contexts can be formally represented by means of ontologies, which adopt a number of descriptive approaches, namely:

- An *event-driven approach* can be applied to the description of the context of the object. Binary relations, e.g. object and datatype properties, and unary relations, i.e. individuals representing events (e.g. creation, shot) can be further annotated. Such an approach is widely used in the cultural heritage domain and it is particularly embraced by CIDOC-CRM.

- An *interpretation-driven approach* can be applied to represent provenance of statements where events are expressed in the form of unary relations. A unary relation representing the hermeneutic activity can be linked and annotated with all the features describing how the statement was reached.

- *Named graphs* can be used to gather machine-readable statements about the two aforementioned scenarios, and be further annotated with meta-level information.

The *Historical Context Ontology (HiCO)*[1] [Daquino and Tomasi, 2015], is an OWL 2 DL ontology developed for representing features related to the context of questionable information. HiCO is a task ontology, meaning it addresses aspects related to a single, domain-independent, representational issue.

In particular, it addresses features characterising hermeneutic activities underpinning context information. Context information refer to all those pieces of information that describe events that were attributed by an actor (e.g. a historian, a cataloguer) to a subject of interest (e.g. a cultural object) and that are recorded in a source of information (e.g. a cataloguing record). For example *being created by somebody* or *being created at a certain time* are events related to an artwork that are claimed by a cataloguer, recorded in a cataloguing record, stored as a metadata document, and transformed into RDF statements. Secondly, HiCO addresses all the features that contribute to define the authoritativeness of a statement on the context information. For instance, the cataloguer cites a bibliographic source or a scholar's opinion to support the statement, and records the date of the attribution.

As a good practice, existing ontologies have been directly reused in HiCO (prefix `hico`) so as to represent specific aspects: an OWL DL 2 formalization of the FRBR model (prefix `frbr`) was considered for describing sources of information such as cataloguing records or cited sources; the PROV-O ontology (prefix `prov`) was used to describe the provenance of a statement and it was extended so as to describe features for validating the authoritativeness of statements, such as motivations, criteria, and primary sources; the CiTO

[1] `http://purl.org/emmedi/hico`

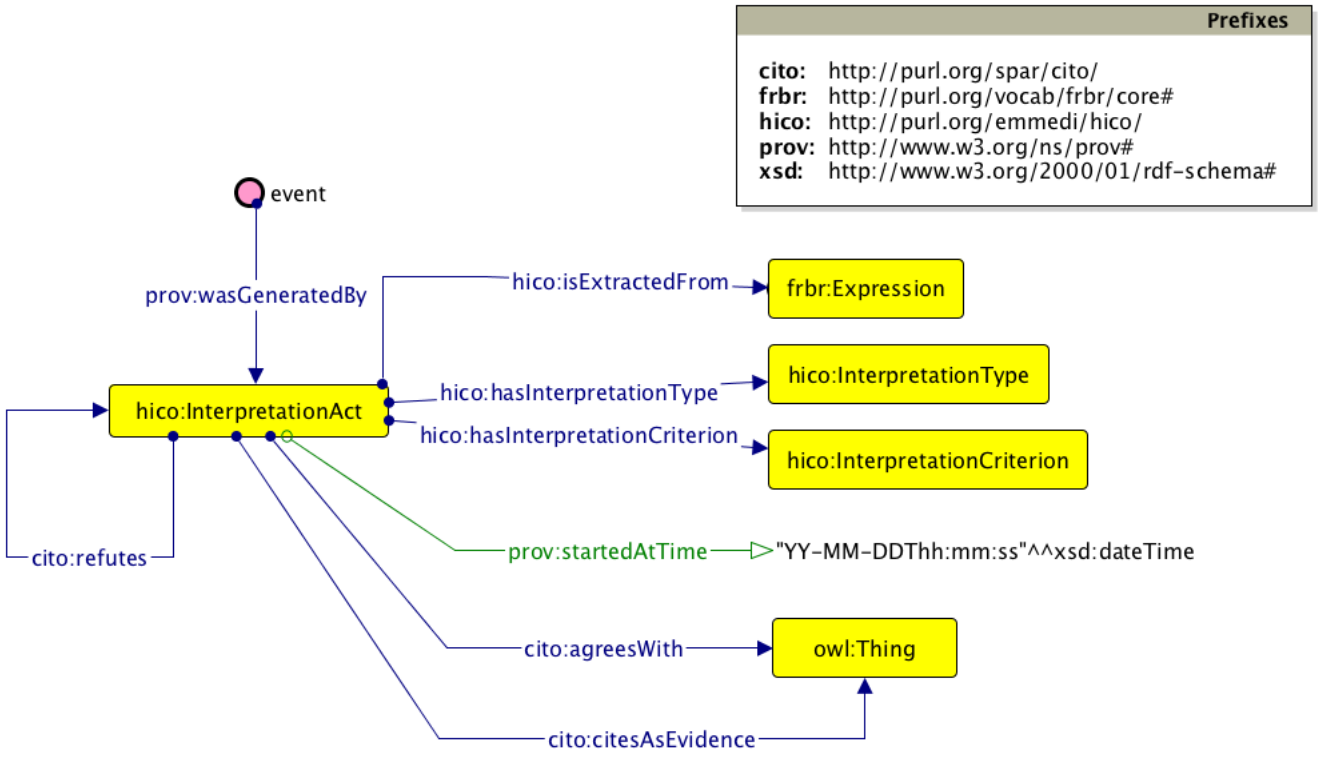


Figure 5.1: The HiCO Ontology main classes and properties

ontology (prefix `cito`) is imported to describe relations between interpretations, sources of information, and involved agents.

Figure 5.1 shows classes (yellow rectangles), object properties (blue dotted lines beginning with a solid circle and ending with a solid arrow), and assertions among classes (black lines ending with a solid arrow). The main class of HiCO is `hico:InterpretationAct`. An interpretation act is a situation in which a statement about the context of an object is linked to all the pieces of information necessary to validate its reliability (i.e. the context of the statement). This includes the following aspects:

- The classification of the interpretation, e.g. being an authorship attribution
- The description of criteria motivating the statement, e.g. bibliography, scholar's attribution
- Cited sources of information, e.g. a bibliographic source, an oral communication
- Temporal extent of the attribution, i.e. when it was first recorded
- The document wherefrom RDF statements are extracted, i.e. the cataloguing record.

In details, an individual representing an event, e.g. the creation of an artwork, is related to an individual of the class `hico:InterpretationAct` by means of a property of the PROV-O Ontology, i.e. `prov:wasGeneratedBy`. Individuals of `hico:InterpretationAct` class are defined by means of a number of object properties, namely:

- `hico:hasInterpretationType`. The value describes an arbitrary classification of the interpretation, such as being an authorship attribution or a date attribution.
- `hico:hasInterpretationCriterion`. The value describes the criterion used to support the questionable information, e.g. usage of bibliography, quotation of a scholar's opinion, prior attributions. Terms should be taken from a controlled vocabulary on methodological aspects characterising a given research field.
- `cito:citesAsEvidence`. Values identify sources of information
- `cito:agreesWith`. Values identify scholars' opinions that agree with the statement.
- `cito:refutes`. Links contradictory attributions on the same subject
- `prov:startedAtTime`. The value is the date to the attribution
- `hico:isExtractedFrom`, a subproperty of PROV-O `prov:wasInfluencedBy` property. The value is an individual representing the content of a document (class `frbr:Expression`), where questionable statement are taken from, e.g. a cataloguing record. Assuming the author of the document is also responsible for the interpretation act, the author of the attribution can be easily inferred.

Statements on context of the object and the context of the statement can be included in named graphs further annotated with information related to the creation of the machine-readable data. The responsible entity for the creation of the RDF statements is the value of the property `prov:wasAssociatedWith`, and the datetime of the transformation is the value of the property `prov:atTime`.

The HiCO Ontology was applied in the Zeri & LODE project for representing authorship attributions and to formalise a controlled vocabulary of terms describing the methodology of photo archives. A detailed

example of the usage of the model is provided in the next section, together with other models and a complete real world example.

5.2 The FEntry Ontology and the OAEntry Ontology

As detailed in Chapter II, CIDOC-CRM is deemed the golden standard for representing cultural heritage objects. The model allows to describe most of the entities and relations that naturally characterize the Photography and Arts domain. However, it lacks of the description of some peculiar of information produced in art historical photo archives, namely:

- aspects characterizing the cataloguing process of photographic reproductions;
- roles held by actors that intervene in the photograph and artwork lifecycles;
- relations (such as influence, derivation) between cultural objects and their conception
- attributions and related hermeneutic aspects
- citations and characterisation of the intention of citations (e.g. agreement, disagreement)

As a good practise, we (1) reused CIDOC-CRM as much as possible to describe cataloguing information in art historical photo archives, (2) we reused other ontologies to add terms for describing aspects not available in CIDOC-CRM, and (3) we created new terms when no existing ontology included those. Terms actually used for describing art historical photo archives were gathered in two specular ontologies, called *F Entry Ontology* and *OA Entry Ontology* for describing respectively the Photography and Arts domain related aspects. They provide terms for representing information related to artefacts catalogued according to ICCD-F and ICCD-OA standards.

The first ontology developed for representing Photography and Arts in the Cultural Heritage domain is the *F Entry Ontology*, which deals with aspects related to the Photography domain. Likewise, the *OA Entry Ontology* provides entities and relations for representing the Arts domain. Several aspects are in common between the original ICCD-F and ICCD-OA content standards, and therefore the two ontologies present similarities as well. The development of the OA Entry Ontology resulted in a revision of the F Entry

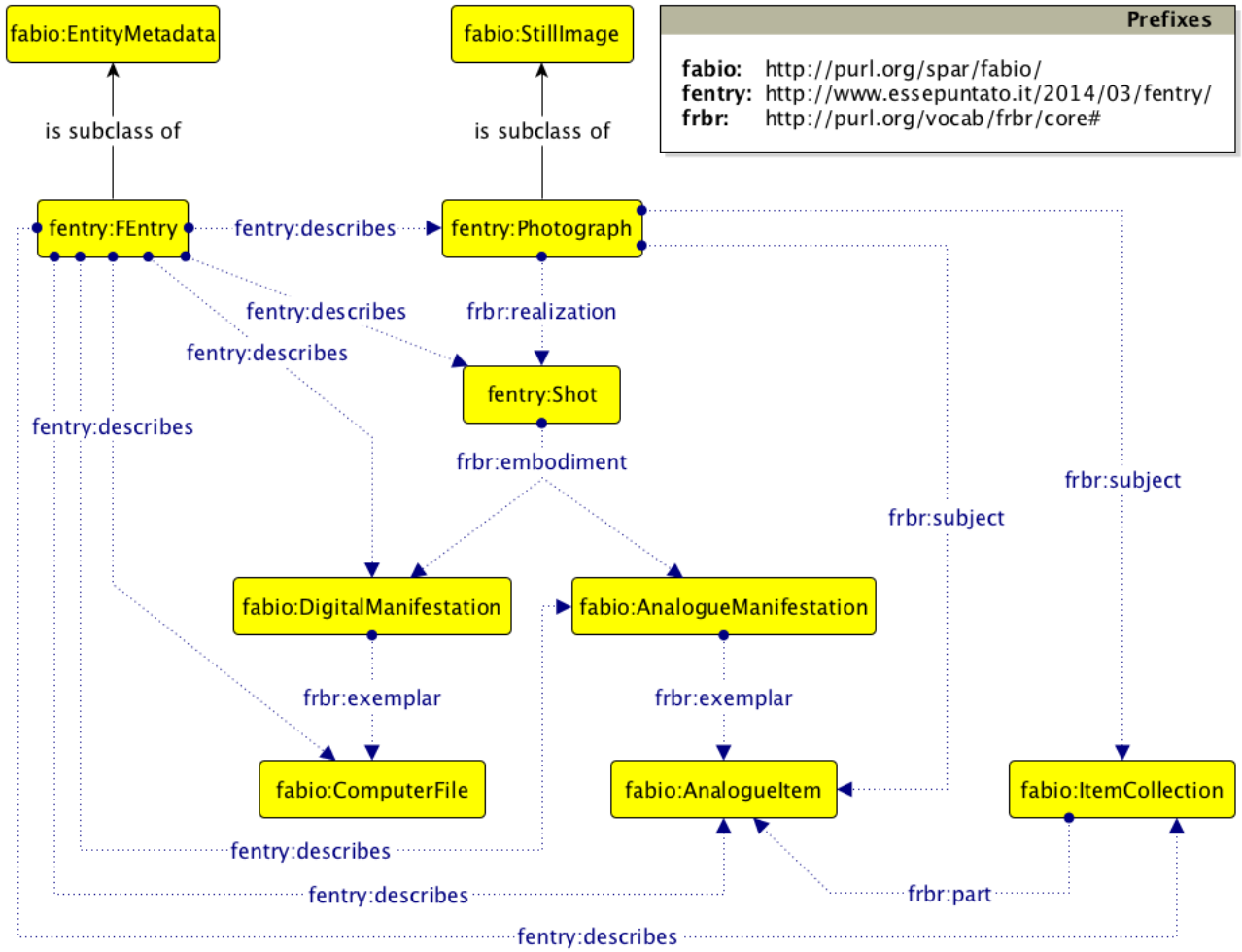


Figure 5.2: The F Entry Ontology main classes and properties

Ontology, so as to provide a specular description of same issues when appropriate, and include aspects related to hermeneutic approach that were not previously covered. For the sake of brevity, common aspects of the two ontologies are detailed only once when describing the F Entry Ontology.

The F Entry Ontology. The current version of F Entry Ontology (FEO)[2] revises the previous version introduced in [Gonano et al., 2014]. The Graffoo diagram of FEO in Figure 5.2 provides an overview of its main entities.

FEO introduces classes and properties needed to characterize three main concepts: (1) the photograph, (2) the subject portrayed in the photograph, and (3) the F Entry describing the photograph and its subjects. Each of the aforementioned entities is characterised in terms of FRBR. In particular:

[2] `http://www.essepuntato.it/2014/03/fentry`

- the photograph is represented as an FRBR Work when describing its essence, as an FRBR Expression when dealing with information about its realization (i.e. the shot), as an FRBR Manifestation when describing each tangible form of the photograph, and as FRBR Item for each individual copy with different features;

- in art historical photo archives the portrayed subject of a photograph is always a work of art. Therefore, it can be represented as an FRBR Work when defining its essence, and as FRBR Item when describing the physical object;

- the F Entry is a work containing metadata about a photograph and its cataloguing; it is subject to several revisions, each of which is related to responsible entities (i.e. cataloguers and supervisors); we are not interested in how it is preserved, what formats and how many copies there are. Thus an F Entry is represented as an FRBR Work when describing its creation and as an FRBR Expression when describing its contents and revisions.

Existing ontologies have been imported into FEO so as to provide a precise description of specific aspects of the domain. In particular, we imported the FRBR-aligned Bibliographic Ontology (FaBiO, prefix `fabio`), the Publishing Roles Ontology (PRO, prefix `pro`), the Historical Context Ontology (HiCO, prefix `hico`), and the Citation Typing Ontology (CiTO, prefix `cito`). In addition to terms from these ontologies, terms from an OWL 2 DL version of FRBR (prefix `frbr`) are also used, so as to represent hierarchical and associative relations between the main entities, as well as terms defined in the PROV Ontology (prefix `prov`).

In the next paragraphs the usage and the extensions of such ontologies are described along with examples.

Extending FaBiO to define the cultural object. FaBiO was originally developed for describing bibliographic entities according to the FRBR conceptual model. It mainly addresses issues related to published texts, by introducing wide taxonomies of possible kinds of works, expressions, manifestation and items. We refined the model in order to represent our main entities (i.e. F/OA entries, photographs, and works of art).

A F Entry describes a photograph in each phase of its lifecycle, such as its creation, its realization (the shot), its development into a visible image (negative, positive, slide, digital image) and its publishing and

reproduction. Each phase of the lifecycle of the photograph corresponds to an instance of a class defined in terms of FRBR.

The cataloguing record is defined in terms of FRBR Work (an instance of the class `fentry:FEntry`, subclass of `fabio:EntityMetadata`). Subjects of the entry are considered FRBR Works as well, including (1) the photograph, that is an instance of the class `fentry:Photograph`, subclass of `fabio:StillImage`, and (2) the work of art portrayed in the photograph, that is an instance of the class `fabio:ArtisticWork`. The object property `fentry:describes` links an instance of the class `fentry:FEntry` to instances of classes `fentry:Photograph` and `fabio:ArtisticWork`.

For example, the natural language scenario "an F Entry describes the photograph portraying the painting called Jesus's baptism", may be expressed, in Turtle syntax, as follows:

```
:fentry-72486 a fentry:FEntry ;
        fentry:describes :jesus-baptism-photo-work , :jesus-baptism-photo-item ,
                :jesus-baptism-work , :jesus-baptism-item .
:jesus-baptism-photo-work a fentry:Photograph .
:jesus-baptism-photo-item a fabio:AnalogItem .
:jesus-baptism-work a fabio:ArtisticWork .
:jesus-baptism-item a fabio:AnalogItem .
```

The shot is described as an instance of the class `fentry:Shot`, subclass of `fabio:Expression`. It is a realization of a photograph, which can take several forms when developed in analogue/digital formats - that are defined as instances of the classes `fabio:AnalogManifestation` or `fabio:DigitalManifestation`. Physical objects can be described as instances of the classes `fabio:AnalogItem` or `fabio:DigitalItem`.

For example, the natural language scenario "The shot of the photograph portraying the Jesus's baptism painting that had been taken by Brogi before 1940 was published by himself in 1940", can be expressed, in Turtle syntax, as follows:

```
:jesus-baptism-photo-work frbr:realization :jesus-baptism-photo-shot .
:jesus-baptism-photo-shot a fentry:Shot ;
        frbr:embodiment :jesus-baptism-photo-positive .
:jesus-baptism-photo-positive a fabio:AnalogManifestation .
```

Using PRO to describe the lifecycle of the object. PRO allows one to describe scenarios in which agents hold roles with respect to a particular time and context. An instance of the class `pro:RoleInTime` is created every time we need to specify these kinds of situation.

For instance, Brogi is both the photographer that made the shot (an FRBR Expression) and the publisher of the positive of the photograph (an FRBR Manifestation). These relations can be represented as follows:

```
:brogi a foaf:Agent ;
        pro:holdsRoleInTime :brogi-photographer-jesus-baptism-photo-shot ;
        pro:holdsRoleInTime :brogi-publisher-jesus-baptism-photo-positive .
:brogi-photographer-jesus-baptism-photo-work a pro:RoleInTime ;
        pro:withRole scoro:photographer ;
        pro:relatesTo :jesus-baptism-photo-shot ;
        tvc:atTime :jesus-baptism-photo-shot-date .
:brogi-publisher-jesus-baptism-photo-positive a pro:RoleInTime ;
        pro:withRole pro:publisher ;
        pro:relatesTo :jesus-baptism-photo-positive ;
        tvc:atTime :jesus-baptism-photo-publishing-date .
```

There may be some situations in which the creator and the realizer of the shot are not the same person. While the creation of a work is described by means of CIDOC-CRM terms (explained in the next section), terms belonging to PRO are used for describing other roles than the creator.

Using HiCO to describe provenance of assertions. HiCO was developed to describe hermeneutic aspects underlying questionable information, including provenance of potentially contradictory statements. For instance, the scenario described in the previous sub-section is a questionable information. The relation between an individual of the class `pro:RoleInTime` and the RDF-defined interpretation act is introduced by means of the object property `prov:wasGeneratedBy`. The situation where a questionable information is generated is defined by an individual of the class `hico:InterpretationAct`, which allows to specify the scope (property `hico:hasInterpretationType`), criteria (`hico:hasInterpretationCriterion`) and eventually sources and dates of the statement. Instances of `hico:InterpretationAct` are linked by means of the property `hico:isExtractedFrom` to the text source of the the F/OA Entry where such questionable information is stated in natural language, i.e. an instance of the class `fabio:MetadataDocument` representing a FRBR Expression. Such information is described in a named graph that is in turn annotated with the property `prov:wasAssociatedWith`, defined in PROV-O, so as to link the attribution to the author of RDF statements. The author of the attribution itself is the author of the original text.

For instance, consider the following natural language scenario: "the attribution of Brogi as the publisher of the photograph portraying the Jesus's baptism painting, was motivated by a formal analysis of the photograph itself, which revealed on its verso an inscription naming Brogi as publisher."

This natural language scenario can be expressed in RDF by using the ontological entities introduced above, as follows (in Turtle syntax):

```
:brogi-publisher-jesus-baptism-photo-positive-graph {
  :brogi-publisher-jesus-baptism-photo-positive
        prov:wasGeneratedBy :jesus-baptism-photo-publisher-attribution .
  :jesus-baptism-photo-publisher-attribution a hico:InterpretationAct ;
        hico:hasInterpretationType :role-attribution ;
        hico:hasInterpretationType :zeri-preferred-attribution ;
        hico:hasInterpretationCriterion :formal-analysis ;
        hico:hasInterpretationCriterion :inscription ;
        hico:isExtractedFrom :fentry-72486-expression ;
  :role-attribution a hico:InterpretationType .
  :zeri-preferred-attribution a hico:InterpretationType .
  :formal-analysis a hico:InterpretationCriterion .
  :inscription a hico:InterpretationCriterion .
  :fentry-72486-expression a fabio:MetadataDocument .
  :crr-mm a foaf:Agent .
  :brogi-publisher-jesus-baptism-photo-positive-graph prov:wasAssociatedWith :crr-mm .
}
```

In the excerpt, the instance `:zeri-preferred-attribution` is provided in order to distinguish the current interpretation chosen by the cataloguing institution from discarded attributions specified elsewhere.

Using CiTO for relating documents and attributions. The relation between an interpretation and its sources can be defined as a proper (even implicit) citation. CiTO allows one to mark citation links between citing and cited entities and to specify the intent of such citations by means of a wide set of object properties. To this end, we can use the object properties provided in CiTO for linking an individual of the class `hico:InterpretationAct` to the original textual interpretation from which the interpretation act was derived.

For instance, in the prior example the cataloguer cites as an evidence an inscription on the verso of the photograph recording the publisher's name, which can be represented by means of the object property `cito:citesAsEvidence` as follows:

```
:jesus-baptism-photo-publisher-attribution
        cito:citesAsEvidence :jesus-baptism-photo-verso .
```

The OA Entry Ontology. While ICCD-F content standard provides just few elements with regard to the work of art that may be portrayed in a photograph, ICCD-OA aims to be an exhaustive reference

document providing a complete description of any work of art. For this reason, aspects peculiar of the work of art portrayed in a photograph have been modelled in the OA Entry Ontology. The Graffoo diagram in Figure 5.3 provides an overview of its main classes and properties.

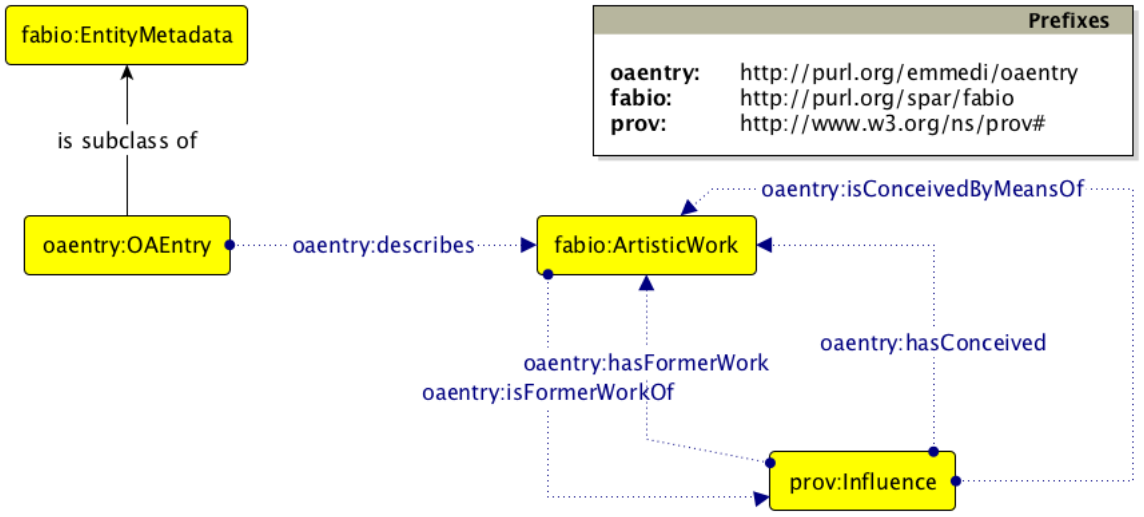


Figure 5.3: The OA Entry Ontology main classes and properties

The OA Entry Ontology introduces three main concepts: (1) the work of art, (2) the OA Entry that contains metadata about the work of art, and (3) the relations between artworks detailed in the cataloguing record. In particular:

- a work of art can be described in different phases of its lifecycle (creation, restoration, location, ownership, custody etc.). We decided to represent a work of art as an FRBR Work when describing its essence, as an FRBR Manifestation when providing information about its physical features that may change over time, and as an FRBR Item when dealing with information about legal aspects and its location;

- the OA entry is a document containing metadata about a work of art and its cataloguing. An OA entry is defined as an FRBR Work when describing its creation and as an FRBR Expression when addressing issues related to its content;

- the relation between two works of art concerns the way one artwork affects the conception of another artwork. Such an influence regards conceptual entities (the FRBR Work level).

In detail, an OA Entry and the artwork here described are both defined as FRBR Work, respectively an instance of the class `oaentry:OAEntry`, subclass of `fabio:EntityMetadata`, and instance of the class `fabio:ArtisticWork`. The object property `oaentry:describes` links an OA Entry to the work of art, represented as a conceptual work (the FRBR Work level), the embodiment of the work of art (the FRBR Manifestation level), and the physical object (the FRBR Item level).

Like F Entry Ontology, the OA Entry Ontology includes other models, namely FaBiO, PRO, CiTO, HiCO, FRBR and PROV-O. These are reused or extended to represent relations between artworks and to provide a controlled vocabulary of roles characterising the Arts domain.

Extending PROV-O and HiCO for describing the influence between works. Original-to-derivative relations between two artworks are represented by means of the class `prov:Influence` from PROV Ontology. Different kinds of influence can exist between artworks, hence we extend PROV-O in the OA Entry Ontology by adding appropriate subclasses to `prov:Influence` (e.g. `oaentry:Copy`, `oaentry:Derivation`, etc.). The taxonomy is derived from the vocabulary adopted at the Zeri Photo Archive. The object property `oaentry:hasFormerWork` (i.e. a sub-property of `prov:entity`) allows one to link an individual of any of the influence classes to the original work of art. The object property `oaentry:hasConceived` enables one to link an individual of any of the influence classes to the derivative work of art in consideration.

For instance, the scenario described in the sentence "the anonymous drawing of Sistine Chapel is a drawing conceived as a derivative work of Michelangelo Buonarroti's frescos in Sistine Chapel" can be represented in Turtle syntax as follows:

```
:anonymous-drawing-sistine-chapel-work
        oaentry:isConveivedByMeansOf
:michelangelo-fresco-sistine-chapel-drawing .
:michelangelo-fresco-sistine-chapel-drawing a oaentry:Drawing ;
                oaentry:hasFormerWork :michelangelo-fresco-sistine-chapel-work ;
                oaentry:hasConceived :anonymous-drawing-sistine-chapel-work .
```

It is worth noticing that such an assertion is a questionable information (the author is unknown, the derivation is observed by the cataloguer). An instance of the class `hico:InterpretationAct` can be created to specify that the claimed influence was actually derived from a cataloguer's subjective choice. To this end, the OA Entry Ontology extends HiCO with terms for defining types of interpretation characterising

the Arts domain - e.g. the attribution of an influence between works may be represented as the individual `oaentry:influence-between-works-attribution`. This can be defined in RDF as follows:[3]

```
:michelangelo-fresco-sistine-chapel-drawing a oaentry:Drawing ;
        prov:wasGeneratedBy :drawing-attribution .
:drawing-attribution a hico:InterpretationAct ;
        hico:hasInterpretationType
                oaentry:influence-between-works-attribution;
        hico:hasInterpretationType :zeri-preferred-attribution ;
        hico:hasInterpretationCriterion :cataloguer-choice ;
        hico:isExtractedFrom :oaentry-15429-expression ;
        prov:wasAssociatedWith :crr-mm .
oaentry:influence-between-works-attribution
        a hico:InterpretationType .
:zeri-preferred-attribution a hico:InterpretationType .
:cataloguer-choice a hico:InterpretationCriterion .
:oaentry-15429-expression a fabio:Expression .
:crr-mm a foaf:Agent .
```

Extending PRO with a controlled vocabulary of roles in the Arts domain. OA Entries generally require the cataloguer to provide information about the artist responsible for the creation of the work, including the role s/he had in the creative process (e.g. colourist, painter). Moreover, when cataloguing the work of art, several responsibilities are attributed to cataloguers as well. We extended individuals of the class `pro:Role` in the OA Entry Ontology so as to describe additional roles proper to the Arts and cataloguing domains. In particular, we created two new subclasses of `pro:Role`, namely `oaentry:ArtisticRole` and `oaentry:CataloguingRole`, each including specific individuals:

- Those defined as instances of `oaentry:ArtisticRole` have been recognized by means of the open vocabulary adopted by the Zeri Foundation for describing the roles of artists and the controlled vocabulary provided by ICCD-OA, such as `oaentry:antiquarian`, `oaentry:architect`, etc.;

- those defined as instances of `oaentry:CataloguingRole` have been recognized as the main roles involved in the cataloguing process, such as `oaentry:cataloguing-institution` or `oaentry:-cataloguer`.

[3]The complete example is available at `http://dx.doi.org/10.6084/m9.figshare.3175048`

5.3 Mapping ICCD-OA and ICCD-F cataloguing standards to RDF

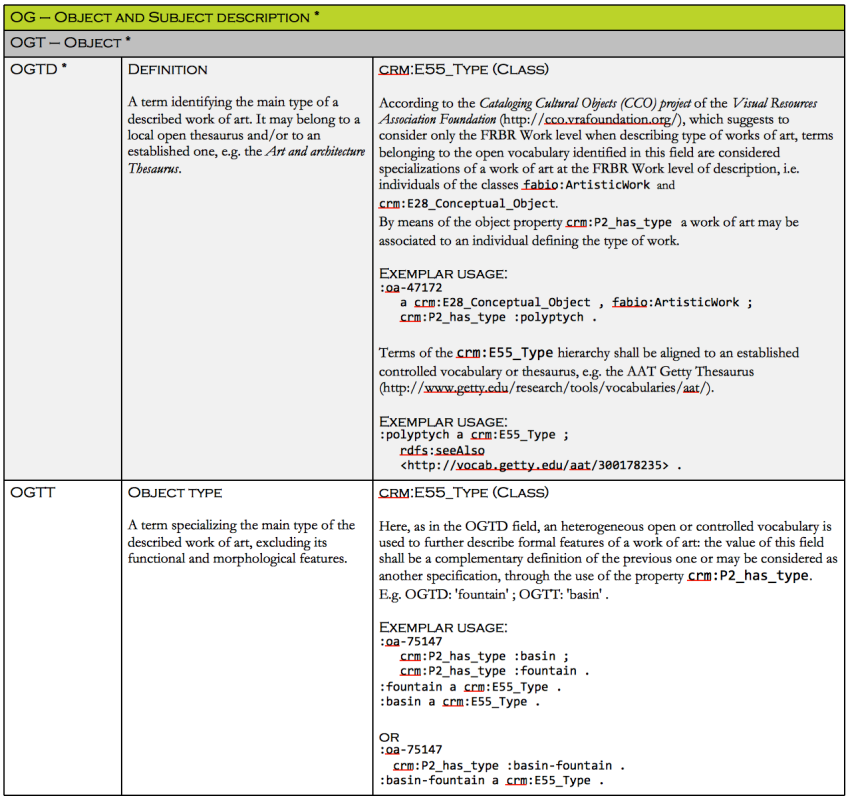
In this section we give an overview of the mapping of ICCD-OA and ICCD-F content standards to RDF according to the ontologies introduced in the previous section including omitted terms coming from the CIDOC-CRM specification. In this work we focus particularly on the mapping between all the fields in ICCD-F/OA that are actually used in the Zeri Photo Archive Catalogue - which include about 118 fields out of more than 300 provided by ICCD-F, and about 97 fields out of 280 provided by ICCD-OA. Rather than detailing all the fields, we focus on entities that are relevant to connoisseurship activities, that is: the cataloguing process, documents and primary sources, relations between cultural objects, relations between objects and actors.

Operatively, a first round of mapping was performed by (1) looking at all the aforementioned fields in the ICCD-F and ICCD-OA content standards one by one, (2) creating a first mapping to RDF accompanied by meaningful examples of usage. Secondly, a consultant (Silvio Peroni) double-checked the mapping. The resulting mapping document was analysed by the other three authors, including a member of the Zeri Photo archive responsible for Zeri's catalogue (Francesca Mambelli), a digital humanist (Francesca Tomasi), and a computer scientist (Fabio Vitali), so as to gather additional feedback. A new version of the mapping document was then released.

The mapping process resulted in the creation of two documents, i.e. *F Entry to RDF*[4] and *OAEntry to RDF*,[5] accompanied by exemplar data that represent contents of an F Entry[6] and an OA Entry in RDF[7], which were created according to such mappings. ICCD-F and ICCD-OA content standards are organized in sections. The two mapping documents contain tables structured as shown in Figure 5.4. Tables reproduce the structure of the content standards they refer to, and are organized in three columns. The first and the second column contain the name of the field in ICCD-F/OA and a brief description. The third column details the mapping to RDF terms and accompanies it with examples of usage.

The following paragraphs describe entities and relations mapped to RDF. For the sake of simplicity, paragraphs are organized on the basis of the four FRBR levels related to photographs and artworks, with an

[4] https://dx.doi.org/10.6084/m9.figshare.3175273
[5] https://dx.doi.org/10.6084/m9.figshare.3175057
[6] https://dx.doi.org/10.6084/m9.figshare.3175252
[7] https://dx.doi.org/10.6084/m9.figshare.3175048



OG – Object and Subject description *		
OGT – Object *		

OGTD *	Definition
	
	A term identifying the main type of a described work of art. It may belong to a local open thesaurus and/or to an established one, e.g. the *Art and architecture Thesaurus*.
		crm:E55_Type (Class)
		
		According to the *Cataloging Cultural Objects (CCO) project* of the *Visual Resources Association Foundation* (http://cco.vrafoundation.org/), which suggests to consider only the FRBR Work level when describing type of works of art, terms belonging to the open vocabulary identified in this field are considered specializations of a work of art at the FRBR Work level of description, i.e. individuals of the classes `fabio:ArtisticWork` and `crm:E28_Conceptual_Object`.
		By means of the object property `crm:P2_has_type` a work of art may be associated to an individual defining the type of work.
		
		Exemplar usage:
		`:oa-47172`
		`a crm:E28_Conceptual_Object , fabio:ArtisticWork ;`
		`crm:P2_has_type :polyptych .`
		
		Terms of the `crm:E55_Type` hierarchy shall be aligned to an established controlled vocabulary or thesaurus, e.g. the AAT Getty Thesaurus (http://www.getty.edu/research/tools/vocabularies/aat/).
		
		Exemplar usage:
		`:polyptych a crm:E55_Type ;`
		`rdfs:seeAlso`
		`<http://vocab.getty.edu/aat/300178235> .`

OGTT	Object type
	
	A term specializing the main type of the described work of art, excluding its functional and morphological features.
		crm:E55_Type (Class)
		
		Here, as in the OGTD field, an heterogeneous open or controlled vocabulary is used to further describe formal features of a work of art: the value of this field shall be a complementary definition of the previous one or may be considered as another specification, through the use of the property `crm:P2_has_type`.
		E.g. OGTD: 'fountain' ; OGTT: 'basin' .
		
		Exemplar usage:
		`:oa-75147`
		`crm:P2_has_type :basin ;`
		`crm:P2_has_type :fountain .`
		`:fountain a crm:E55_Type .`
		`:basin a crm:E55_Type .`
		
		OR
		`:oa-75147`
		`crm:P2_has_type :basin-fountain .`
		`:basin-fountain a crm:E55_Type .`

Figure 5.4: An excerpt of the mapping document "OA Entry to RDF"

introduction on top-level relations between entries and subjects described therein. Terms from the F Entry Ontology and the OA Entry Ontology are directly used without further explanation, while the use of CIDOC-CRM in the RDF excerpts is detailed.

The entry and its subject. Any F/OA Entry can be defined in CIDOC-CRM as an instance of the class `E31 Document`, a broader class than `fentry:FEntry` class. Individuals are linked to subjects by means of the object property `P70 documents`. An explicit relation between the F Entry describing a photograph and an OA Entry describing the work of art portrayed in that photograph can be represented by using `P67 refers to`. An entry can have several identifiers. Identifiers of the entries are instances of the class `E42 Identifier`, characterized with the property `P2 has type`. We used terms belonging to PRO and individuals of the class `oaentry:CataloguingRole` for describing the cataloguing process.

The following Turtle excerpt provides an example of all the aforementioned aspects:

```
:fentry-72486 a fentry:FEntry , crm:E31_Document ;
        fentry:describes :photo-72486 , :oa-47172 ;
        crm:P67_refers_to :oaentry-43677 .
:fentry-72486-creation a crm:E65_Creation ;
        crm:P14_carried_out_by :cataloguer ;
```

```
        crm:P4_has_time_span :2016-
:oaentry-43677 a oaentry:OAEntry , fabio:Work , crm:E31_Document ;
        oaentry:describes :oa-47172 ;
        crm:P140i_was_attributed_by
          :oaentry-43677-catalog-level-assignment ,
          :oaentry-43677-nctr-assignment , :oaentry-43677-nctn-assignment ;
:md-cataloguer-oaentry-43677 a pro:RoleInTime ;
        pro:relatesTo :oaentry-43677 ;
        pro:isHeldBy :md ;
        pro:withRole oaentry:cataloguer ;
        tvc:atTime :2012-11-04 .
```

The Work level. Both the photograph and the work of art can be defined in CIDOC-CRM terms as instances of the class `E28 Conceptual Object` when describing their essence and their creation. A direct relation between the photograph and the depicted work of art can be established relating the photograph (the FRBR Work) to the concrete object of art (the FRBR Item) by means of `frbr:subject`.

`E65 Creation` describes the authorship of photographs and works of art. Creators are instances of `E39 Actor` or one of its subclasses, such as`E21 Person` and `E74 Group`. The authorship attributions are specified by using the object property `P14 carried out by`. The object property `P4 has time span` is used to specify the duration of the creation event. The creation can be associated to a place (class `E53 Place`) by using the object property `P8 took place at`, and to a specific occasion (instance of the class `E4 Period`, further specified in `E5 Event`) by means of the object property `P10 falls within`. Creators can be associated to a cultural context (for instance a school of painters or a workshop) by using the property `P107i is current of former member of`, which relates them to an individual of the class `E74 Group`.

The archival description of the photograph, i.e. the hierarchical organization of the containers that include the catalogued object, can be described by means of transitive object properties `P106 is composed of` for relating conceptual entities. Titles can be attributed to the entities by means of the object property `P102 has title` associated to an instance of the class `E35 Title`, further specialized by using `P2 has type` to define whether the title is attributed, traditional or an alternate one.

When a bibliography or other sources are provided to support the cataloguing, e.g. letters, audio-recorded works, catalogues, entries etc., a generic relation can be represented with `P70i is documented in`,

linking to an individual of the class `E31 Document`.

The following Turtle excerpt introduces an example of the aforementioned aspects:

```
:photo-72486 a fentry:Photograph , crm:E28_Conceptual_Object ;
        crm:P94i_was_created_by :photo-72486-creation ;
        crm:P106i_forms_part_of :folder-leonardo ;
        crm:P102_has_title :jesus-baptism-verrocchio ;
        frbr:subject :oa-47172-item ;
        crm:P70i_is_documented_in :document-f2336 .
:photo-72486-creation a crm:E65_Creation ;
        crm:P94_has_created :photo-72486 , :photo-72486-expression ;
        crm:P7_took_place_at :florence ;
        crm:P10_falls_within :exhibition-of-paintings ;
        crm:P4_has_time_span :1926-1932 ;
        crm:P14_carried_out_by :brogi-studio .
:folder-leonardo a fabio:Work , crm:E90_Symbolic_Object ;
        crm:P106i_forms_part_of :subseries-leonardo .
:subseries-leonardo a fabio:Work , crm:E90_Symbolic_Object ;
        crm:P106i_forms_part_of :series-leonardo .
:series-leonardo a fabio:Work , crm:E90_Symbolic_Object ;
        crm:P106i_forms_part_of :zeri-photo-archive .
:zeri-photo-archive a fabio:WorkCollection .
:oa-47172 a fabio:ArtisticWork , crm:E28_Conceptual_Object ;
        crm:P94i_was_created_by :oa-47172-creation ;
        fabio:hasPortrayal :oa-47172-item .
```

The Expression level. While in the Arts domain we are not interested in representing contents of an artwork separately from its conception, in the Photography domain we separate the FRBR Expression, which is realized at the same time as the creation of the work, but can involve other actors than the main photographer that conceived the work. None of this information is precisely covered by the CIDOC-CRM, hence terms from FRBR OWL and F Entry Ontology are used. The following Turtle excerpt introduces an example of the aforementioned aspects:

```
:photo-72486-creation a crm:E65_Creation ;
        crm:P94_has_created :photo-72486 , :photo-72486-expression .
:photo-72486-expression a fentry:Shot ;
        crm:P94i_was_created_by :photo-72486-creation ;
        frbr:realizationOf :photo-72486 .
```

The Manifestation level. On one hand, photographs may appear in different formats, e.g. digital images, slides, negatives, and positives. Each manifestation belongs to the class `E22 Man-Made Object` and

represents a specific form that the work may have. On the other hand, artworks can be described as a FRBR Manifestation any time a relevant change affects the object, e.g. a restoration intervention.

Both the photograph and the work of art may be described in terms of the material they are made of, i.e. an instance of the class `E57 Material`. The class `E16 Measurement` is used to annotate the event of measuring various dimensions (`E54 Dimension`) related to such manifestations (e.g. weight and height). Other specific features characterizing the manifestation, e.g. colour, are linked with the object property `P56 bears feature`. The following Turtle excerpt introduces an example of all the aforementioned aspects:

```
:photo-72486-positive
        a crm:E22_Man-Made_Object , fabio:AnalogManifestation ;
        crm:P45_consists_of :gelatin-silver ;
        crm:P56_bears_feature :black-and-white ;
        crm:P39i_was_measured_by :photo-72486-positive-measurement .
:photo-72486-positive-measurement a crm:E16_Measurement ;
        crm:P40_observed_dimension :height-194mm ;
        crm:P40_observed_dimension :width-250mm .
```

The Item level. The photograph can be defined as an individual of the class `E22 Man-Made Object` (also inferred as `E84 Information Carrier`) linked to the portrayed work of art by means of the object property `P62 depicts`. Several actors with a specific role may be involved in the production of photographs and works of art. Terms belonging to PRO Ontology are here preferred.

Features regarding the object recorded by cataloguers during an assessment (individual of the class `E14 Condition Assessment`) are defined as instances of the class `E3 Condition State`, further specialized by using the `P2 has type` property.

Different locations (individuals of the class `E53 Place`) can be associated to physical items by using the object property `P55 has current location`. The current keeper (`E39 Actor`) is related to the place of conservation by using the object property `P74 has current or former residence`. A complete description of the current location of an object allows to describe transfers of custody (`E10 Transfer of Custody`) or changes of location (`E9 Move`).

The ownership of the object can be defined by using the property `P52 has current owner`. Each owner (`E39 Actor`) could have acquired the work (property `P22i acquired title through`) as

the consequence of an acquisition event (`E8 Acquisition`). Finally, exhibitions (`E5 Event`) may be recorded and linked to (1) the object by means of `P12 occurred in the presence of`, (2) the location (`P7 took place at`), (3) the date (`P4 has time span`) of the event, and (4) to a formal appellation (`E41 Appellation`) specified through the property `P1 is identified by`.

The following Turtle excerpt introduces a partial example of the aforementioned aspects regarding a photograph:

```
:photo-72486-positive-item a fabio:AnalogItem , crm:E22_Man-Made_Object ;
        crm:P62_depicts :oa-47172-item ;
        crm:P57_has_number_of_parts "1" ;
        crm:P34i_was_assessed_by :photo-72486-positive-item-condition ;
        crm:P55_has_current_location :large-formats-room ;
        crm:P140i_was_attributed_by :photo-72486-invn-assignment ;
        crm:P52_has_current_owner :university-of-bologna ;
        crm:P12i_was_present_at :exhibition-london-1987 ;
        crm:P30i_custody_transferred_through :photo-72486-item-provenance-1 .
:photo-72486-positive-item-condition a crm:E14_Condition_Assessment ;
        crm:P35_has_identified :photo-72486-positive-item-condition-state .
:photo-72486-positive-item-condition-state a crm:E3_Condition_State ;
        crm:P2_has_type :discrete ;
        crm:P3_has_note "silver mirror" .
:large-formats-room a crm:E53_Place ;
        crm:P89_falls_within :ex-convent-santa-cristina .
:photo-72486-item-provenance-1 a crm:E10_Transfer_of_Custody ;
        crm:P28_custody_surrendered_by :villa-i-tatti ;
        crm:P29_custody_received_by :zeri-foundation ;
        crm:P30_transferred_custody_of :photo-72486-positive-item ;
        crm:P4_has_time_span :1989 .
```

Chapter 6

A Conceptual Framework for Measuring Authoritativeness in Art Historical Data

In this chapter is presented the conceptual framework of dimensions and measures for assessing authoritativeness of contradictory attributions provided by art historical photo archives. Dimensions address *textual authoritativeness* of secondary sources and *cognitive authoritativeness* of cited scholars. Dimensions of textual authoritativeness are extracted and validated by means of data analysis over three datasets belonging to photo archives and are confirmed by domain experts' consultancy. Bespoke indexes for cognitive authoritativeness are developed so as to evaluate scholars' relative authoritativeness in a relatively narrow context. Metrics for evaluating dimensions are grouped into a ranking model. Results contribute to validate hypotheses *H3 (Analytical data and domain experts' feedback can be used to formalize the criteria underpinning the methodology of art historical data providers when publishing authorship attributions)*, *H4 (The evaluation of textual authoritativeness of sources recording authorship attributions can be based on a documentary, evidence-based approach)*, and *H5 (Measuring scholars' authoritativeness in the arts field can be achieved by developing bespoke metrics)*. Lastly, strategies for improving metadata quality in art historical photo archives are presented.

6.1 Approach to define authoritativeness in art historical photo archives

In the Arts field data providers offer competing information about the same artefacts. A peculiar aspect of art historical photo archives is that catalogues include detailed information on contradictory authorship attributions. Instead, museums and galleries rarely record such precious information in their catalogues. Therefore, assessing the validity of such questionable information is challenging. Despite providers offer high quality information, the lack of documentation, not-updated sources, and the disagreement with other providers may affect the authoritativeness of statements.

This study foresees the analysis of internal grounds of cataloguing records including motivations and argumentations (also called criteria) around authorship attributions. The aim is to define a set of dimensions characterising textual authoritativeness of sources and see how these interact with each other when validating authoritativeness of a statement. The first phase of this study includes the analysis of motivations supporting attributions in photo archives, the definition of a rating of those, and the validation of the latter through to a number of analyses. The approach to define and validate dimensions includes the following steps.

- **S1. Definition of criteria motivating attributions.** Criteria motivating authorship attributions are extracted from definitions and requirements outlined in the Italian content standard ICCD-OA.
- **S2. Comparison between definitions and actual implementation in a use case.** Analytical data on the actual usage of terms is extracted from the Zeri photo archive dataset, so as to define an initial set of dimensions to be validated.
- **S3. Domain experts' revision.** A photo archivist with a background in art history from the Zeri photo archive double-checked the list of criteria and provided a first rating of those, according to her knowledge and the actual implementation in the Zeri cataloguing data.
- **S4. Validation of domain experts' rating.** The rating is validated by performing an internal analysis on the Zeri photo archive data so as to check the consistency of the proposed rating.
- **S5. Validation of the rating in other sources.** Step S4 is iterated over two similar datasets to double-check the rating is shared in the community.

- **S6. Comparison of the three photo archive methodologies.** A comparative analysis over the three datasets is performed for reviewing the list of criteria and addressing flaws in archival policies.

The datasets analysed are subsets of the catalogues of the Federico Zeri photo archive, Villa I Tatti - Berenson Library, and the Frick Art Reference Library. The latter were chosen because (a) of the same scope of data, and because (b) they present similar cataloguing policies when recording information about authorship attributions. Results of the analysis are available online [Daquino, 2018a].

Secondly, a broader set of dimensions and metrics with regard to textual authoritativeness are addressed to balance the importance of criteria in cataloguers' final decision.

- **S7. Dimensions and metrics.** Selection of domain-dependent and domain-independent measures and metrics that apply to art historical data quality assessment.
- **S8. Ranking model.** Development of a ranking model gathering metrics to assess textual authoritativeness.

As a preliminary work, in step S7 we also define metrics for addressing cognitive authoritativeness. However, such indexes do not affect the final ranking model, since these are in a too early stage, due to a general infancy of citation indexes for scholars in the Humanities.

- **S7.1. Definition of scholars' citation indexes.** Citation indexes for representing scholars' credibility and trustworthiness are selected and tuned so as to present additional information to users.

6.2 Assessment of the methodology of art historical photo archives

Cataloguers of art historical photo archives are generally required to record motivations to support an authorship attribution, as prescribed by cataloguing standards. The three surveyed photographic collections record a broad range of detailed motivations and their favourite sources, including bibliographic sources, museum, auction, and scholars' opinions. For such reasons photo archives are chosen as subjects of the analysis.

The three archives adopt three different metadata standards, respectively: the ICCD-OA content standard, a custom metadata format based on Visual Resource Association (VRA) standard, and the MARC format compliant with Resource Description and Access (RDA) cataloguing standard. Among the three standards, only ICCD-OA, prescribes a controlled vocabulary of motivations. In the other cases, an open vocabulary is adopted by cataloguers, which overlaps with some ICCD-OA terms.

S1. Definition of criteria motivating attributions. The ICCD-OA controlled vocabulary called "AUTM" includes twenty terms for describing motivations that may support an authorship attribution (similar terms were grouped under the same label for the sake of brevity, hence the actual number is slightly higher). However, the actual usage of terms may differ in real scenarios.

In table 6.1, terms of the ICCD-OA vocabulary are listed (both in italian and in english), along with a brief description, and the usage in the Zeri photo archive, Villa I Tatti, and Frick Art reference Library, as resulted from the data analysis over the three datasets. The analysis is performed over a subset of RDF data gathered on a topic base, i.e. attributions of artworks of Modern Era, which includes 19.061 cataloguing records from the Zeri photo archive, 12.256 from Villa I Tatti, and 10.207 from the Frick Art Reference Library. "A" represent the usage of the term, while "N/A" identifies missing terms.

S2. Comparison between definitions and actual implementation in a use case. Terms defined in ICCD-OA vocabulary are modified by cataloguers of the Zeri photo archive to better fit their purposes, and new terms are included. Such new terms also appear in the other two surveyed archives. The list of 9 terms that are not included in ICCD-OA but are actually mentioned in the three photo archives is shown in table 6.2.

When analysing the difference between the guidelines and its actual usage in real scenarios, two interesting groups of motivations appeared, namely: motivations provided by scholars or other authorities, and motivations related to the appraisal of photographic documentation. Such two groups reflect peculiarities of photo archives, where connoisseurs used to study, share their opinions with photo archivists, and annotate photographs, and where photographs in turn became evidences of the different scientific methods pursued by scholars.

S3. Domain experts' revision. ICCD-OA rules do not provide guidance on the usage of motivations, nor on the extent to which a criterion should be deemed more reliable than another one when contradic-

N.	Term	Description	Zeri	I Tatti	Frick
1	Analisi diagnostiche / Diagnostic measures	Infrared ray and other non-invasive techniques are adopted for analysing the technique and the support of the artwork.	N/A	N/A	N/A
2	Analisi iconografica / Iconographic analysis	The study of themes depicted in the artwork.	N/A	N/A	N/A
3	Analisi stilistica / Stylistic analysis	The study of artist's techniques and style.	A	A	A
4	Analisi storica / Historical analysis	The study of the historical context of the artist/artwork.	N/A	N/A	N/A
5	Analisi tipologica / Type analysis	The study of the formal aspects of the artwork.	N/A	N/A	N/A
6	Bibliografia / Bibliography	The usage of articles/books as sources of information.	A	A	A
7	Bollo, Punzone / Stamp	A stamp of the museum, collection or owner records the attribution.	N/A	N/A	N/A
8	Confronto / Comparison	Comparison between similar artworks and analysis.	N/A	N/A	N/A
9	Contesto / Context	The artistic context provides insights on the authorship.	N/A	N/A	N/A
10	Documentazione / Documentation	Reports and expertises recording the assessment of the attribution.	A	N/A	A
11	Esame intervento / Analysis of the artist's intervention	The analysis of the contribution given by an artist to the production of the artwork.	N/A	N/A	N/A
12	Firma / Signature	The artwork is signed by the artist.	A	A	A
13	Grafia / Handwriting style	The artwork is annotated by the artist.	N/A	N/A	N/A
14	Fonte archivistica / Archival documentation	Correspondence, notes, or archival classification recording the attribution.	A	A	A
15	Iscrizione / Inscription	An inscription on the artwork.	A	N/A	A
16	Marchio, Timbro / Mark	A mark on the artwork.	A	N/A	N/A
17	Monogramma, Sigla, Simbolo / Monogram	A monogram on the artwork.	A	N/A	A
18	Nota manoscritta / Handwritten note	A handwritten note on the photograph depicting the artwork.	A	A	A
19	Tradizione orale / Traditional (oral) attribution	Traditional attribution ascribing the artwork to an artist, that may have been revised.	A	N/A	A
20	NR / Not recorded		A	A	A

Table 6.1: Usage of ICCD-OA controlled vocabulary of criteria supporting attributions in Zeri, I Tatti, and Frick photo archives

N.	Term	Description	Zeri	I Tatti	Frick
1	Scholar's attribution	The scholar officially ascribed the artwork to an artist. In general, the date of the attribution is recorded.	A	A	A
2	Scholar's note on photograph	The scholar ascribed the artwork to an artist by annotating the photograph. It is not sure this is the definitive attribution, which could have changed over time.	A	A	A
3	Museum attribution	The museum preserving the artwork officially ascribed the artwork to an artist.	A	A	A
4	Auction attribution	The auction house that sold the artwork at hand officially ascribed the artwork to an artist. Economic interests and lack of expertise may hinder the reliability of the attribution.	A	A	A
5	Collection attribution	The collection (private or public) preserving the artwork officially ascribed the artwork to an artist. Economic interests and lack of expertise may hinder the reliability of the attribution.	A	N/A	A
6	Market attribution	The market official attribution for the artwork at hand. Economic interests and lack of expertise may hinder the reliability of the attribution.	A	A	A
7	Anonymous note on photograph	An anonymous scholar ascribed the artwork to an artist by annotating the photograph. It is not sure whether this is the definitive attribution, which could have changed over time, nor the authoritativeness of the scholar, or the date.	A	A	A
8	False signature	The artist's signature is recognized as potentially false by means of the appraisal of the photographs of the artwork.	A	N/A	N/A
9	Caption on photograph	The appraisal of photographic documentation shows a caption recording the name of the artist, possibly provided by the photographer.	A	A	A

Table 6.2: Terms not included in ICCD-OA Controlled vocabulary used in Zeri photo archive, Villa I Tatti, and Frick Art Reference Library

tory attributions are compared. In art historical photo archives such a decision is made on the basis of cataloguers' expertise and their subjective interpretation of available sources. The list of criteria extracted from ICCD-OA specifications and from the three catalogues has been reviewed by photo archivists of the Zeri photo archive, who provided an initial rating. Table 6.3 lists terms in descending order of reliability, along with a description of their usage and the initial rating.

S4. Validation of domain experts' rating. To validate assumptions made by domain experts, the Zeri photo archive RDF dataset[1] is analysed first. In particular, given the list of accepted attributions and discarded attributions (when recorded in data), criteria that support accepted attributions are compared one-by-one to criteria that support discarded attributions. The aim is to double-check that the rating provided by cataloguers is coherent with the actual usage of criteria when contradictory attributions are compared.

Secondly, the analysis focuses on the nature of sources of information. These may be external (museums, auctions, scholars) or may depend on the the archive creator's influence on the cataloguing process (e.g. notes, attributions, publications at cataloguers' hand). Being the Zeri photo archive a personal archive (i.e. Federico Zeri's bequest), we want to address if accepted attributions are potentially biased, and how such criteria affect the rating.

The analysis is performed over the aforementioned subset of RDF data gathered on a topic base, i.e. attributions of artworks of Modern Era, which includes 19.061 artworks. Among these, 18.826 accepted attributions are annotated with one or more motivations, and 5.356 attributions include the description of discarded attributions. The latter subset is the subject of the comparative analysis. The subset is likely to be representative of the archival policies and the accuracy of the rest of the dataset, which includes artworks of diverse periods and may be characterised by different degrees of accuracy in the cataloguing process. We assume the comparative analysis may slightly differ if applied in different contexts. According to photo archivists of the Zeri Foundation, the subset is a representative demonstration of the average (or low) standard of accuracy of data.

Figure 6.1 shows the distribution of criteria and a comparative analysis of their usage. Rows represent criteria supporting accepted attributions, and columns represent criteria supporting discarded attributions.

[1] The dataset is available at `https://w3id.org/zericatalog/`

N.	Term	Description
1	Documentation	Reports and expertises provided by connoisseurs. In general, the date of the report is recorded.
2	Artist's signature	The artist's signature recognized as original by analysing photographs.
3	Bibliography	Bibliographic references recording information on the attribution.
4	Archival classification	The archive creator or photo archivists ascribed the artwork because of the arrangement of photographs in folders dedicated to the artist at hand.
5	Scholar's attribution	The scholar officially ascribed the artwork to an artist (verbally or in a source). In general, the date of the attribution is recorded.
6	Museum attribution	The museum preserving the artwork officially ascribed the artwork to an artist.
7	Scholar's note on photograph	The scholar ascribed the artwork to an artist and annotated it on a photograph. It is not sure this is the definitive attribution, which could have changed over time.
8	Inscription	The photograph shows an inscription is recorded on the support of the artwork, which can corroborate the attribution.
9	Sigla	The photograph shows a sigla is recorded on the support of the artwork, which can corroborate the attribution.
10	Auction attribution	The auction house that sold the artwork at hand ascribed the artwork to an artist. Economic interests, the date of the sale, and the lack of expertise may hinder the reliability of the attribution.
11	Collection attribution	The collection (private or public) preserving the artwork officially ascribed the artwork to an artist. Economic interests, the date of acquisition, and the lack of expertise may hinder the reliability of the attribution.
12	Market attribution	The market attribution for the artwork at hand. Economic interests and lack of expertise may hinder the reliability of the attribution.
13	Traditional attribution	The generally accepted attribution. It may be not updated, or it may be overcome by more recent discoveries.
14	Stylistic analysis	The attribution is performed by photo archivists by means of the appraisal of photographic documentation.
15	Anonymous note on photograph	An anonymous person ascribed the artwork to an artist and annotated the photograph. It is not sure whether this is the definitive attribution, which could have changed over time, nor the authoritativeness of the scholar, or the date.
16	False signature	The artist's signature is recognized as potentially false by means of the appraisal of the photographs of the artwork.
17	Caption on photograph	The photograph shows a caption recording the name of the artist, possibly provided by the photographer.
18	Other	Other criteria.
19	None	No criterion is recorded.

Table 6.3: Criteria rated by photo archivists at the Zeri photo archive

tot.	ACCEPTED \ DISCARDED	documentation	artist's signature	scholar's attribution	bibliography	archival classification	scholar's note on photo	museum attribution	inscription	sigla	auction attribution	collection attribution	market attribution	traditional attribution	stylistic analysis	anonymous note on photo	false signature	caption on photo	other	none
	tot.	108	8	783	2547	334	318	88	6	4	701	153	227	8	1	1259	11	56	4	0
34	**documentation**	15		21	12	13	3				3		4			4				
26	**artist's signature**		3	5	13	2	3				1	1				3				
2629	**scholar's attribution**	69	3	527	973	194	144	42	4	3	436	76	142	2	1	798	6	24	3	
1697	**bibliography**	17	2	253	1288	199	81	29			96	32	16	5		201		24	1	
5322	**archival classification**	108	8	795	2585	328	315	88	6	4	700	153	226	8	1	1246	11	56	4	
471	**scholar's note on photo**	6	1	49	218	53	110	5	1		58	12	21			102		6		
1	**museum attribution**			1		1		1												
3	**inscription**				3															
2	**sigla**				1											1				
73	**auction attribution**			35	17	39	8				48	1	4			3				
13	**collection attribution**			5	1	6	3					13				1				
28	**market attribution**			13	5	11					2	1	13			3				
0	**traditional attribution**																			
8	**stylistic analysis**			1	1		1				4					1	1			
110	**anonymous note on photo**			43	37	56	8				8	2	1			72		1		
0	**false signature**																			
5	**caption on photo**			4	1	5												5		
111	**other**	8		9	26			1			16	4	7			46				
132	**none**	6		14	45		4	2			17	6	5			45	1	2		

Figure 6.1: Distribution and comparison of criteria adopted by the Zeri photo archive

More than one criterion may support an attribution, hence there is an overlap in the usage of criteria. Values in columns "tot." represent the total number of records that use the criterion at hand which differs from the number of occurrences of the term, that can be calculated by summing numbers in the row.

For instance, the criterion "documentation" supports 34 accepted attributions and 108 discarded attributions; in 21 attributions the criterion "documentation" is accepted over the criterion "scholar's attribution" (first row, third cell), which in turn is accepted 69 times over the criterion "documentation" (third row, first cell).

The distribution of the nineteen criteria immediately highlights photo archivists' preferences. Decisions

recorded during the cataloguing process mainly rely on the archival classification, which may overlap with Federico Zeri's original arrangement. The criterion supports the 99% attributions (5.322). Secondly, scholars' attributions (2629, i.e. 49%), and bibliographic references from the vast art library (1697, i.e., 32%) are the main tools of cataloguers.

Some criteria are not well represented in the dataset, such as museums attributions, collection attributions and traditional attributions. In such cases we trust the original rating provided by archivists.

Other relevant criteria provide interesting insights on the peculiarities of the archival policies at the Zeri Foundation. According to archivists, the criterion documentation is deemed the most reliable. However, it happens to be mostly discarded when the accepted attribution is supported by "archival classification" (108), "scholar's attribution" (69), bibliography (17), and "scholars' note on photographs" (6). Analysing the provenance of scholars' attributions we notice that 64 out of 69 are Federico Zeri's attributions, 1 out of 17 is Zeri's bibliography, and 6 out 6 are Zeri's notes. We assume in this case cataloguers are influenced by the archive creator's opinion and the other documentation at hand might be less recent.

Scholars' attributions and bibliography are supposed to be higher rated than decisions taken during the archival classification. Nonetheless, these are often deemed less reliable than the archival classification: respectively, "scholar's attribution" is accepted 194 times and discarded 795 times; "bibliography" is accepted 199 times and discarded 2585 times. We assume in such cases the influence of the cataloguing process is predominant and does not respect the original rating. However, such criteria are consistently preferred over lower rated criteria.

A similar inconsistency is found between "scholar's note on photo" and "bibliography". The former is often preferred over the second (218 times accepted / 81 times discarded), despite an annotation is deemed less reliable than publications by archivists. In particular, we notice that 81 annotations out of 81 are signed by Federico Zeri. Again, we assume in this case cataloguers are influenced by the archive creator's opinion, whose notes may be more recent than the bibliography at hand.

According to data, a philological approach pursued by archivists (i.e. recording the will of the archive creator) and researches performed by cataloguers during the cataloguing process guide the attributionship process. Nonetheless, the criterion archival classification is mainly accompanied by other motivations, and is rarely the only reason of attribution. As above explained, the table includes an overlap of concurring

criteria that support the same attribution. Figure 6.2 is represented the distribution of criteria that appear along with the criterion "archival classification" when supporting an accepted attribution.

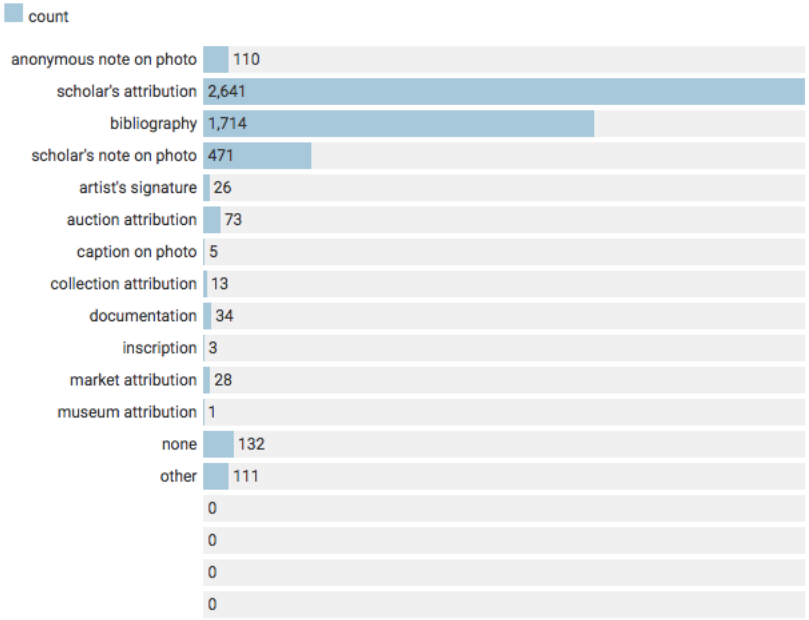


Figure 6.2: Distribution of criteria that appear along with "archival classification" in the Zeri photo archive

In details, "archival classification" mainly appears along with the following criteria:

- Scholar's attribution (2641 times out of 5322). The criterion is originally meant as more reliable than archival classification. When the two criteria are used together, the attribution at hand is consistently accepted as favourite over other low rated criteria.
- Bibliographic reference (1714 times). Such a criterion is deemed more reliable than archival classification, and the attribution is consistently supported with high rated criteria.
- Scholar's note on the photograph (471 times). In this case the criterion is deemed less reliable than "archival classification". We assume the acceptance of attributions supported by such criteria is based on a cumulative approach (i.e. two concurring criteria supporting the attribution make it more reliable than an attribution supported by one criterion only, despite that may be more reliable) or depends on other factors, such as more recent attributions compared.

In summary, 5033 records out of 5322 attributions are supported by more than one criterion or source of information, which increases their degree of reliability. We deduce that a historical approach is mainly

pursued by cataloguers at the Zeri photo archive (i.e. detailed researches on the available literature are performed by cataloguers). To validate such an assumption, three criteria that often include reference the archive creator Federico Zeri are compared. In details, when "scholar's attribution", "bibliography", and "scholar's note on photo" appear along with "archival classification" to support an accepted attribution, we notice that:

- 2513 (out of 2641) scholars' attributions are Federico Zeri's attributions.
- 169 (out of 1714) bibliographic references are Federico Zeri's publications.
- 471 (out of 471) annotations on photographs are made by Federico Zeri.

We revise the original rating so as to include such three situations, and clearly distinguish when the decision recorded takes into account an extended literature or relies only on the archive creator's bequest. The archive creator's influence is found when he is cited as (1) an official source of the attribution, (2) the author of a cited bibliographic reference, or (3) has annotated a photograph. We define such situations as follows:

- Archive creator's attribution. According to domain experts such criterion should be deemed less relevant than expertises and documentation, but more relevant than the archival classification. Other scholars' attributions are in turn lowered in the rating, since their acceptance highly depends on the decision taken during the cataloguing process (i.e. archival classification).
- Archive creator's bibliography. According to domain experts such criterion should be deemed less relevant than documentation and archive creator's official attributions (which are generally updated and may revise prior publications), while it is more relevant than other bibliographic references and the archival classification.
- Archive creator's note on photograph. Assuming that (1) notes may be overtaken by official statements and publications, (2) the author may have changed opinion over time, and (3) the cataloguing process must ensure an accurate review of existing sources, this criterion is deemed less reliable of archival classification but more reliable than other scholars' notes.

The three new criteria are included in the rating according to both archivists preferences and their actual usage. It is worth to notice that such exceptions are also found in Villa I Tatti catalogue (that is a personal archive too), but not in Frick Art Reference library - which is not a personal photo archive. For the sake of brevity we illustrate the final rating after showing the comparative analysis performed over the other two datasets.

S5. Validation of the rating in other sources. To ensure the proposed rating is valid and shareable, data analysis is performed over Villa I Tatti - Berenson Library and the Frick Art Reference Library datasets.

Villa I Tatti - Berenson Library was chosen because of the similarity with the Zeri photo archive, meaning they are both personal photo archives, their photographic collections overlap (i.e. they describe same artworks, and sometimes share the same pictures), and the methodology to assess authorship attributions is likely to be similar, or comparable. In particular, the influence of the archive creator's opinion in resulting data is investigated.

The Frick Art Reference Library is chosen to validate the rating over another type of photo archive, since it is not characterised by the influence of a particular scholar. Since it is not a personal archive, the methodology does not include references to a predominant scholar, and is supposed to be slightly different. Nonetheless, it is an interesting test case for evaluating what we called cognitive authoritativeness with regard to archive creators, i.e. Federico Zeri and Bernard Berenson.

The two datasets are first evaluated singularly, so as to review or confirm the validity of the proposed rating.

Villa I Tatti - Berenson Library. The dataset provided by Villa I Tatti includes information on a group of images of Renaissance Italian paintings that Bernard Berenson famously classified as "homeless,"[2] that is, works that were documented by a photograph but whose current location was unknown to him. Bernard Berenson, intended to use such information to bring up to date his continuously revised "Lists" of the works of Italian painters of the Renaissance, those indispensable manuals used by generations of students and art historians. As a result, most of the attributions recorded in the dataset rely on such lists as primary sources of information. The dataset includes 12.256 cataloguing records of unique artworks providing information on the accepted authorship attribution. Among these, 5.384 records include information on

[2] `https://itatti.harvard.edu/berenson-library/collections/photograph-archives/homeless-paintings`

alternative attributions. Not all the criteria adopted by the Zeri Foundation are used in the context of Villa I Tatti. Stylistic analysis, traditional attribution, and false signature are not included.

Data is provided as a unique .csv file including the complete set of records. Three fields were useful for the sake of the survey, namely: ID of the artwork, name of the artist, the descriptive note, which includes both accepted and discarded attributions. Data are transformed into RDF according to CIDOC-CRM when representing the creation of the artwork, and according to the HiCO Ontology when describing attributions.[3]

In Figure 6.3 is illustrated the distribution of criteria and the comparison of criteria supporting accepted and alternative attributions.

Like in the Zeri photo archive, preferences reflect peculiarities of the photo archive, namely: (1) an extensive usage of "archival classification", (2) the influence of scholars' opinions emerged from the appraisal of photographic documentation recording many annotations, and (3) the usage of bibliography, especially Berenson's references.

The archival classification affects most of the decisions taken at Villa I Tatti. Figure 6.4 shows the interaction between the "archival classification" criterion and others that are used along with it to support accepted attributions. As shown in the picture, only few times (out of 680) the criterion archival classification is accompanied by other criteria explaining how the attribution was chosen. We assume the cataloguing process is biased by the archive creator's opinion, which would require more investigation.

The archive creator is mostly cited as author of a bibliographic references (e.g. the aforementioned lists) or as author of annotations on the back of the photographs. Official statements (e.g. verbal communications) do not apply in this case (i.e. it is cited only once). Berenson often recorded a number of alternative attributions on the back of photographs, which were examined by cataloguers and only the last accepted attribution was recorded. Therefore, we found a high number of both accepted and discarded attributions referencing Berenson's notes as main reason of attribution (256 times).

Like in the Zeri photo archive, the archive creator's opinion appears more reliable than other scholars' attributions and notes on the photographs. In particular, the archive creator's bibliography is always pre-

[3]Data sources are available at `https://github.com/marilenadaquino/mauth/tree/master/data/itatti`

tot.	ACCEPTED \ DISCARDED	documentation	artist's signature	archive creator's attrib.	archive creator's bibl.	bibliography	archival classification	archive creator's note	scholar's attribution	scholar's note on photo	museum attribution	inscription	sigla	auction attribution	collection attribution	market attribution	traditional attribution	stylistic analysis	anonymous note on photo	false signature	caption on photo	other	none
	tot.	39	0	20	279	290	4	710	483	396	19	23	0	202	14	1	0	0	422		8	7	0
3	**documentation**					1		1	1										1				
12	**artist's signature**							8	1			1							3				
1	**archival creator's attrib.**																		1				
368	**archival creator's bibl.**	1		3	72	118		130	93	53	2	3		27	5				52		2	3	
284	**bibliography**	3		6	133	36		111	48	39	1			19	3				14			1	
680	**archival classification**	17		4	8	92		185	177	131	3	13		73	3	1			184		2	2	
514	**archive creator's note**	11		6	3	36	1	256	92	110	11	2		44	5				104		1	1	
61	**scholar's attribution**				11	10	2	18	6	15				5					15			1	
194	**scholar's note on photo**	5		1	3	17	1	59	51	70	1	1		22	2				37		1		
93	**museum attribution**			2	29	4		50	19	25		1		3					9		1		
0	**inscription**																						
0	**sigla**																						
95	**auction attribution**	1		1	13	3		42	17	12		3		15	1				13				
11	**collection attribution**				5			7						2					2				
0	**market attribution**																						
0	**traditional attribution**																						
0	**stylistic analysis**																						
63	**anonymous note on photo**	1			2	4		13	19	11	1			7					27		1		
0	**false signature**																						
0	**caption on photo**																						
0	**other**																						
0	**none**																						

Figure 6.3: Distribution and comparison of criteria adopted by Villa I Tatti photo archive

ferred over scholars' notes on the photo, and it is preferred 93 times over "scholar's attribution", while it is discarded only 11 times. Similarly, "archive creator's note on photo" is preferred 92 times over "scholar's attribution" (discarded 18 times), and it is preferred 110 times over "scholar's note on photo" (discarded 59 times). Lastly, the usage of Berenson's bibliography over other references seems balanced, since it is preferred 118 times and discarded 133 times. In such cases we assume more recent references are taken into account to update attributions stated by the archive creator in older works.

Other less cited criteria do not provide more insights on the actual preferences since these are either not well-represented in the dataset, or are completely absent.

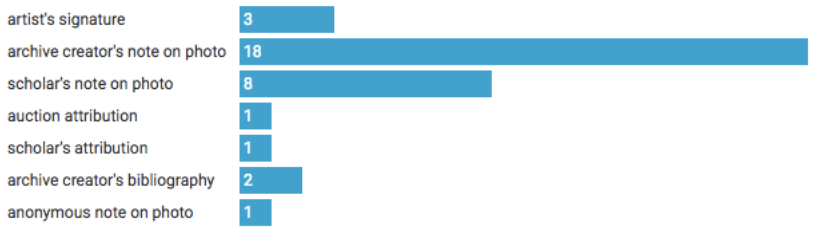


Figure 6.4: Distribution of criteria that appear along with "archival classification" in Villa I Tatti photo archive

In conclusion, the core of criteria characterising the methodology of photo archives seems to be shared between Zeri and I Tatti. The actual usage of criteria confirms the prior rating, but no further information can be deduced on other criteria.

The Frick Art Reference Library. The third dataset is provided by the Frick Art Reference Library of New York[4]. It includes about 10.207 records mainly referring to artworks that have not been ascribed to an artist yet. Among these, 941 records include also discarded attributions. For the artworks whose authorship is still debated, the dataset includes a rich literature of alternative attributions.

Since the Frick Art Reference Library is not a personal archive, few of the above listed criteria do not apply to the definition of a shared methodology, namely: "archive creator's attribution", "archive creator's bibliography", and "archive creator's note on photograph". However, the two aforementioned archive creators, Federico Zeri and Bernard Berenson, are often cited as sources of attribution, here mentioned as scholars rather than as archive creators. We look into such third party dataset to understand whether archive creators can be deemed more authoritative sources of attribution when their opinion is compared to other scholars' opinions or criteria.

Data is provided as a unique .csv file including the complete set of records. Four fields were useful for the sake of the survey, namely: ID of the artwork, name of the artist, descriptive note including both accepted and discarded attributions, and sources of information (including bibliography and verbal opinions). Data are transformed into RDF according to CIDOC-CRM when representing the creation of the artwork, and according to the HiCO Ontology when describing attributions.[5]

Figure 6.5 shows the distribution of criteria and their comparison when supporting accepted and alterna-

[4] `https://www.frick.org/research/library`
[5] Data sources are available at `https://github.com/marilenadaquino/mauth/tree/master/data/frick`

tot.	ACCEPTED \ DISCARDED	documentation	artist's signature	archive creator's bibl.	archive creator's attrib.	bibliography	archival classification	archive creator's note	scholar's attribution	scholar's note on photo	museum attribution	inscription	sigla	auction attribution	collection attribution	market attribution	traditional attribution	stylistic analysis	anonymous note on photo	false signature	caption on photo	other	none
	tot.	1	1	0	0	121	123	0	561	18	73	2	0	40	12	0	7	0	39	0	1	3	115
1	**documentation**	1																					
0	**artist's signature**																						
0	**archival creator's attrib.**																						
0	**archival creator's bibl.**																						
434	**bibliography**					25	13		298	2	32				5				15				44
375	**archival classification**		1			41	60		142	6	24	2		25	2		3		7			2	60
0	**archival creator's note**																						
167	**scholar's attribution**					18	29		86	2	7			7	3		2		9			1	3
7	**scholar's note on photo**					1			3	3													
37	**museum attribution**						12		11		3	1		1	1		2						6
0	**inscription**																						
0	**sigla**																						
6	**auction attribution**													2	1				1				2
4	**collection attribution**						1				2			1									
0	**market attribution**																						
0	**traditional attribution**																						
15	**stylistic analysis**					2			7	4				2									
69	**anonymous note on photo**					34	6		13	1	5			2	1				7				
0	**false signature**																						
0	**caption on photo**																						
2	**other**						2																
0	**none**																						

Figure 6.5: Distribution and comparison of criteria adopted by Frick photo archive

tive attributions.

Like in Villa I Tatti, the low number of available comparisons of some criteria, e.g. scholar's note on photograph, auction attribution, market attribution, and so on, does not allow to extract information useful to revise the rating. Despite this, many similarities in the usage of highly rated criteria are found between the Frick dataset and the former ones. In particular, the usage of bibliography is consistent with the original rating, especially when compared to a scholar's attribution (accepted 298 times and discarded only 18 times). The extensive usage of the Gernsheim Corpus Photographicum (which is cited about 4.000 times as main source of information regardless discarded attributions are recorded) confirms

the usefulness of photographic catalogues as fundamental sources of information in connoisseurship - as discussed in Chapter I.

The criterion "archival classification" is consistently used when compared to lower rated criteria, while it is less consistent when compared to bibliography (accepted 41 times and discarded 13 times).

Figure 6.6: Distribution of criteria that appear along with "archival classification" at Frick photo archive

Figure 6.6 shows the interaction of the criterion archival classification with other supporting criteria. We deduce that when alternative attributions are recorded, the archival classification is rarely supported by other criteria. Only in few cases cataloguers' expertises on the artwork at hand (13 times out of 301), and few other sources of information (16 times out of 301) are adopted to support statements. This scenario confirms the predominant role of the decisions taken by cataloguers during the cataloguing process.

It's worth to notice that several updates in the cataloguing process are recorded by cataloguers. Indeed, when new attributions revising prior ones are recorded by cataloguers (described by the criterion "archival classification"), the prior attribution is in turn recorded among the discarded ones (60 times), so as to preserve the history of attributions created by the institution.

As aforementioned, archive creators are here referenced as sources of attributions. Federico Zeri is mentioned 115 times as source of information of an accepted attribution, while his opinion is discarded only 14 times. Among the reasons for not being chosen, we found his opinion was discarded 6 times because of a more recent revision of the cataloguing process lead to a new attribution (i.e. "archival classification"): 5 times because of a more recent scholar's attribution; 2 times a museum attribution is preferred; and once because of a contradictory annotation on a photograph is preferred. Instead, Bernard Berenson's attributions are accepted 22 times, and discarded 41 times. 16 times a more recent cataloguing intervention ("archival classification") revises older attributions. 21 times a more recent scholars' attribution is chosen instead. Likewise, 2 museums attributions, 2 annotations on photographs, and 1 stylistic analysis are preferred.

We can assume that the three biased criteria introduced in the original rating apply when the archive

creator's attribution is more recent than the other available attributions. According to the data analysis in the Frick dataset, the revised rating is likely to be valid when applied to Zeri's attributions, whose attributions are generally more recent, and less valid when applied to Berenson's attributions. We present the final rating in the next section along with the other measures.

S6. Comparison of the three photo archive methodologies. The analysis performed on contradictory statements is useful to understand how to rate the reliability of criteria when alternatives are available. However, the distribution of criteria chosen by data providers regardless competing attributions are recorded provides insights on the archival policies and the flaws in the publication of art historical data. The aim of this comparison is to highlight which features can be always deemed valid, say *a priori*, and which would need instead a continuous assessment so as to ensure the textual authoritativeness of attributions.

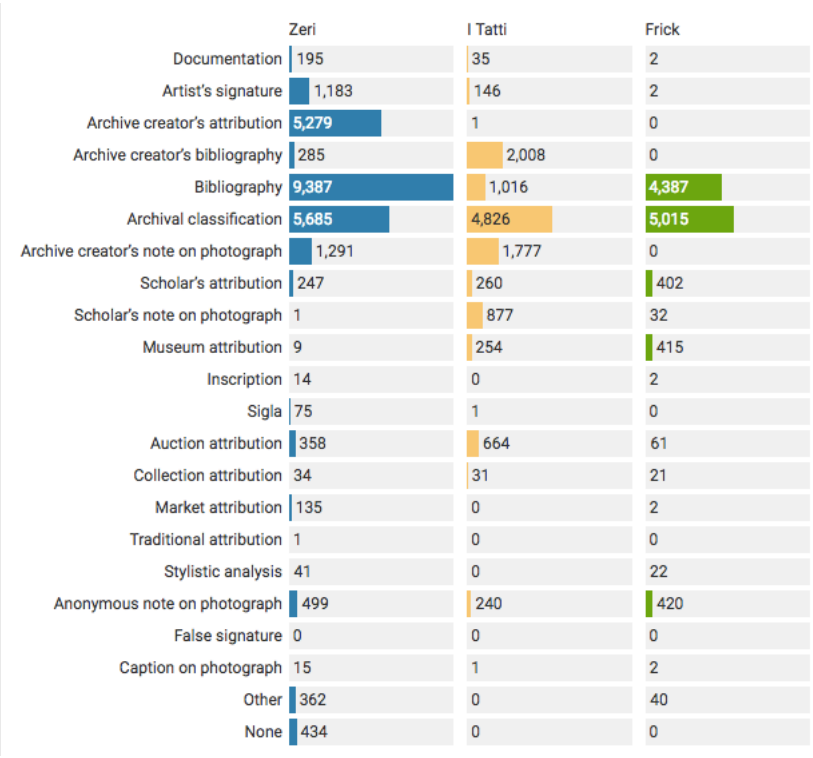


Figure 6.7: Distribution of criteria in Zeri, I Tatti, and Frick photo archives

In Figure 6.7 is illustrated the distribution of criteria adopted by the three photo archives when supporting accepted attributions, regardless alternative attributions are recorded. The scenario here presented does not differ significantly from the ones highlighted in prior comparative tables, meaning that proportions in the usage of criteria do not differ significantly whether the history of attributions is recorded or not. The criterion archival classification is the most used in all of the three archives, showing how the cataloguing

process significantly affects the decision made on attributionship. Although in some cases it is accompanied by other concurring criteria when supporting an attribution, the number of records that do not present sources of information is still high, especially in I Tatti and Frick.

Secondly, bibliography appears to be the main tool of cataloguers in most of the cases. Although the criterion could be considered valid a priori, an accurate analysis on cited authors and the validity of references over time would deserve more attention. In particular, the Zeri photo archive mainly privileges bibliographic references not belonging to Zeri (9.387), that are part of the art library. Villa I Tatti mainly relies on Berenson's published lists (2.008), which may be less recent, and external bibliographic references (1.016) which are likely to be more recent. Frick relies on heterogeneous bibliographic sources including photographic catalogues (4.387), museum and exhibition catalogues (415), and auction catalogues (61).

As already mentioned several times, the third aspect characterising photo archives is the impressive number of scholars' attributions gathered by cataloguers, either as official statements or as notes recorded on the back of photographs. Federico Zeri's official attributions selected by the Zeri photo archive amount to 5.279, followed by his 1.291 notes. Moreover, 247 scholars' attributions and 499 not identified annotations suggest the importance of the Zeri photo archive as a research centre for art historians. Likewise, I Tatti recorded around 1.777 Berenson's notes on photographs, 260 scholars' official attributions, 877 scholars' notes, and 240 anonymous notes on the same photographs. Lastly, the Frick provides 402 scholars' attributions, 32 scholars' notes, and 420 not identified notes on photographs.

Other sources of information are more or less similarly represented in the three datasets, even though some archives may prefer certain types of sources rather than others. For instance, museums attributions are well represented in I Tatti (254) and Frick (415), while are underrepresented in Zeri (9). In turn, Zeri relies on market attributions (135), which seem to be almost absent in the other two datasets. Auction attributions are mainly cited by I Tatti (664) and Zeri (358), and less in Frick (61). It is worth to notice that auction attributions are unanimously mostly cited in discarded attributions.

In conclusion, we assume the proposed rating of criteria is valid over the three photo archives, except for the three biased criteria that apply to personal photo archives only. Therefore it can be proposed as a minimum common denominator when comparing contradictory attributions. Lastly, the representativeness of the Zeri photo archive use case for performing the preliminary analysis on the methodology of photo archives

is confirmed. Indeed, it presents the highest number of alternative attributions on which the rating is based, the highest number of cited criteria (both in accepted and discarded attributions), and it is consistent with the final rating.

6.3 Dimensions and measures for evaluating textual authoritativeness

In Chapter II we listed dimensions of Information Quality (IQ) on the web and metrics for assessing data quality. According to Naumann and Rolker [Naumann and Rolker, 2005], dimensions of IQ can be grouped in Subject Criteria (features characterising users' expectations), Object Criteria (features characterising internal grounds of sources), and Process Criteria (features of the retrieval process).

Both domain-dependent and domain-independent dimensions apply to the Arts field. In Chapter II we highlighted a number of dimensions that affect the consistency of the above described rating of criteria, which is not sufficient alone to define textual authoritativeness alone. The aim is to compensate potential inconsistencies in the rating of criteria by adding other variables taken into account when assessing the most authoritative source and refine a ranking model of contradictory attributions accordingly.

S7. Dimensions and metrics. Table 6.4 shows a selection of the dimensions described by Naumann and Rolker [Naumann and Rolker, 2005] that are implemented in this study. In the first column is listed the grouping class of dimensions. In the second column the selected dimensions are listed, according to their original definition provided in [Naumann and Rolker, 2005]. In the third column the assessment methods are defined. The original dimensions are pruned so as to include only the ones that apply to the arts domain, which is characterised by a high degree of uncertainty of both Subject and Object Criteria. The selection is made according to online guidelines [Baca and Harpring, 2009], domain experts' consultancy, and aspects highlighted by the corpus analysis. The selection includes only the dimensions useful to address aspects that characterise questionable information in connoisseurship.

Subject Criteria assessment. Subject Criteria are generally hard to be assessed, since the assessment is accurate only for individual groups. However, metrics for measuring relevance and reputation help to

IQ Group	IQ Measure	Assessment Method
Subject criteria	Relevance	Counting of agreements
	Reputation	List of trusted data providers, citation indexes
Object criteria	Reliability	Rating of criteria
	Timeliness	Distance between date of attribution and retrieval date
Process Criteria	Accuracy	Cleansing techniques, user assessment
	Amount of data	Continuous assessment

Table 6.4: Dimensions of IQ in the arts field and related metrics

overcome uncertainty in the assessment.

Relevance. Relevance is the extent to which information is applicable and helpful for the task at hand. For the sake of the study here proposed, we ensure relevance of results of a research is always respected by relying on a list of trusted data providers, which are likely to include the information sought. Secondly, in order to verify that attributions retrieved are helpful, we group them by ascribed artists, and the counting of sources in agreement is calculated, so as to show how each single data provider contributes to define the acceptance of the selected attribution.

Reputation. Reputation is the extent to which information is highly regarded in terms of source or content. We distinguish providers' reputation from cited scholars' reputation. Reputation of data providers can be evaluated by relying on third party opinions, and by annotating the aforementioned list of trusted providers with a rating of these. The bucket of selected data sources useful to validate this study includes three photo archives, i.e. the Zeri Photo Archive, Villa I Tatti - Berenson Library, and the Frick Art Reference Library, and three multipurpose datasets, namely: Wikidata, DBpedia, and VIAF. The selection includes multipurpose datasets that can support the assessment of authoritative attributions when no sufficient authoritative sources are available. Providers are flagged with a label describing those as domain experts or not. In this study photo archives are flagged as domain experts, while multipurpose datasets are not.

When dealing with reputation of cited historians that ascribed the artwork first, two bespoke indexes are associated to the attribution, namely the *artist-related index*, and the *acceptance-rating* of the scholar (S7.1).

The artist-related index. The artist-related index is inspired by the well-known *h-index* metric. H-index is a metric that uses the number of an author's publications along with the number of times those publications have been cited by other authors in an attempt to gauge an author's perceived academic authority in their fields of research [Mitchell et al., 2011].

Since the Arts field is not a bibliometric research field, the h-index of most scholars is not available. Secondly, most of the art historians that worked on Modern Era belong to the first half of the 20th century, and most of their works are not indexed in any citation index. Moreover, scholars' attributions are acknowledged in cataloguing records in many ways other than listing their bibliographic references. Such types of citation include "verbal communication" or "note on the verso of the photograph", which are hard to be linked to the correct bibliographic reference (if it exists). Hence, the h-index should apply to all of these forms of citation too.

To overcome such issues, we assume that (i) the number of publications can be substituted by the number of artists addressed by the art historian in the course of her/his activity, and (ii) the number of citations can be addressed by the number of times the scholar is cited with regard to artworks of a certain artist. This assumption is motivated by a common belief in the Arts field, where a historian that worked on several artists is likely to be an expert of the period at hand. In order to apply a citation-based metric to art historians, the following parameters are taken into account:

- The number of artists to whom the scholar ascribed some artworks. The number includes all the artists retrieved in the three photo archives, whose artworks where ascribed by the scholar at least once, and therefore s/he is cited as preferred source of information (discarded attributions that cite the scholar are not taken into account).
- For each artist, the number of the scholar's attributions correspond to his citations, i.e. the number of artworks that the scholar ascribed to the artist. The number includes all the scholar's accepted attributions retrieved in the three photo archives.

The following listing (in Python) exemplifies how to calculate the artist-related index given a list of citations counts:

```
def artist_related_index(citationsArray):
        n = len(citationsArray);
        count = [0] * (n + 1)
        for x in citationsArray:
                if x >= n:
                        count[n] += 1
                else:
                        count[x] += 1
```

```
ar_index = 0
for i in reversed(xrange(0, n + 1)):
        ar_index += count[i]
        if ar_index >= i:
                return i
return ar_index
```

For instance, in the course of his activities, Bernard Berenson ascribed artworks to 8 artists. For each of these artists he has been cited as favourite source of attribution respectively 10, 9, 9, 8, 8, 3, 2, 1 times. In details, he has been cited 10 times for having ascribed 10 different artworks to the first artist, 9 times for 9 different artworks to the second artist, and so on. His artist-related index is 5, because he has been cited at least five times with regard to 5 artists.

The limits of this metric are evident. The number data sources is limited to the three photo archives, and multipurpose datasets do not contribute to increment the grounds of this metric, since references to scholars are never included in data. Hence we cannot have a clear overview of the actual number of citations of the scholar, and a traditional h-index would not be representative (i.e. it would never be greater than 3). However, assuming artists studied by a scholar are representative samples of scholars' activity, some conclusions can be drawn on the academic reputation.

Secondly, the number of citations depends on the number of artworks that the considered artists have actually created. This means that historians who focused on minor artists, who are likely to have produced a lower number of artworks, will be characterised by a lower number of possible citations. The limited number of sources and limited number of potential citations affect the perception of the scholar, whose quality of scientific production cannot be deemed less relevant than the quality of scholars' scientific production that ascribed a greater number of artworks. However, being the scope of the three photo archives narrow, i.e. artworks of the Modern Art, and assuming historians focused on same artists belonging to the same period or schools, we expect that the number of potential attributions would not differ significantly between scholars, hence the indexes will be based on a similar number of possible tuples artists/artworks.

In order to compensate the potentially misleading perception of scholars' authoritativeness, another metric is defined to shed light on providers' perception of scholars, i.e. the acceptance-rating.

The acceptance-rating. The acceptance-rating is a simple and flexible measure that uses the number of

a scholar's accepted attributions with regard to a certain artist, along with the total number of possible attributions for that artist (i.e. the total number of artworks ascribed to the artist that are surveyed in the three photo archives). Precisely, given a list of tuples `(historian, artist)` the rating is calculated for each tuple as the proportion between the number of scholar's citations for that artist over the three photo archives `(numberOfCitations)` and the number of artworks that are ascribed to the latter in the three photo archives `(totalNumberOfArtworks)`. The following listing shows how the proportion is calculated (in percentage):

```
a_index = numberOfCitations / totalNumberOfArtworks * 100
```

For instance, Bernard Berenson has been cited 10 times with regard to Titian's artworks (i.e. 10 of his attributions were accepted by data providers). The three photo archives surveyed 20 Titian's artworks. The acceptance-rating of Bernard Berenson's attributions with regard to Titian is 50%.

Like the artist-related index, the acceptance rating presents some limits. In fact, not only scholars are cited as primary sources of attributions. Other criteria than scholars' attributions (described in the next section) can support the choice of an authorship attribution, e.g. captions or anonymous notes on the photograph of the artwork, auction attributions. Therefore a low score may not be representative all the times. This is the case of artists that are not well-known and studied, hence the literature is scarce and other less reliable criteria are adopted by data providers to ascribe the artworks than scholars attributions.

Object Criteria assessment. Object criteria can be mostly assessed by using automatic methods.

Reliability. Reliability is the extent to which information is correct and trusty. The rating of criteria supporting an authorship attribution is the main tool for assessing reliability of the information. In particular, twenty two shared criteria were extracted from the three surveyed photo archives catalogues, and sorted according to their relevance. Table 6.5 includes terms of the controlled vocabulary of criteria and their related score (from 10 to 1).

The rating is defined according to domain experts' opinion, who provided an initial rating secondly validated on the actual usage of criteria in the three photo archives. The rating has been normalised between 1 and 10 to balance its importance among the set of dimensions here defined. Nonetheless this dimension is the one that mostly affects the final ranking of authorship attributions.

N.	Term	Score
1	Documentation	10
2	Artist's signature	10
3	Archive creator's attribution	9
4	Archive creator's bibliography	8
5	Bibliography	7
6	Archival classification	7
7	Archive creator's note on photograph	7
8	Scholar's attribution	6
9	Museum attribution	5
10	Scholar's note on photograph	5
11	Inscription	5
12	Sigla	5
13	Auction attribution	4
14	Collection attribution	4
15	Market attribution	4
16	Traditional attribution	4
17	Stylistic analysis	3
18	Anonymous note on photograph	3
19	False signature	2
20	Caption on photograph	2
21	Other	2
22	None	1

Table 6.5: The controlled vocabulary of criteria and the rating

Timeliness. Timeliness is the distance between the date of the information and the retrieval date. A common belief in the Arts domain is that the latest recorded attribution - assuming it is also well-documented - is likely to be the most reliable. Hence, timeliness is calculated by comparing available dates of retrieved attributions and scoring those in descending order.

Process Criteria assessment. Process Criteria regard all the features related to the data integration process. Several degrees of accuracy may affect final results of a research. Hence, for the sake of the proof of concept, semi-automatic methods are mostly applied to ensure a high rate of confidence in retrieved information.

Accuracy. Accuracy is the extent to which data are correct, reliable and certified free of error. Cleansing methods and manual double checks are applied to eliminate a variety of data errors that may occur in the integration of data sources, especially in data reconciliation. While images depicting artworks have been matched by means of image similarity tool Pastec, URIs identifying artists and scholars mentioned in several datasets have been matched automatically and manually double-checked by using methods detailed in Chapter IV.

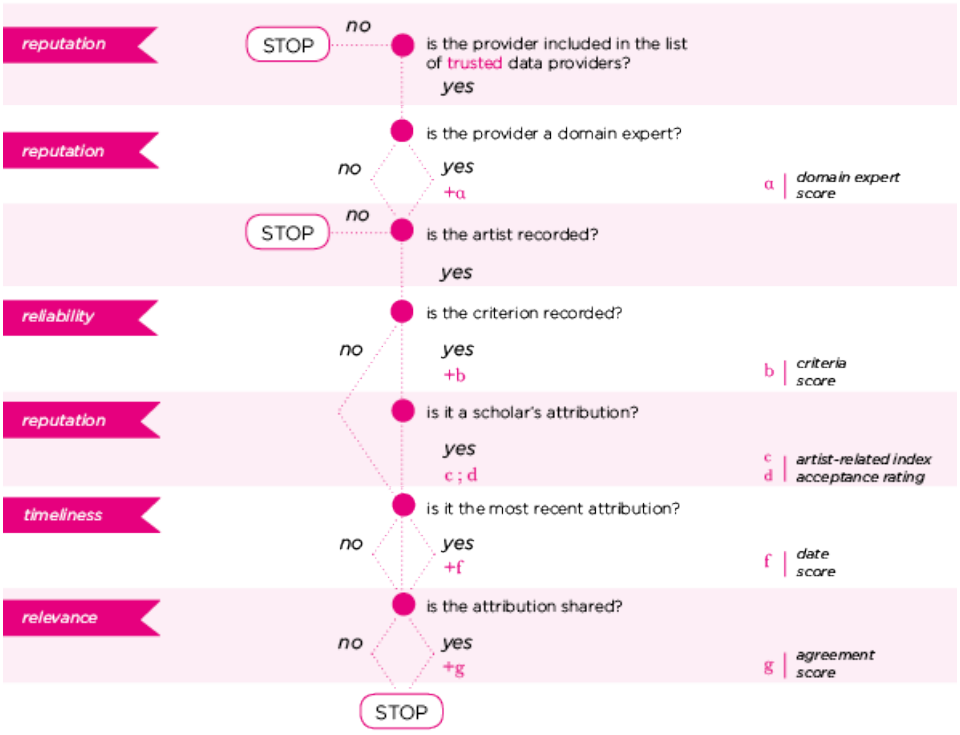


Figure 6.8: The ranking model for textual authoritativeness of authorship attributions

Amount of data. Amount of data indicates the extent to which the quantity or volume of available data is appropriate. As aforementioned, at the moment the collection of data sources includes a restricted number of six providers. These are selected on (1) a topic base, i.e. they provide attributions of artworks of Modern Era, (2) because of the richness of metadata (this apply to photo archive catalogues mainly), and (3) because of the social acceptance of data recorded (which applies to multipurpose datasets). Assuming the number of sources increases over time, a continuous assessment is required to revise other measures that highly depend on the amount of data (i.e. artist-related index and acceptance rating).

6.4 The ranking model for art historical data sources

Figure 6.8 summarizes how the described dimensions characterising authoritativeness of authorship attributions interact with each other in the final ranking model (S8).

The ranking model elaborates a number of steps and incrementally associates a score to attributions recorded in data sources. It accepts in input the URI identifying an artwork, for which the history of authorship attributions is harvested in a number of data sources. The input URI belongs to one of the six

data providers listed in the list of trusted data providers, namely:

- The Zeri photo archive (University of Bologna)
- Villa I Tatti - Berenson Library (University of Harvard)
- The Frick Art Reference Library
- DBpedia (in particular, the italian, english, and french datasets)
- Wikidata
- VIAF

In the first step reputation of potential data providers is evaluated. An algorithm developed for the sake of the proof of concept (detailed in Chapter VII) looks at the URI base to recognise the data provider. If the URI base is included in the list of trusted data providers the process proceeds, otherwise it stops, i.e. the data provider is not included in the set of results.

Secondly, if the data provider at hand is a domain expert, the *score a* is added, otherwise the process moves to the next step. Data providers included in the list of trusted providers are flagged as being domain experts or not. Such a decision is based on a third-party opinion.

In the third step harvested data sources are queried to see whether these include statements on the authorship. If so, the process moves to the next step, otherwise it stops and no further information are harvested.

In the fourth step statements on the criteria supporting the attribution are parsed. If one or more terms included in the vocabulary of criteria are retrieved, the *score b* is assigned, otherwise a default score is assigned.

In the fifth step, if the motivation supporting the attribution includes the name of a scholar, values of two metrics are calculated, namely: artist-related index - *c score*, and acceptance rating - *d score*. As aforementioned, such indexes do not increment the final score associated to the attribution, while the information on the cognitive authoritativeness is provided as an aid for the final user.

In the sixth and seventh steps, a comparison of retrieved information sources is performed. Dates of attributions - whether years are explicitly recorded in data or these can be deduced by context information - are parsed, and attributions are sorted in descending order. The *score f* is assigned to the date, and it increments proportionally according to the timeliness of the attribution. Lastly, attributions are grouped by artist, and the *score g* of each attribution increments according to the number of agreements found on the same artist.

The final score is calculated as the sum of the prior partial scores and it is associated to each retrieved attribution. Precisely, a score is associated to each attribution recorded in the information source, including also discarded attributions, so as to judge the textual authoritativeness of the record or web page. Finally, sources are sorted in descending order according to the calculated score.

Scores and ranges. To compute the aforementioned partial scores and the final score associated to information sources, different units of measure apply, and partial scores lie on different ranges of values.

Subject Criteria and Object Criteria can be automatically assessed by relying on definitions of scores and weights. These scores influence the sorting of results. However, Subject Criteria and Process Criteria highly depend on the user's perception and experience, which can be assessed by means of a user-centered evaluation. This intervention is discussed in Chapter 8, where we evaluate the soundness of the conceptual framework.

Table 6.6 shows measures, associated score, and range of values that are implemented in the ranking model.

IQ Group	IQ Measure	Score	Range
Subject criteria	Relevance	agreement (*g*)	[0-(n-1)] where *n* is the total number of retrieved information sources
	Reputation	domain expert (*a*)	[0 or 1] boolean
Object criteria	Reliability	criteria (*b*)	[1-10]*n where *n* is the total number of recorded criteria
	Timeliness	date (*f*)	Range: [0-1]

Table 6.6: IQ dimensions, scores and ranges

The dimensions that affect the ranking of authorship attributions are: Relevance (agreement score), Reputation (domain expert score), Reliability (criteria score), and Timeliness (date score).

Relevance is addressed by relying on a list of a fixed number of trusted data providers. Once attributions

are retrieved these are grouped by the selected artist. The *agreement score (g)* is measured according to the total number of providers minus the selected source, i.e. having six data providers, the range is between 1 and 5.

Likewise, reputation relies on the same list of providers, in which providers are flagged as being domain experts or not. The *domain expert score (a)* is either 1 when the provider is a domain expert, or 0 if it is not a domain expert. The score is intentionally low so as to not penalise less scholarly sources, such as DBpedia, Wikidata, and VIAF, and balance the final score with other partial scores, such as the number of agreements itself and the timeliness.

Reliability is measured by relying on the rating of criteria that support the attribution. According to domain experts' opinions, the *criteria score (b)* is the one that mostly affects the ranking of results of a research, hence must weight significantly more than the others in scope. Precisely, when more than one criterion is provided in the information source to support the selected artist, the reliability score increments accordingly, i.e. different criteria supporting the same attribution are treated as concurring attributions. The score is defined as the the sum of all the scores associated to each single criterion. However, only the criteria score increments accordingly, while the domain expert score does not apply to cited primary sources. For instance, if a scholar is cited as source of attribution, the domain expert score of the information source does not increment nor decrease.

Finally, timeliness is measured by the *date score (f)*, that is obtained by comparing the dates of retrieved attributions. The most recent attribution is scored 1, while attributions missing the date are scored 0. Scores for other less recent attributions are sorted in descending order and scored accordingly. The score is normalised between 1 and 0 so as to compensate criteria with a lower rating, e.g. scholar attributions, that may be offer more relevant attributions whether these are updated.

As an example, the following scenario illustrates how the ranking is affected by the so weighted scores. Figure 6.9 shows a debated artwork described at the Zeri photo archive.[6]

Authorship attributions about the artwork are recorded in the Zeri photo archive and in Villa I Tatti online catalogues. Both of the providers are deemed domain experts (a=1 in both cases). The two attributions are in disagreement (g=0 in both cases): Zeri records Granacci Francesco, while Villa I Tatti records Mainardi

[6]See the cataloguing record at `https://tinyurl.com/yb5yg8yy`

Figure 6.9: A debated artwork described at the Zeri photo archive catalogue

Bastiano. Zeri motivates the attribution with a scholar's attribution made by Everett Fahy (b=6), and with the archival classification of the photograph (b=7, that summed to the former gives in total b=13). Villa I Tatti relies only on the archival classification of the photograph (b=7). Villa I Tatti does not provide a date for the attribution (f=0), while the archival classification at the Zeri Foundation happened around 1990 (f=1). In conclusion, I Tatti's attribution is scored 8, and Zeri's attribution is scored 15. The latter results more documented than the former, and relies on two concurring and more recent attributions.

6.5 Strategies for data quality assessment and improvement in art historical photo archives

The comparative analysis shows trends in archival policies adopted by the three surveyed photo archives and the ranking model offers a means to evaluate accuracy of cataloguing processes. From the analysis results that "archival classification" is the most recurrent and influential criterion. This value demonstrates the high subjectivity of statements resulted from the cataloguing process. Although the three providers are deemed domain experts - hence their statements benefit of their cognitive authoritativeness - the lack of well-documented and researched data - say the internal grounds that define textual authoritativeness of their statements - may affect their reliability in the long term. In other terms, if cataloguing data have be treated as research data, data must comply with data quality requirements shared in the research community.

Secondly, bibliography is the second most used criterion by the three archives (including archive creators' bibliography). In this case, an in depth analysis on the cognitive authoritativeness related to authors would deserve attention. However, citation-based analyses are mainly related to scientific research products, and lack in the Arts and Humanities field. Hence, at this stage, it is not possible to track the evolution of scholars' authoritativeness over time. Moreover, bibliography is likely to be often out of date, and updates in cataloguing records are required continuously. While the validity of the criterion itself is confirmed by the effective usage in three real scenarios, statements based on it may be contradicted by more recent attributions, supported by even less reliable criteria.

Likewise, scholars, museums, collection, and auctions' attributions are likely to be overcome by new attributions and new researches. The high questionability of such types of statements can only partially be addressed, since the goodness of statements can only be evaluated by domain experts. While the automatisation of such aspects is out of scope in this research, addressing and improving textual authoritativeness of secondary sources is deemed a mandatory requirement for data providers that want to provide high quality art historical research data.

Activities underpinning the cataloguing process are expensive, time-consuming, and a continuous data (and information) quality assessment is required. Many strategies for data quality improvement have been defined in the literature [Batini et al., 2009]. In particular, interorganizational cooperation systems [Scannapieco, 2006] aim to leverage the potential of external data sources for improving data quality of single partners. The literature provides a wide range of techniques to assess and improve the quality of data, such as record linkage, business rules, and similarity measures. Due to the diversity and complexity of these techniques, research has recently focused on defining methodologies that help to select, customize, and apply data quality assessment and improvement techniques.

However, data quality assessment and improvement is not easy to be integrated in cultural institutions daily practices, because of the aforementioned barriers (time, resources, and accuracy of research). Among the hypotheses of this research is that Linked Open Data can help to lower barriers derived from the expensive and time-consuming cataloguing activities by providing means for integrating missing or partial information about a subject at hand (i.e. *H6. Linked Open Data and Semantic Web technologies can support and satisfy common requirements of research activities in the Arts and Humanities*). For instance, the analysis

performed on the Zeri dataset revealed an overlap of around 940 cataloguing records with Villa I Tatti, which could be easily integrated. Likewise, Zeri shares around 175 records with Wikidata, around 400 records with VIAF, and 412 with DBpedia, which can provide contextual information not recorded in the Zeri dataset, e.g. information on artists (VIAF), additional - more recent - pictures of the artworks (DBpedia), and identifiers of the artwork in other relevant institutions (Wikidata).

Methodologies for data quality improvement generally adopt two types of strategies, namely data-driven and process-driven. Data-driven strategies improve the quality of data by directly modifying the value of data, e.g. replacing obsolete data values by refreshing a dataset with data from a more updated source. Process-driven strategies improve quality by redesigning the processes that create or modify data. For example, a process can be designed to include new information acquired from external sources when a defined threshold of data quality is not exceeded. The latter approach is the one here proposed for improving data quality of cataloguing records including information on authorship attributions. Likewise, museums, galleries, and multipurpose datasets could potentially benefit of photo archives information and integrate motivations and sources of information when recording authorship attributions. Indeed, the latter providers rarely include information on motivations and primary sources used to assess the veracity of an authorship attribution.

Strategies for data quality improvement apply a variety of techniques, such as algorithms, heuristics, and knowledge-based activities. Here we propose a methodology that relies on the developed methods, whose implementation is detailed in Chapter VII, for harvesting authorship attributions and achieve the following objectives:

- Integrate cataloguing records with supporting motivations and sources of information retrieved in cataloguing records belonging to other photo archives.
- Extend the history of attributions related to artworks with other attributions available in existing datasets, including both accepted and discarded attributions.
- Support cataloguers in the decision-making process, who may, eventually, revise their attributions on the basis of the extended literature retrieved during the integration process.

The strategy is based on existing surveyed strategies [Batini et al., 2009], and is tailored on the Federico

Zeri photo archive use case, but it consistently applies to other art historical photo archives. The strategy includes the following list of steps:

- **Acquisition of new data.** New datasets including high-quality data on the same subject are acquired to integrate or eventually replace the values that raise quality problems.
- **Standardization (or normalization).** New data sources which present nonstandard data values are normalised with corresponding values that comply with the shared standard. For example, terms of the controlled vocabulary of criteria supporting attributions are associated to every authorship attribution retrieved in the new data sources.
- **Record linkage and data reconciliation.** Data on real-world objects described in multiple datasets are reconciled so as to allow comparison and integration. In particular artworks, artists, organisations (including museums and auction firms), and scholars are reconciled to the same authority files, i.e. VIAF, DBpedia, and Wikidata, and then cross-linked.
- **Data and schema integration.** The integration defines a unified view of the data provided by heterogeneous data sources. To overcome technological interoperability problems all data are transformed into Linked Open Data and stored in the same triplestore, in bespoke named graphs. To overcome issues related to semantic interoperability, data transformed into RDF for the sake of the integration is represented according to the same ontologies, namely CIDOC-CRM and HiCO Ontology. Other Linked datasets that are already adopting existing vocabularies are mapped to CIDOC-CRM and HiCO ontology by means of an extendible mapping document. Data heterogeneity at instance-level is overcome by creating bespoke linksets, including only similarity links between respectively artists, artworks, organizations, and scholars. Linksets are stored along with data in bespoke graphs. The integration process is annotated according to the PROV ontology and stored in another named graph. The latter includes information on the retrieval of information sources addressing the description of the same artwork.
- **Source trustworthiness assessment and ranking.** The ranking model described in Section 6.4 allows to select which data sources offer the higher quality data contents to integrate in the original dataset.

- **Error detection and correction.** The user-centered evaluation and further data cleansing intervention are meant to identify and eliminate data errors, e.g. by detecting the records that do not correctly match. In particular false positives (records that are wrongly matched as describing the same artwork) are deleted, and, eventually, false negatives (records that actually match but that were not found by means of the image recognition tool) are matched.

- **Cost optimization.** Effective quality improvement actions along with a set of thresholds are defined so as to minimize and optimise costs of data integration in the archive catalogue.

In particular, to optimise cleansing activities and facilitate error detection and correction, a web application leveraging the developed methods is developed, which allows both scholars and cataloguers to (1) review the retrieved history of attributions of selected artworks, and (2) provide feedbacks on the correctness of results.

Data resulted from data sources integration can be integrated in current online photo archive catalogues. The inclusion of information in the catalogue can be performed by means of a bespoke API (Application Programming Interface), which queries the aforementioned triplestore including results of the data integration and returns the list of attributions and related metadata in a defined format (i.e. JSON). The implementation of both the web application and the underlying API is described in Chapter VII.

Secondly, in order to optimise the actual implementation of data quality improvements, a number of policies are defined for selecting records in the original catalogues that would benefit of data integration, namely:

- **Attributions motivated by low rated criteria or without motivation.** Records wherein authorship attributions are supported by criteria whose rating is equal or less than 5, or that are not supported by any motivation are automatically integrated with attributions found in external sources that are motivated by higher rated criteria and that confirm the same attribution.

- **Potentially outdated attributions.** Records including information that is likely to be outdated, such as bibliography, scholars and organisations' attributions, are compared. If retrieved external resources include more updated information, this is integrated into the original data.

- **Contradictory attributions.** When contradictory attributions are found, cataloguers review those one-by-one and eventually revise their attribution. In this case, they record the change in the history of attributions, and add the reference to retrieved sources of information. Otherwise, cataloguers may decide to add retrieved contradictory statements in the list of discarded attributions.

These actions are required to ensure textual authoritativeness is granted for those cataloguing records that may be affected by information quality issues over time. It is worth to notice that the first two actions do not entail an actual modification of the original dataset, meaning that new (integrated) data can be accessed and served by relying on the API services. Only in the last case, i.e. when contradictory information are found in other authoritative sources, cataloguers may decide to review or update their data.

Chapter 7

mAuth. A Framework for Discovering and Comparing Authorship Attributions

In this chapter we present mAuth, a framework based on a semantic crawler that harvests authorship attributions, provenance information and related sources of information, in the Web of Data. In the following sections requirements, architecture, and implementation of the framework are described. In particular, the following components are presented: (1) the crawling process that harvests authorship attributions and stores information in the mAuth triplestore; (2) the process that queries the triplestore, ranks results, and assigns authoritativeness scores to the retrieved information sources; (3) the API that serves data to applications, and (4) the web application that aggregates and serves sorted results to the final user.

7.1 Scope, restrictions, and requirements

Frameworks for crawling the web have been developed for long time and their usefulness is demonstrated. Nonetheless, when dealing with art historical data and authorship attributions in particular, general-purpose crawlers are not sufficient to address so specific types of information. A focused crawling process is necessary to guide the crawler through relevant information, discarding immediately irrelevant resources, and saving time. Focused crawlers are also called preferential or heuristic-based crawlers. The heuristic we

use in the proposed solution is based on ontology mapping. All the resources harvested are semantically annotated, served as RDF data, and represented according to one or more vocabularies. Vocabularies are mapped to a crawling schema, and fetched data are stored in a central triplestore.

The three main objectives of mAuth are:

- Discover relevant authorship attributions in the Semantic Web with regard to an input artwork.
- Analyse and rank observed attributions on the basis of their textual authoritativeness.
- Provide final user/application with a sorted list of attributions, accompanied by context information, a score of authoritativeness and, eventually, metrics describing scholars' authoritativeness.

The framework is tailored on the use case of this research, i.e. art historical photo archives and multi-purpose datasets that record authorship attributions. Indeed, mAuth comes as a proof of concept of the conceptual framework for assessing textual authoritativeness of information sources recording authorship attributions detailed in Chapter VI.

The crawler retrieves sources including information related to an artwork, whose URI is the starting point of the crawling process. Assuming that (1) such specific types of information can be found in a restricted number of sources, (2) only a selected number of sources is deemed relevant and accurate enough to review an initial assertion (i.e. an attribution), and that (3) the crawling scheme for different sites can differ dramatically, we restrict the number of sources to be fetched by relying on a list of trusted providers. This allows us to have a high accuracy of information harvested, and reduce time-consuming and error-prone activities related to customise the crawler for many non-relevant websites. Such a supervised and highly curated approach ensures the bucket of results always satisfy user's expectations, and saves time otherwise spent to retrieve the aforementioned sources one-by-one.

Precisely, six data sources are used for the sake of the evaluation of the framework, namely: the Zeri photo Archive, Villa I Tatti Berenson Library, the Frick Art Reference Library, DBpedia (the Italian, English and French datasets), Wikidata, and VIAF. Three out of six providers serve RDF data related to artworks of Modern Era only. Therefore, the crawling (and the consequent evaluation of the framework) is based on such a subset of available data. Nonetheless, the framework comes with a number of components that

can be customised according to different needs, e.g. the list of data sources to be harvested, the data to be mined, scores and ranges of the ranking model.

Since art historical data do not change significantly over time, fetching data is a task that can be executed on a given time interval, rather than being executed on-the-fly every time a user inputs a new URI. Indeed, the number of trusted sources is fixed (although easily extendible), and we assume also the number of artworks to be retrieved is limited. In order to optimise the query time and ensure accuracy of the fetched results, the crawling process is performed on a monthly basis on a given collection of URIs and results are stored for being analysed.

The main contribution of the mAuth framework is the advanced level of data analysis performed on the data extracted from the sources. Specifically, information resources retrieved are rated and sorted according to a ranking model, described in Chapter VI, and then served to applications by means of an API, and to users by means of a Web application that shares the same logic of the API. The ranking model highlights the most documented and well-researched information sources, as based on their textual authoritativeness. Secondly, bespoke metrics provide scores for describing the cognitive authoritativeness of scholars cited as primary sources of an attribution.

The aim is to facilitate users' tasks related to knowledge discovery in the Arts field, and support the decision-making process when reviewing artworks attributions. Three types of users may benefit of this approach, namely:

- Photo archivists, cataloguers, and data collection managers, who want to update their data by integrating information retrieved in other authoritative online sources.
- Connoisseurs that are seeking for the complete history of attributions with regard to a specific artwork.
- Scholars in the Humanities that are looking for documentation related to artworks and artists.
- Auction firms and art business representatives may benefit of the tool to quickly retrieve the most authoritative attribution. However, at the moment these actors were not involved in the user-centered evaluation of the framework.

In terms of requirements, the crawler responds to the following tasks:

- The crawler is started from a command-prompt, with a number of given components, namely: (1) a linkset of URIs identifying artworks, (2) a settings file including instructions on how to access data sources (content negotiation, SPARQL endpoint, or Linked Data Fragments servers), (3) a list of trusted providers to be harvested, defined by the URI base to be matched, and (4) an ontology mapping document, including triple patterns for query rewriting.
- The crawler queries a linkset including a collection of URIs identifying artworks, it parses the URI bases, and looks for matches in the list of trusted providers.
- The crawler looks into a settings file providing instructions on how to access the data sources.
- The crawler looks into a mapping document to collect the triple patterns to be parsed, rewrites a query to be performed against some endpoint, and returns results annotated according to the crawling schema.
- The crawler stores retrieved triples in a dedicated named graph of the local triplestore.

Data stored in the triplestore are queried and analysed by a number of algorithms that return a sorted list of results, which are grouped by data provider and identified by a score representing their textual authoritativeness. In particular, both accepted and discarded attributions recorded in data sources are retrieved. The ranking model is applied to data in order to associate attributions with four partial scores, namely: (i) domain expert score, (ii) date score, (iii) criterion score, and (iv) agreement score. Such algorithms respond to the following requirements:

- The algorithm identifies the data provider of each attribution as being a domain expert or not, and assigns the domain expert score.
- The algorithm sorts attributions by date, calculate their timeliness, and associates the date score.
- The algorithm queries a controlled vocabulary of terms including criteria and related rating, and associates the criterion score.

- The algorithm identifies whether an attribution cites a scholar as primary source, and performs a number of statistical analyses to return the scholar's artist-related index and the acceptance rating.
- The algorithm groups attributions by artist, queries the linkset including equivalence statements on artists, and calculates the agreement score, i.e. the number of sources in agreement on the same artist.
- The algorithm sums all the scores and associates a final score to each attribution.

The triplestore stores harvested data that are have not been manipulated by the ranking model yet. The aforementioned algorithms compute the scores on-the-fly and serve the sorted list of results on demand. Manipulated data can be consumed in two ways according to the nature of the request, namely:

- An API accepts as input the persistent URI identifying an artwork, and returns the ranked list of results as a JSON object.
- A Web application provides a web interface for querying the triplestore. The interface accepts the URL of the cataloguing record or web page describing an artwork at hand.

7.2 Architecture of the framework

Fig. 7.1 provides an overview of the architecture of mAuth. At a higher level, mAuth is made out of three components that aim at achieving the three aforementioned objectives, namely: (1) a crawler for discovering and mining relevant data sources describing artworks, (2) a stack of technologies for analysing and ranking data sources, (3) bespoke software solutions - an API and a web app - to serve ranked data according to the request.

Figure 7.2 shows how components of the framework interact with each other. The framework consists of the following components:

- Settings file
- List of trusted providers

- Image similarity index
- Equivalence lookup service
- URI stack
- Domain filter
- Mapping rules
- Data miner
- Crawling schema
- Observation graph
- Data Analyser
- Ontology-based ranking model
- Controlled vocabulary of rated criteria
- Statistics graph
- mAuth API
- mAuth Web application

Settings file. The settings file includes an extended number of data providers that are potentially relevant to the research, and the instructions for accessing data sources to be mined. In detail, it includes (1) the URI base of resources belonging to a domain, (2) the data access strategy, namely content negotiation, SPARQL endpoint, or Linked Data Fragments, and (3) the access point, whether it is the URI of the SPARQL endpoint, or a rewriting rule to fetch RDF documents by content negotiation. The settings file comes as a JSON file that can be easily substituted or extended for including new data sources.

List of trusted data providers. To restrict the focus of the evaluation of the framework, mAuth relies on a list of six trusted data sources to be fetched and analysed. The list is used by the domain filter to prune the URI stack from not relevant domains in the mining process.

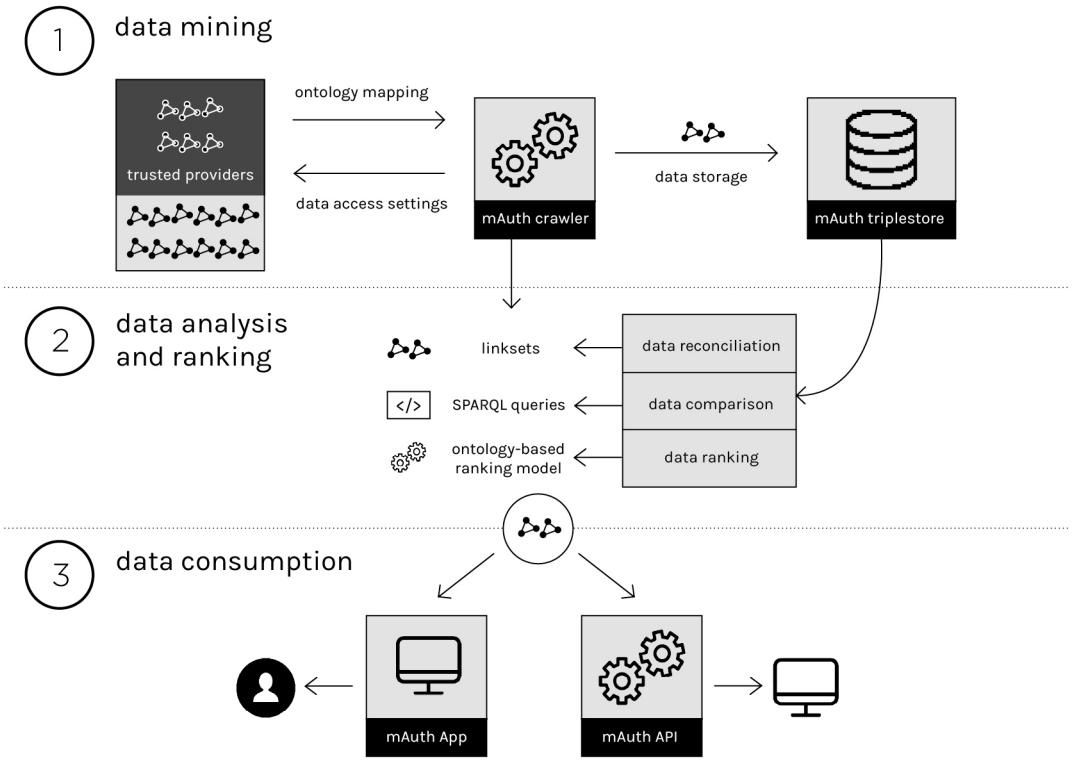


Figure 7.1: Overview of mAuth architecture

Images similarity index. The image similarity index includes the results of the image recognition process performed over the images of the trusted data providers. Images were first retrieved on online catalogues. Images retrieval was a one-time task performed by means of a web scraping script, which fetched images from online catalogues of the three aforementioned photo archives. Pastec[1] search engine is used to compare images and create the index of matching images. The URIs of the artworks whose images have a similarity score is greater than 30.0 are included in the URI stack, and the link between similar artworks is included in the linkset of artworks.

Equivalence lookup service. The lookup service retrieves equivalences for the URIs relevant to the mining process, namely the URIs identifying (i) artworks, (ii) artists, (iii) organisations and scholars retrieved in the six trusted data providers.

An initial linkset for each of the above mentioned entities is created by using several approaches. Artworks are matched by using the aforementioned computer vision tool Pastec. Artists, organizations, and scholars are reconciled to authority records, namely VIAF and DBpedia, by means of a semi-automatic approach.

[1] `http://pastec.io/`

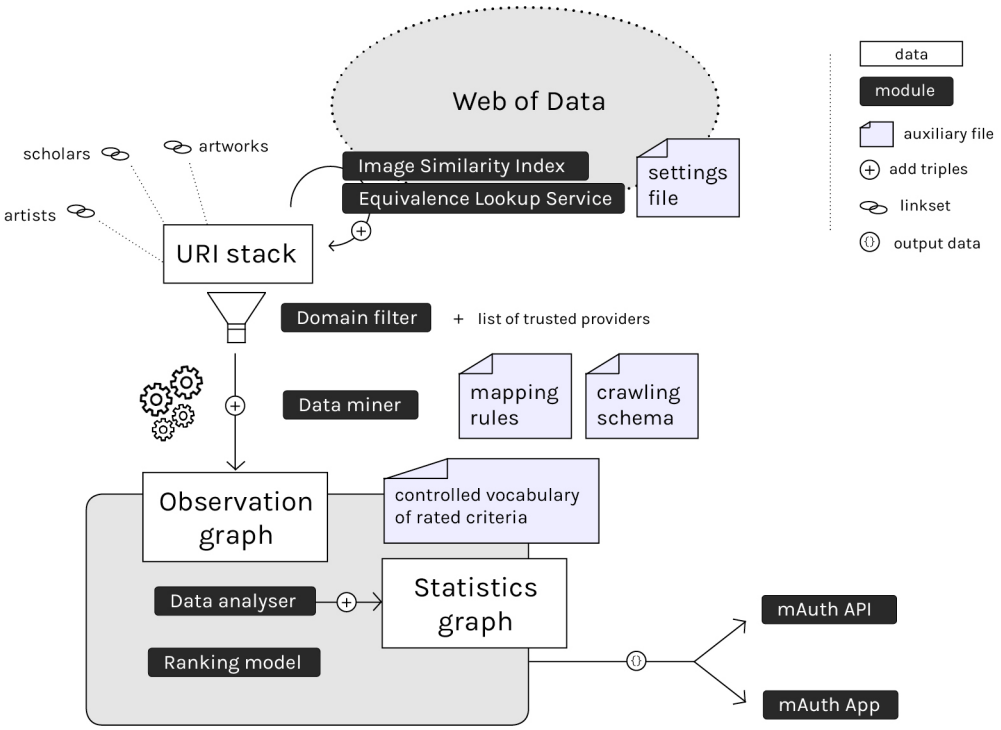


Figure 7.2: Components of the framework mAuth

In particular, methods for fuzzy matching strings[2] are applied to (a) a list of input labels associated to URIs in the six data sources and (b) labels in VIAF and DBpedia. A similarity score is computed and links are stored in bespoke linksets. Transitive links are created for the entities reconciled to the same VIAF or DBpedia records.

The lookup service queries the linksets and uses instructions detailed in the settings file to access data sources of all the URIs. It performs two iterations of equivalences lookup on the so created linksets. In detail, it first looks for equivalences explicitly stated in data sources, i.e. `owl:sameAs` properties, and stores the equivalence link in the linkset; then the service accesses retrieved URIs in the first round, and performs a second round of equivalences lookup. All the new links are stored in the linkset. Lastly, it creates transitive links between URIs that result being equivalent to a common URI. The so created linksets are stored in the triplestore and define the URI stack to be used to initialise the crawling process.

URI stack. The URI stack is the initial collection of URIs identifying artworks, artists, organisations and scholars, that are stored in the mAuth triplestore in three dedicated named graphs. These include URIs

[2] `https://github.com/seatgeek/fuzzywuzzy`

extracted from the bucket of six trusted data providers and their equivalence links. The URI stack can be easily extended by adding new equivalences to the linksets. Each URI identifying an artwork is sent in a *first in first out* order to the Domain Filter for further processing.

Domain filter. The domain filter checks whether URIs included in the URI stack belong to a domain in scope or not, so as to restrict the scope of the crawler to the six specified domains. Indeed, links extracted by the lookup service added to the URI stack belong to heterogeneous sources - which can be useful whether the scope of the research has to be broader. For the sake of this evaluation, all the URIs that are not in scope are filtered out. For instance, only three of the diverse national DBpedia datasets harvested by the lookup service are taken into account, namely the Italian, English, and French DBpedia. The domain filter is extendible as long as new data providers are included in the list of trusted providers, in the settings file, and in the mapping document.

Mapping rules. The mapping document includes the ontology mapping rules for all the properties relevant to identify an authorship attribution in the domain in scope. In particular, it includes triple patterns useful to retrieve: (1) artist, (2) title of the artwork, (3) criteria motivating the attribution, (4) date of the attribution, (5) sources of information, (6) cited scholar and institutions, and (7) images. Triple patterns are used by the data miner for rewriting a number of SPARQL queries to be performed against either triplestores of RDF documents. Several mapping rules may coexist and be called by the data miner according to the nature of data to be retrieved. The mapping document is provided as a JSON file where triple patterns for each data provider are listed.

For instance, the following JSON excerpt represents triple patterns used to query Wikidata in the mapping document. We assign a human-readable label as keys and a property path as value. The object of such property paths are the value we are looking for. When rewriting the SPARQL query values are retrieved and returned to the Data miner component.

```
"http://www.wikidata.org/entity/": {
        "label": "Wikidata",
        "artworkTitle": "http://www.wikidata.org/prop/direct/P373",
        "artist": "http://www.wikidata.org/prop/direct/P170",
        "artistTitle": "http://www.wikidata.org/prop/direct/P170 ?c .
                        ?c http://www.w3.org/2000/01/rdf-schema#label",
        "other_artist": "",
        "other_date": "",
```

```
        "other_criterion": "",
        "other_biblio":"",
        "biblio": "",
        "source": "http://www.wikidata.org/prop/P170 ?c.
                        ?c http://www.w3.org/ns/prov#wasDerivedFrom ?d.
                        ?d http://www.wikidata.org/prop/reference/P143",
        "notes": "",
        "criterion": "",
        "others": "",
        "scholar": "",
        "images":"http://www.wikidata.org/prop/P18 /
                        http://www.wikidata.org/prop/statement/P18",
        "date": "http://schema.org/dateModified"
}
```

Data miner. The data miner is the core component of the framework, which integrates all the previous components. It accepts in input (1) a URI identifying an artwork taken from the URI stack, (2) a settings file, and (3) a document of mapping rules. The data mining algorithm is iterated over all the URIs in the URI stack. Data are accessed by means of the rules specified in the setting file. For each property path listed in the mapping document the miner rewrites a SPARQL query to be performed against the dataset, by using the method specified in the settings file. Data fetched are stored in the Observation graph, which will be subject of further analyses by means of the Data analyser. Results stored in the Observation graph are represented according to the Crawling schema.

Crawling schema. The crawling schema is based on the Observation ontology pattern[3] and the PROV Ontology.

The following listing in turtle syntax shows an exemplar of data retrieved and transformed according to the crawling model. The example describes the observation of an authorship attribution fetched in the cataloguing record n. 39459 of the Zeri photo archive. The record describes an attribution made by Everett Fahy, who ascribed the artwork “San Pietro Martire in preghiera e sante” to Francesco Granacci. A second criterion, archival classification, confirms the same attribution, which is dated 1990. Moreover, links to three photographs depicting the artwork are fetched. Provenance information of the fetching process is recorded by means of two PROV properties, associating a date time and an agent to the mining process.

```
mauth:39459-artist-granacci-francesco-obs
```

[3] http://www.ontologydesignpatterns.org/cp/owl/observation.owl

```
rdfs:label "Zeri Foundation (University of Bologna) accepted attribution" ;
mauth:hasType   mauth:accepted ;
mauth:hasObservedArtist zeri:granacci-francesco ;
mauth:hasObservedArtwork zeri-artwork:39459 ;
mauth:hasObservedCriterion criteria:scholar-attribution ;
mauth:hasObservedCriterion criteria:archival-classification ;
mauth:agreesWith zeri:e-fahy ;
mauth:hasAttributionDate "1990-01-01T00:00:00.001Z"^^xsd:dateTime ;
mauth:hasSourceOfAttribution
        <http://w3id.org/zericatalog/artwork/39459> ;
mauth:image
        <http://catalogo.fondazionezeri.unibo.it/foto/120000/82800/82571.jpg> ,
        <http://catalogo.fondazionezeri.unibo.it/foto/120000/82800/82572.jpg> ,
        <http://catalogo.fondazionezeri.unibo.it/foto/120000/82800/82573.jpg> ;
prov:atTime "2018-07-22T22:35:48.767Z"^^xsd:dateTime ;
prov:wasAttributedTo     mauth:md ;
```

Observation graph. The Observation graph represents snapshots of authorship attributions related to artworks of the Modern Era available in six trusted data sources. All data fetched in the aforementioned sources are here stored according to a unique crawling model. Data are queried and further elaborated by the Data analyser so as to extract information on textual authoritativeness and create indexes for cognitive authoritativeness. The benefits of storing snapshots of retrieved attributions rather than querying data sources on-the-fly every time a user asks for a URI are two: to speed up the query phase and to preserve changes in attributions over time (i.e. the versioning).

Querying remote triplestores may be time-consuming, affected by time-outs or other limits of third-party softwares. For this reason, when an alternative Linked Data Fragments server was available it was preferred over SPARQL endpoints. Despite such a solution halved the query time, the time required to query heterogeneous sources was still high. Therefore we preferred the intermediate storage in a bespoke triplestore. The Observation graph is currently stored in a Blazegraph triplestore,[4] which was chosen because of its scalability and high-performance.

Data Analyser. The Data analyser consists of a number of scripts that query data stored in the Observation graph, so as to (1) sort results of a query according to the ranking model, (2) calculate scholars' citation indexes, and (3) send the final list of attributions to the API and the web application. It operates on the basis of a user input, who queries either the API or the web application with a URI identifying the

[4] https://www.blazegraph.com/

artwork to be retrieved. The Data analyser looks into the Observation graph for matches and retrieves a list of observations of authorship attributions. In order to associate attributions with a score of textual authoritativeness it performs four operations, namely:

- Checks data provenance against the List of trusted providers and attributes a *domain expert score* to the attribution.
- Extracts from the Controlled vocabulary or rated criteria the rating associated to criteria supporting the attribution and computes the *criterion score*.
- Compares retrieved attributions so as sort them by date (when available). It calculates the timeliness of the attribution, i.e. a function that compare the distance between the current date and the date of the attribution, and associates the *date score* to the attribution.
- Groups attributions by artist, and calculates the *agreement score*. It queries the linkset of artists to identify overlaps between the lists of equivalences related to each artist mentioned in retrieved attributions and increments the score for each match.

The Data analyser sums the partial scores and includes the final textual authoritativeness score in the list of results for each attribution. Ranges of values to be used when computing partial scores are provided by the ontology-based ranking model. Moreover, the Data analyser performs a one-time task for creating citation indexes related to scholars. It analyses scholars' citations across data providers, so as to compute the artist-related index and the acceptance rating - whose rationale is detailed in Chapter VI. Results are stored in the Statistics graph. When a scholar is cited as primary source of an attribution, the Data analyser looks into the Statistics graph for such indexes and includes the values in the list of results. Finally, it returns the list of results as a JSON file.

Ontology-based ranking model. The ontology-based ranking model is the component that provides ranges of values of the aforementioned scores. The Data analyser is in charge to weight retrieved information accordingly. The ranking model takes as input a number of property values, namely: the name of the data provider, the label of criteria, the position of the attribution date in the list of sorted attributions, the number of agreements. Property values are defined according to the crawling schema. The rationale of the ranking model is detailed in Chapter VI.

Controlled vocabulary of rated criteria. The Controlled vocabulary of rated criteria is a named graph stored in the mAuth triplestore describing the twenty-two criteria that can motivate an attribution, as deduced from the data analysis performed over three representative photo archives. Criteria are individuals of the class `hico:InterpretationCriterion`, and the rating is associated by means of the DBpedia ontology property `dbo:rating`. For instance, the criterion "documentation" is described as follows (in turtle syntax):

```
criteria:documentation rdf:type hico:InterpretationCriterion ;
      rdfs:label "documentation" ;
      dbo:rating "10.0"^^xsd:float .
```

Statistics graph. The Statistics graph is the result of the analysis performed over the Observation graph in order to extract information on scholars' authoritativeness. The Data analyser queries the Statistics graph to retrieve indexes to be associated to the attributions, and include them in the list of results to be sent to the API/app. It is updated on a monthly base so as to record indexes changes. However, temporal snapshots are not preserved.

mAuth API. The mAuth API provides functionalities that go beyond data access. Indeed, it is a means for relationship discovery and data integration. It is accessible through HTTP calls, and accepts in input the persistent URI identifying an artwork included in one of the six aforementioned providers. It reuses the described components so as to look into the Observation graph for a match, retrieves attributions and indexes, ranks and sorts results, and serve the list of results as a JSON file. For instance, the following call retrieves all the attributions related to the Wikidata entity "Venus of Urbin", unanimously attributed to Tiziano Vecellio in 4 information sources.

```
curl http://163.172.177.79:8000/full/http://www.wikidata.org/entity/Q727875
```

mAuth Web application. The mAuth web application shares the same logic of the mAuth API, i.e. all the aforementioned components and auxiliary files, and serves ranked data to users that look for the history of attributions related to a single artwork. It works as an aggregator of results, and it is used for the user-centered evaluation of the conceptual framework described in Chapter VIII. Figure 7.3 shows the interface and the list of results of a query.

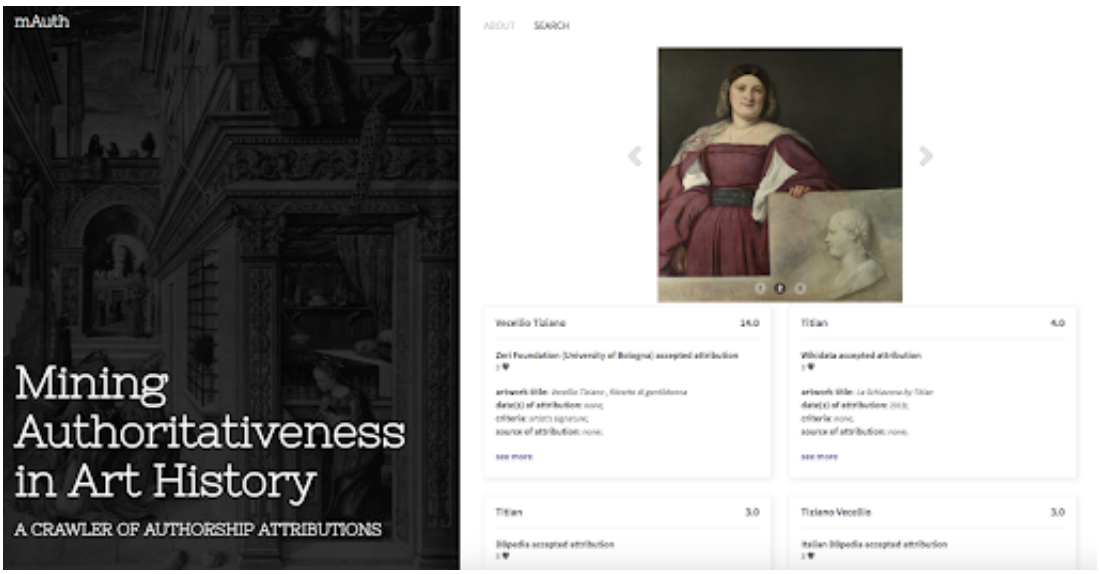


Figure 7.3: Screenshot of the mAuth Web application, including results of a research

7.3 Implementation

We have chosen to implement the conceptual framework for retrieving and ranking authorship attributions in Python 2.7. Python benefits of a high number of available open libraries for interacting with RDF data, such as RDFlib[5] (for creating and manipulating RDF data), SPARQLWrapper[6] (for querying remote triplestores), Hydra.py[7] (for querying Linked Data Fragments), and bespoke modules to interact with a number of triplestores. Blazegraph is a performant triplestore used to store mAuth graphs and linksets. It interacts with Python framework by means of the library Pymantic[8]. Unfortunately, some of the aforementioned libraries work only with Python 2, hence the crawler is built works with Python 2 only, while both the API and the Web application are compatible with Python 3. Both server-side (e.g. crawler, analyser, ranking model) and client-side components (web interface) are developed by using the microframework Flask.[9]

mAuth comes as a toolkit for art historians that want to retrieve the history of attributions of pieces of art related to the Modern Era. Moreover, it offers two services to access and compare data (the web app), and for integrating data in other applications (the API). The toolkit includes all the resources resulted from

[5] https://github.com/RDFLib/rdflib
[6] https://rdflib.github.io/sparqlwrapper/
[7] https://github.com/pchampin/hydra-py
[8] https://github.com/blazegraph/blazegraph-python
[9] http://flask.pocoo.org/

the data integration process performed over six data providers, and a number of flexible python modules that can be reused in different contexts (such as different providers and different scopes of information to be gathered). The toolkit includes the following components:

- Three linksets including results of data reconciliation on artists, artworks and historians related to paintings of the Modern Art. The linkset of artists includes around 37.386 equivalences links between 12.227 individuals. The linkset of artworks includes 7.284 links between 2.474 artworks described by the six aforementioned providers. The linkset of historians include 33.676 links between 11.996 entities.
- A dataset of observed attributions with regard to 1.269 unique pieces of art.
- A dataset on citation indexes for the aforementioned scholars.
- An ontology mapping document including triple patterns for retrieving information in six data sources.
- A settings file with instructions on how to access six data sources.
- An instance of Blazegraph triplestore storing the aforementioned graphs.

All these elements can be easily extended or substituted in order to make changes in the bucket of data sources to be retrieved and analysed. The source code of mAuth and all related resources are stored in a Github repository available online.[10]

The API and web application. Both the API and the Web application share the same application logic. They both depend on the knowledge base realised for the purpose. Figure 7.4 shows the class diagram of the project.

The classes `Connoisseur` and `Mauth` are the two main classes of the project. `Connoisseur` is the class that creates the knowledge leveraged by the `Mauth` class. It has methods like `fetchData()`, which looks into the web of data and fetches data. The method `findAttributions()` uses `fetchData()` for each of the triple patterns outlined in the mapping document and sends results to the `updateAttributions()`.

[10] `https://github.com/marilenadaquino/mauth`

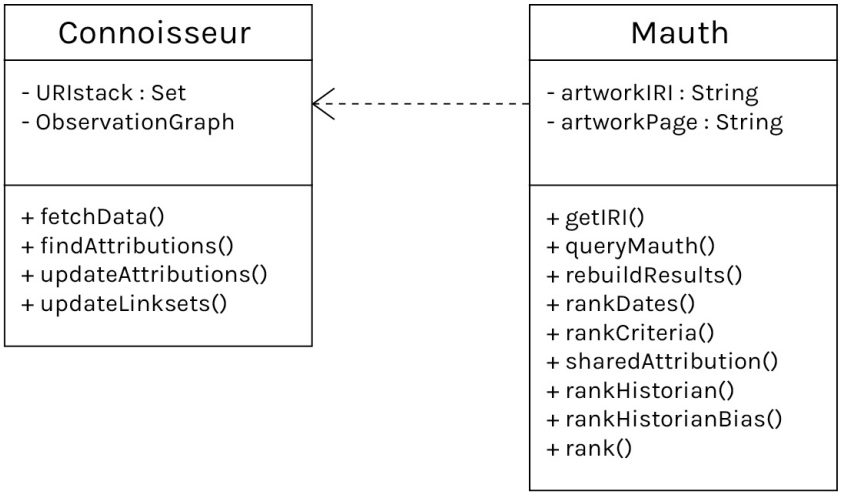


Figure 7.4: Class diagram of mAuth

The methods `updateAttributions()` and `updateLinksets()` send data to the triplestore to store the knowledge base.

The `Mauth` class calls the other class in order to perform the analysis or the knowledge base and returns a ranked list of results. It has methods such as `getIRI()` to be used when the user inputs the URL of a web page describing an artwork instead of the IRI of the artwork itself. The method `queryMauth()` performs a SPARQL query against the mAuth SPARQL endpoint and returns a list of results in JSON, that is reorganised by using the `rebuildResults()` method. Then a number of methods calculate partial scores - `rankDates()`, `rankCriteria()` and `sharedAttribution()` - and scholars' indexes - `rankHistorian()` - to be associated to elements of the list. Finally, the method `rank()` gather all the information and returns the list of results including the results of the analysis.

Chapter 8

Evaluation of artefacts

In this chapter the evaluation of the artefacts developed in this research is presented. The ontologies presented in Chapter V are evaluated first. In particular the FEntry Ontology and the OAEntry Ontology are evaluated by means of the comparison to golden standards and by means of a data-driven approach. The HiCO Ontology is evaluated separately by means of the evaluation of the mAuth application that leverages the model as the basis for the ranking model. Such an evaluation is meant to validate hypotheses H1 and H2. The conceptual framework, the dimensions and the ranking model presented in Chapter VI are evaluated by means of a user-centered evaluation. The aim is to validate hypotheses H3 and H4, and provide a preliminary evaluation of H5. The mAuth framework is evaluated as a proof of concept of the aforementioned elements and to demonstrate that Semantic Web technologies can effectively support common tasks in the Humanities and can respond to sophisticated needs such as supporting the decision-making process. The evaluation contributes to validate hypotheses H6 and H7.

8.1 Ontologies evaluation

The research project here presented resulted in the creation of two ontologies and the revision of an existing one. In particular, we developed two domain ontologies, i.e. the OAEntry Ontology for describing the Arts domain, and the FEntry Ontology for describing the Photography domain (which is a revision and an extension of an earlier version of the same), and a task ontology, i.e. the HiCO Ontology, for describing

questionable information. Various approaches to the evaluation of ontologies have been proposed in the literature, depending on the type of ontologies to be evaluated and for what purpose [Brank et al., 2005]. Evaluation approaches mainly fall into the following categories:

1. Comparison of the ontology to a *golden standard* [Maedche and Staab, 2002].
2. Task-based evaluation, based on using the ontology in an application and evaluating the results [Porzel and Malaka, 2004].
3. Data-driven evaluation, comparing sources of data (e.g. a collection of documents) about the domain to be covered by the ontology [Brewster et al., 2004].
4. Evaluation done by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements, etc. [Lozano-Tello and Gómez-Pérez, 2004].

According to Brank, Grobelnik, and Mladeni [Brank et al., 2005] evaluation methods may apply to several levels of features to be evaluated, including: lexical vocabulary level, hierarchy level, other relations level, context or application level, syntactic level, structure and design level. As suggested by authors, we decided to rely on all of the four approaches according to the feature to evaluate. In detail, we evaluate the following three features:

- **Vocabulary layer.** The focus is on concepts, instances, and facts included in the ontology, and the vocabulary used to represent or identify these concepts. Evaluation on this level involves the comparison with data domain-specific datasets (i.e. methods 3). This method is applied for the evaluation of vocabulary completeness and data consistency related to the FEntry and OAEntry ontologies.
- **Hierarchy and taxonomy layer.** Refers to the consistency of hierarchical is-a relation between concepts. This is evaluated by comparing the FEntry and OAEntry ontologies to golden standards such as CIDOC-CRM and FRBR (i.e. method 1).
- **Context or application layer.** An ontology may be part of a larger collection of ontologies, and may be referenced by various ontologies. A form of context is the application where the ontology is

used (i.e. method 2). The evaluation looks at how the results of the application are affected by the use of the ontology and whether these are satisfying for users (i.e. method 4). This is the evaluation method chosen for the task ontology HiCO.

Vocabulary layer. Although several methods potentially apply to the evaluation of the ontologies, the methodology we have chosen mainly relies on the methodology adopted for ontology development, i.e. SAMOD [Peroni, 2016]. SAMOD is characterised by a strong data-oriented approach for the creation of ontologies, and encourages developers to create well-documented resources that are iteratively evaluated. The logical consistency of every single module that compose the final models represents the answer to a competency question, which is first evaluated by domain experts, and then tested over a representative data source.

The evaluation of the domain ontologies is performed over the Zeri photo archive RDF dataset, and is validated by domain experts, i.e. photo archivists of the Zeri photo archive. Results of this evaluation, meaning all the competency questions and data tests, are detailed in bespoke documents, available online along with the FEntry Ontology[1] and OAEntry Ontology.[2]

As an example of the data-driven evaluation process, we describe the steps performed to validate two competency questions addressed when developing the OAEntry Ontology. The first input is a motivating scenario, where we provide a name, a description, and an example in natural language:

```
## NAME:Description of works of art: metadata and relations
between works of art.

## DESCRIPTION: The OA Entry is a document containing
metadata about a work of art. Any OA Entry may be described
in terms of some entities of the Functional Requirements
for Bibliographic Records, i.e. Work (referring to
the essence of the entry), and Expression (referring to
its contents and its possible revisions). The work of art
may be described as a FRBR Work (referring to
the essence of the work of art), as a FRBR Manifestation
(referring to any transformation the work of art may undergo)
and as a FRBR Item (referring to the physical object).
The described work of art may be an original work of art
or a derivate one. When the described work of art
```

[1] http://www.essepuntato.it/2014/03/fentry/samod
[2] http://oaentry-ontology.sourceforge.net/samod/OAdevelopment.zip

```
is somehow related to a former/latter work of art,
the relation may be described in terms of
an influence between two works.

## EXAMPLE 1: A OA Entry n. 15429 describes
an anonymous drawing of the ceiling of Sistine Chapel,
which is a copy of the Michelangelo Buonarroti's work of art.
```

Secondly, we address a number of competency questions extracted from the motivating scenario, for which we provide an ID, a description in natural language, and an expected outcome:

```
## COMPETENCY QUESTION
ID:CQ1
Name: OA Entries and their works of art
Question: What are all the OA Entries and the works of art
        they describe?
Outcome: A list of pairs containing the OA Entry
        and the work of art described in the OA Entry.
Example1: OA Entry 1 , anonymous drawing of Sistine Chapel
Example2: OA Entry 2 , the Jesus' baptism by Leonardo da Vinci

## COMPETENCY QUESTION
ID:CQ2
Name: Original and derivative works of art
Question: What are all the original works of art
        and their derivative works and how they are derived?
Outcome: A list of tuples, each describing the original work of art,
        the influence it had on a derivative work, and the derivative work.
Example1: anonymous drawing of Sistine Chapel , copy ,
        Michelangelo Buonarroti's fresco of Sistine Chapel
```

The third step aims at providing an exhaustive glossary of terms that are addressed by the scenario. Domain experts revised the following glossary:

```
TERM: OA Entry
DEFINITION: A document containing metadata about a work of art.

TERM: Work of art
DEFINITION: An aesthetic physical object, result of an artistic creation.

TERM: describes
DEFINITION: the link between an OA Entry and the work of art that
        the OA Entry describes.

TERM: is described by
```

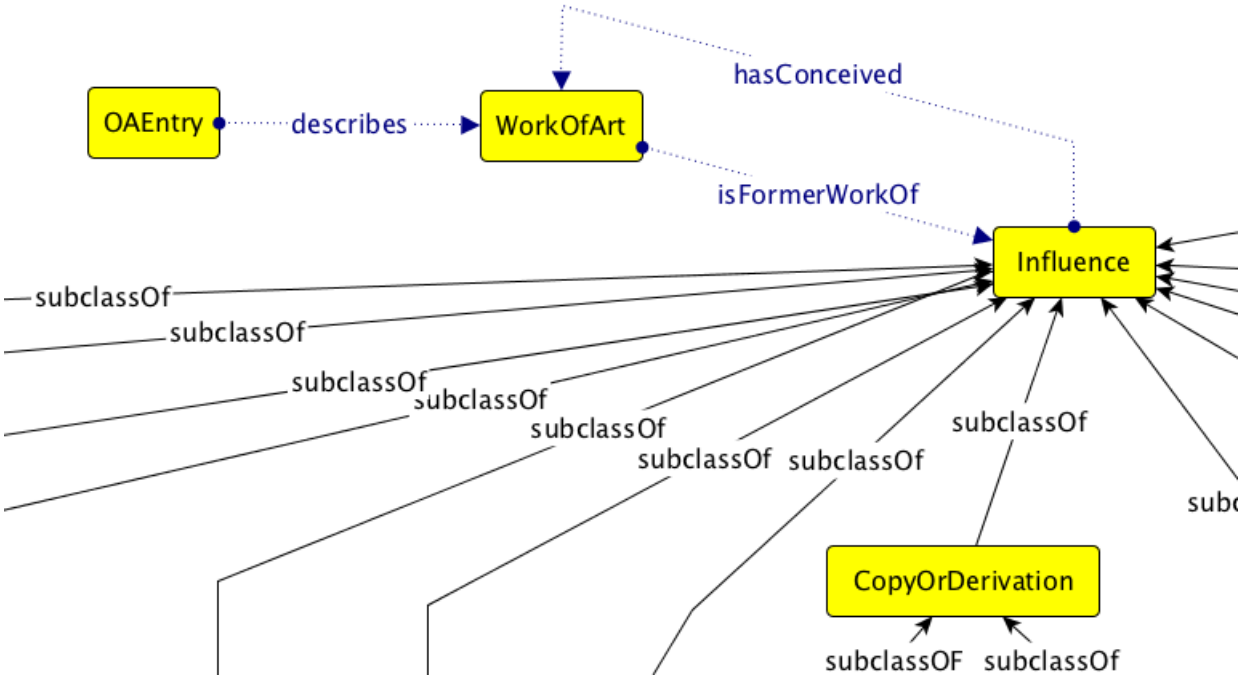


Figure 8.1: Sample of the OAEntry Ontology diagram referencing CQ1 and CQ2

```
DEFINITION: the link between a work of art and an OA Entry
        describing it.

TERM: is former work of
DEFINITION: the link between an original work of art and
        an entity describing the influence it had on a derivative work.

TERM: has conceived
DEFINITION: the link between an entity describing the influence
        a derivative work may undergo and the derivative work.

TERM: Influence
DEFINITION: A relation between two works, where
        the latter is a derivative work of the former.

TERM: Copy
DEFINITION: The influence a work of art may have on another one
        when the latter is considered a copy of the former one.
```

We draw a diagram (Figure 8.1) in order to represent the interaction between the above listed terms, whether these are concepts or relations. For the sake of brevity, we include here the final version of the diagram rather than every stage of the diagram development (which has been re-drawn several times).

We develop a modelet, i.e. an OWL file including classes and properties addressed by the glossary of

terms that describe the motivating scenario at hand. In particular, the modelet does not include any term belonging to external vocabularies. The following listing shows an excerpt taken from the aforementioned modelet, serialised in RDF/XML.

```
<!-- http://purl.org/emmedi/oaentry-modelet/describes -->
<owl:ObjectProperty rdf:about="http://purl.org/emmedi/oaentry-modelet/describes">
    <rdfs:domain rdf:resource="http://purl.org/emmedi/oaentry-modelet/OAEntry"/>
    <rdfs:range rdf:resource="http://purl.org/emmedi/oaentry-modelet/WorkOfArt"/>
</owl:ObjectProperty>

<!-- http://purl.org/emmedi/oaentry-modelet/hasConceived -->
<owl:ObjectProperty rdf:about="http://purl.org/emmedi/oaentry-modelet/hasConceived">
    <rdfs:domain rdf:resource="http://purl.org/emmedi/oaentry-modelet/Influence"/>
    <rdfs:range rdf:resource="http://purl.org/emmedi/oaentry-modelet/WorkOfArt"/>
</owl:ObjectProperty>

<!-- http://purl.org/emmedi/oaentry-modelet/hasFormerWork -->
<owl:ObjectProperty rdf:about="http://purl.org/emmedi/oaentry-modelet/hasFormerWork">
    <rdfs:domain rdf:resource="http://purl.org/emmedi/oaentry-modelet/Influence"/>
    <rdfs:range rdf:resource="http://purl.org/emmedi/oaentry-modelet/WorkOfArt"/>
    <owl:inverseOf rdf:resource="http://purl.org/emmedi/oaentry-modelet/isFormer\-
    WorkOf"/>
</owl:ObjectProperty>

<!-- http://purl.org/emmedi/oaentry-modelet/isConceivedByMeansOf -->
<owl:ObjectProperty rdf:about="http://purl.org/emmedi/oaentry-modelet/isConceivedBy\-
MeansOf">
    <rdfs:range rdf:resource="http://purl.org/emmedi/oaentry-modelet/Influence"/>
    <rdfs:domain rdf:resource="http://purl.org/emmedi/oaentry-modelet/WorkOfArt"/>
    <owl:inverseOf rdf:resource="http://purl.org/emmedi/oaentry-modelet/hasConceived"/>
</owl:ObjectProperty>

<!-- http://purl.org/emmedi/oaentry-modelet/isFormerWorkOf -->
<owl:ObjectProperty rdf:about="http://purl.org/emmedi/oaentry-modelet/isFormerWorkOf">
    <rdfs:range rdf:resource="http://purl.org/emmedi/oaentry-modelet/Influence"/>
    <rdfs:domain rdf:resource="http://purl.org/emmedi/oaentry-modelet/WorkOfArt"/>
</owl:ObjectProperty>

<!-- http://purl.org/emmedi/oaentry-modelet/Copy -->
<owl:Class rdf:about="http://purl.org/emmedi/oaentry-modelet/Copy">
    <rdfs:subClassOf rdf:resource="http://purl.org/emmedi/oaentry-modelet/Partial\-
    Copy"/>
</owl:Class>
```

To validate the consistency of the developed modelet over a real scenario, we transform the Zeri photo

archive dataset into RDF according to the developed modelet, and we design a number of SPARQL queries to be performed over the dataset so as to test data consistency. The following SPARQL queries aim at answering the aforementioned competency questions.

```
PREFIX : <http://www.w3id.org/zericatalog/>
PREFIX oaentry: <http://purl.org/emmedi/oaentry-modelet/>
## CQ1
SELECT DISTINCT ?oa ?work
WHERE {
        ?oa a oaentry:OAEntry ;
        ?oa oaentry:describes ?work .
}
## CQ2
SELECT DISTINCT ?orig ?influence ?deriv
WHERE {
        ?orig a oaentry:WorkOfArt .
        ?orig oaentry:isFormerWorkOf ?influence .
        ?influence oaentry:hasConceived ?deriv .
}
```

We analyse results, and whether these do not comply with requirements defined in the competency questions, we go back to the definition of the glossary, and we formulate new hypotheses. If no errors are found, we proceed to the next step, where terms defined in the modelet are refactored by using terms belonging to existing and well known ontologies. In this case, two existing ontologies apply to the description of cataloguing records and the description of the influence between artworks, namely: the FaBiO Ontology, and the PROV Ontology.

Figure 8.2 shows the final version of the OAEntry model that addresses the aforementioned competency questions, and wherein terms are aligned to existing vocabularies.

Finally, the OWL file representing terms described in the diagram is created.

The described steps are performed for each scenario addressed in the data source, and new samples of the modelet/final ontology are added to the former ones iteratively. So doing, the consistency of the vocabulary is ensured at every iteration, and the completeness of the final model is evaluated in a data-driven approach over the whole dataset at hand.

Hierarchy and taxonomy layer. In certain cases, the adoption of models by relevant peers may be even more important than standardization. CIDOC-CRM is deemed the golden standard for describing cul-

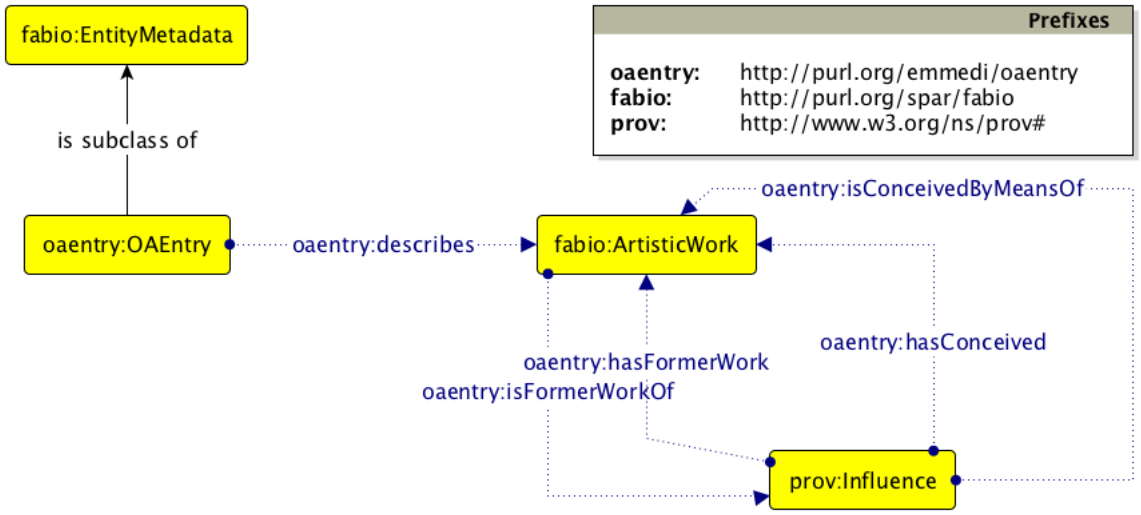


Figure 8.2: Sample of the final OAEntry Ontology

tural objects in the Cultural Heritage domain, especially in the museum domain. The wide adoption of terms belonging to CIDOC-CRM for refactoring terms of FEntry Ontology and OAEntry Ontology ensures consistency of the hierarchy is pursued, and complies with the domain community modelling preferences.

Likewise, terms of the FRBR conceptual model instantiated by the SPAR Ontologies ensure that the correct alignment of the ontologies to current practises in the library domain is respected. The detailed alignment of the FEntry and OAEntry ontologies to CIDOC-CRM and FRBR has already been illustrated in Chapter V.

Context or application layer. When developing the task ontology HiCO we had in mind an evaluation based on its actual implementation in a system for relationship discovery and data integration, i.e. the mAuth framework. To address the evaluation of such an ontology we evaluate the conceptual framework that is based on the ontology. We compare data sources where three out of six providers adopt the HiCO ontology for describing potential contradictory information, namely the Zeri photo archive, Villa I Tatti, and the Frick Art Reference Library, and we ask users to judge whether information is sufficient or not fro assessing the authoritativeness of attributions.

In the next section we present the user-centered evaluation of the conceptual framework described in Chapter VI as applied in the mAuth framework described in Chapter VII.

8.2 User-centered evaluation of the conceptual framework

The goal of the user-centered evaluation is to evaluate the conceptual framework when applied to information retrieval tasks in the Photography and Arts domain. The study addresses the usage of Semantic Web technologies when accomplishing research tasks in the Arts and Humanities, such as (1) gathering information sources recording authorship attributions and (2) comparing them in order to draw some conclusions on the goodness of statements recorded in sources.

In particular, the evaluation here described tackles the research problem *RP3 - Support users' decision-making process when assessing reliability of authorship attributions*. The main objective of the evaluation is to confirm two hypotheses, namely: *H6 - Linked Open Data and Semantic Web technologies can support and satisfy common requirements of research activities in the Arts and Humanities* and *H7 - Automatic and curated methods can support the decision-making process in connoisseurship activities*.

To validate H6 and H7, we compare the usage of mAuth web application with (1) an aggregator of images of artworks called images.pharosresearch[3] and (2) a number of online catalogues, namely: the Zeri photo archive catalog,[4] Wikidata,[5] and Wikipedia.[6] The latter are a part of the list of trusted data providers harvested by mAuth.

Images.pharosresearch gathers information about 61.051 artworks depicted in 97.091 pictures provided by eight members of the PHAROS consortium. Among the images providers there are the Zeri photo archive, Villa I Tatti, and the Frick Art Reference Library. The collections of images they provided to images.pharosresearch include the same subsets of images provided for the development of the proof of concept mAuth. Images.pharosresearch allows users to retrieve information about artworks either by means of a traditional text search or by uploading an image for similarity matching. Both mAuth and images.pharosresearch use the same image recognition tool for matching artworks that are depicted in different photographic collections, i.e. Pastec. We assume mAuth retrieves a subset of results retrieved by images.pharosresearch (since it includes more data providers).

Secondly, the aforementioned online catalogues and web portals are used to reproduce a common research

[3] `http://images.pharosartresearch.org`
[4] `http://catalogo.fondazionezeri.unibo.it`
[5] `http://www.wikidata.org`
[6] `http://www.wikipedia.org`

scenario. A scholar looks for the name of an artwork in several online resources and s/he retrieves some information on the authorship.

In order to evaluate H6, we consider the following sub-problems:

- H6a A user can find relevant information relatively faster when they use mAuth rather than pharos-research and online catalogues and web portals
- H6b A user can find relevant information by accessing a less number of pages when they use mAuth rather than pharosresearch and online catalogues and web portals

The two hypothesis are based on the idea that Semantic Web technologies positively impact users' expectations when performing common tasks such as collecting relevant information on the subject at hand.

In order to evaluate H7, we consider the following sub-problem:

- H7a The user's perception and satisfaction when validating internal grounds of authorship attributions is better when using mAuth rather than pharosresearch and online catalogues and web portals
- H7b The user's perception and satisfaction when validating a sorted list of attributions in mAuth is high

These hypotheses rely on the idea that Semantic Web technologies can respond to users' sophisticated information needs, such as supporting the validation of questionable information.

The user-centered evaluation is performed by using mAuth web application described in Chapter VII. Few tasks are assigned to users, and then they are asked to fill in an evaluation form. A number of questions allow to perform a quantitative and qualitative analysis on the degree of user's satisfaction when using the application. Since the mAuth application is developed as a proof of concept of the conceptual framework only, aspects such as the performance of the crawler, user's experience, and technical features of the framework are not evaluated.

Three tasks were designed to evaluate the aforementioned hypotheses, namely:

- Gather information on a well-known artwork avoiding time-consuming researches.
- Gather information on a less-known debated artwork whose authorship attributions are not sufficiently documented.
- Gather information on a debated artwork whose authorship attributions are well-documented.

Description of scenarios and tasks. In order to evaluate hypotheses outlined in the previous section we set up a task-based evaluation. Users performed assigned tasks remotely and filled in an evaluation form, from which we gathered data presented in the next section. Tasks are designed so as to reproduce the three mentioned common scenarios in connoisseurship, namely:

1. Look for documentation related to a well-known artwork that is currently unanimously ascribed to a certain artist. Only one domain expert's attribution is retrieved, and two less scholarly sources support the attribution. The research is performed in (1) three online catalogues and web portals, (2) Images.pharosresearch, and (3) mAuth. Users record the number of pages visited, the time required to retrieve relevant information, and compare the motivations underpinning the attributions when available.
2. Look for documentation related to an artwork whose authorship attribution is debated between two domain experts, and no sufficient evidences are provided to assess their veracity. Users are asked to rely on the internal grounds of data sources in order to evaluate the goodness of contradictory statements. Indexes representing cited scholars' authoritativeness are served along with results, so as to test their potential in such a situation.
3. Look for documentation related to an artwork whose authorship attribution is debated between three domain experts. In this case more insights are provided in order to evaluate the goodness of contradictory authorship attributions, and two sources are in agreement.

The first scenario. In the first scenario, named "Retrieve authorship attributions and assess their acceptance", we reproduce a realistic scenario where a user is required to complete the same task, i.e. a web research, in the three different systems in order to gather enough information on attributionship. Users are asked to

search for a given artwork, browse related web pages, and gather information on the authorship attributions recorded. The chosen artwork is the well-known painting "La Schiavona", currently ascribed to Titian, although it was formerly attributed to Giorgione. Users are introduced to the artwork, by showing a picture of it at the beginning of the evaluation test.

The research is firstly performed on three online catalogues and web portals, namely: the Zeri online catalogue, Wikidata and Wikipedia. The user is here asked to look for the title of the artwork by typing a string in a search interface. The same research is performed in images.pharosresearch, either by querying by string or by uploading an image and look for similar images. Third, the user is redirected to mAuth. The user is asked to input the URL of the corresponding cataloguing record found in the Zeri photo archive. Since the final aim of mAuth is to be integrated in such a catalogue, the user is supposed to have already reached such a web page before accessing information of the history of attributions. The user will find here the same three results (provided by the Zeri Foundation, Wikipedia and Wikidata) s/he already found in the first research. Retrieved attributions all agree on Tiziano.The Zeri's attribution gets the higher rating since it is the most documented, and it is the first shown in the list of results.

The second scenario. In the second scenario, named "Choose and motivate the most reliable attribution" the focus is on mAuth only. Users are redirected to the results of a research, which include two authorship attributions related to a less known artwork. The attributions are provided by two domain experts, i.e. the Zeri photo archive and Villa I Tatti. Users are asked to evaluate internal grounds of sources and finally to evaluate the goodness of the attributions retrieved.

The scenario here presented is the most complex one. Only two attributions are provided and these are in disagreement. Moreover, both the domain experts rely on scholars' opinions to support their statements. The Zeri photo archive relies on the archival classification of the photograph depicting the artwork, dated around 1990, and on Everett Fahy's attribution. Villa I Tatti relies on the archival creator's bibliography, i.e. Bernard Berenson's lists published in 1968. In this case the artist-related index and the acceptance rating of the two scholars are presented to the user in order to support the validation of the most authoritative attribution - though these are not taken into account in the ranking model. Everett Fahy's artist-related index is 2, and the acceptance rating is less than 1%. Bernard Berenson's artist-related index is 34, and his acceptance rating around 100%. According to the ranking model, the attribution provided by the Zeri

photo archive is more recent, and based on two concurring criteria, while Villa I Tatti provides an older attribution, supported by the archive creator's opinion only.

The third scenario. In the third scenario the user is redirected to mAuth and finds attributions related to the painting "The three graces". Attributions are provided by three trusted domain experts, namely: the Zeri photo archive, Villa I Tatti, and the Frick Art Reference Library. Discarded attributions recorded by the Zeri photo archive are shown too, for a total of four authorship attributions. Zeri and I Tatti agree on the same artist, i.e. Baldassarre Peruzzi, while the Frick Art Reference Library could not ascribe the artwork to a specific artist, hence an anonymous artist is recorded as accepted attribution. The Zeri Foundation motivates the attribution with the archival classification, dated around 1990, and a number of bibliographic references, including a Berenson's bibliographic reference dated 1968, that is in turn cited by Villa I Tatti as the main reason of attribution. The Frick Art relies on the archival classification only, which is dated 1952. The Zeri discarded attribution, which ascribes the artwork to the workshop of Bernardino Luini, is motivated by an auction attribution, dated 1994. Despite this is the most recent attribution, the criterion is less reliable (according to the ranking model) than the ones supporting other attributions, hence it is scored less. The order of results is: (1) Zeri photo archive accepted attribution, (2) Villa I Tatti accepted attribution, (3) Frick Art Reference Library accepted attribution, and (4) Zeri discarded attribution.

Evaluation measures. In all the scenarios users are asked to answer a number of questions related to their satisfaction. In particular, we measured four variables (V) in the first scenario (two quantitative variables and two qualitative variables), and two (qualitative) variables in the second and third scenario. Here below is provided a definition of the variables.

V1. Completion time. The completion time (CT) is the time span between the moment a user begins a task and the moment the retrieval task is accomplished. The CT metric is widely used to measure users' satisfaction with regard to the performance of the retrieval process [Kelly et al., 2009].

V2. Total pages visited. The total pages visited (TPV) measures the number of pages visited by a user in order to get the information required. It is measured for each retrieval task performed with a given system [Su, 1992]. The TPV metric measures the efficiency of the crawling system and the user satisfaction with regard to the retrieval information system.

Such quantitative measures apply to the first described scenario, i.e. retrieval of information sources related to a well-known artwork. However, these might not be sufficient to evaluate users' satisfaction, since the users' perception may vary according to their experience and background, and the difficulty of the task at hand [Cheng et al., 2010].

Two qualitative assessments, described below, aim at filling the gap related to the evaluation of user's satisfaction. Users are asked to provide a subjective feedback on their experience with the three systems, and secondly on the ranking of results.

V3. User satisfaction wrt the information retrieval process. The User Satisfaction of Information Retrieval Results (US) measure quantifies the user's satisfaction with regard to the results of the information retrieval. Specifically, it measures whether retrieved information are useful and sufficient to assess the goodness of an authorship attribution. Participants provide the measure by using a Likert scale from 1 to 5 (Strongly disagree to Strongly agree).

V4. User satisfaction wrt the ranking of results. The User Satisfaction of Ranking measure (USR) allows to quantify the user's satisfaction with respect to the ranking model and the suggested authorship attribution. In particular, two scores contribute to define the USR measure, namely:

- Rank Satisfaction Score (RSS). The RSS measure provides a feedback on the user's satisfaction with respect to the order of presented results and the score associated to each information source. Such a score is the final output of the ranking model outlined in Chapter VI.
- Perception of Authoritativeness Score (PAS). The PAS measure provides a feedback on the user's acceptance of the suggested authorship attribution, i.e. the attribution scored more than the others. It is based on the Net Promoter Score [Reichheld and Markey, 2011] for measuring the likeliness of a user to prefer, and eventually suggest and cite, a certain attribution over the others available.

Like the US measure, participants provide the RSS and PAS measures by using a Likert scale from 1 to 5 (Strongly disagree to Strongly agree).

Table 8.1 summarises the usage of metrics in the three scenarios. As aforementioned, the two quantitative metrics (CT and TPV) apply to the first scenario only, so as to compare the user satisfaction with respect to

Scenario	Online Catalogues	images.pharosresearch	mAuth
1	CT, TPV, US	CT, TPV, US	CT, TPV, US, RSS, PAS
2			US, RSS, PAS
3			US, RSS, PAS

Table 8.1: Metrics used in the user-center evaluation grouped by scenario

the three evaluated systems. The two qualitative metrics (US and USR) apply to all of the three scenarios. It is worth to notice that US applies to the three systems in the first scenario, and to mAuth only in the second and third scenarios. The USR measures apply to the evaluation of mAuth only in the three scenarios, since the other systems do not rank results.

Lastly, we collected feedbacks on users' preferences for improving the ranking model, including insights on their perception of the usefulness of indexes for cognitive authoritativeness. Users are asked to (1) select from a list the dimensions they deem relevant for ranking attributions according to the selected scenario, and (2) to provide a feedback on how scholars' authoritativeness scores would affect the ranking - if taken into account.

To be precise, in the second and third scenarios users were not told that the citation indexes do not affect the ranking, but most of them believed they were actually affecting it or that they should have affected it more. The aim is to study the user's reaction when an automatic method is applied to evaluate cognitive authoritativeness. Such a social experiment provides useful insights on how to tune the current ranking model and enable future work on the inclusion of cognitive authoritativeness and scholars' citation networks in the Arts and Humanities, which is further discussed in the conclusions.

Data collection. The data collection was conducted by using a survey online application, i.e. Google Form.[7] Users filled in the form remotely and submitted their answers to be analysed. The survey form includes the three aforementioned scenarios and is divided in three sections, namely:

- *Context information.* The description of mAuth and details about the ranking model are provided to let the user understand what is the objective of the evaluation. In order to understand the difficulties an user may find when using mAuth and evaluate results accordingly, we asked users to describe their background and profession.

[7] See the form at `https://goo.gl/forms/xDLwvCCaEFWm4D5h2`

- **Introduction to the scenario and related tasks.** The three scenarios are introduced, and the tasks to be performed are detailed. Specifically, users are warned of which data will be gathered for the sake of the evaluation, hence accuracy of answers is encouraged. All the scenarios are described by means of a title, e.g. "Choose and motivate the most reliable attribution", a picture of the artwork to be searched, so as to avoid possible mistakes in the retrieval, and an overall description of the tasks the user will be asked to accomplish. For each task a list of actions to be performed in a sequence is provided.

- *Task-related questions.* When the retrieval process is done, the user answers multiple choice questions with respect to the task accomplished. For each task, respectively three tasks in the first scenario and a single task in the second and third scenarios, a number of questions are presented. Once s/he filled the form s/he submit it and a spreadsheet is automatically populated.

Data collected from the survey are published online [Daquino, 2018b].

Description of users. We collected feedbacks from 25 users.[8] Users participated only once to the evaluation test. They all performed the same tasks and they did not know which systems they would have used, or which researches they would have been asked to perform. Table 8.2 shows users grouped by background and their related affiliation.

Users were invited to fill in the form by sending them a private invitation so as to accurately select both domain experts and relevant stakeholders in the other fields. Indeed, representativeness of participants is the key element of this evaluation. Users belong to some of the most important cultural institutions dealing with art historical data and other stakeholders in the Humanities and Computer Science were involved to get feedbacks from different points of view.

In particular, domain experts including art historians, data collection managers, and photo archivists are the main target users of the tool. We expect them to provide insights on the benefits and the drawbacks of semi-automatic methods for classifying authorship attributions. Digital Humanities researchers and Computer scientists currently working on scholarly data are expected to provide insights on effective strategies for ranking authoritative pieces of information - despite they are not domain experts and cannot judge the

[8] After the publication of this work other six participants filled in the form. Results may slightly differ when reproducing the analysis but no significative changes in percentages are recorded.

Background	N.	Affiliation
Art historian	1	Warburg Institute
	1	Max Planck Inst. for Art History
	1	Frick Art Reference Library
	1	University of Padua
	1	University of Bologna
	4	Italian Public Education System
	1	Getty Research Institute
	1	University of Rome
Collection manager	1	Getty Research Institute
	1	Yale Center for British Art
	1	Italian Ministry of Cultural Heritage and Activities (MiBACT)
	1	Paul Mellon Centre for Studies in British Art
Photo archivist	1	Federico Zeri Foundation
	1	Kunsthistorisches Institut in Florenz
	1	Bibliotheca Hertziana - Max-Planc Institut
	1	Italian Ministry of Cultural Heritage and Activities (MiBACT)
DH scholar	1	University of Bologna
	1	University of Lausanne
Computer Scientist	1	University of Bologna
	1	Knowledge Media Institute - Open University
Other	1	University of Milan
	1	University of Florence

Table 8.2: Population of the User study

goodness of an authorship attribution. Finally, we included few users with heterogeneous backgrounds in the Arts and Humanities that have a basic knowledge of art history, so as to evaluate the soundness and the usefulness of the proof of concept when used in the context of similar researches.

8.3 Results of the user-centered evaluation

In this section are presented the results of the user-centered evaluation of the conceptual framework mAuth. Results are grouped by metric and scenario.

Completion Time (CT) measure. Figure 8.3 shows the time required for participants to perform the tasks included in the first scenario, grouped by system used. The CT measure is calculated for the three systems in scope, namely: three online catalogues, images.pharosresearch and mAuth.

The average time is calculated on the basis of the CT measure, and is respectively: 04:05 minutes for search-

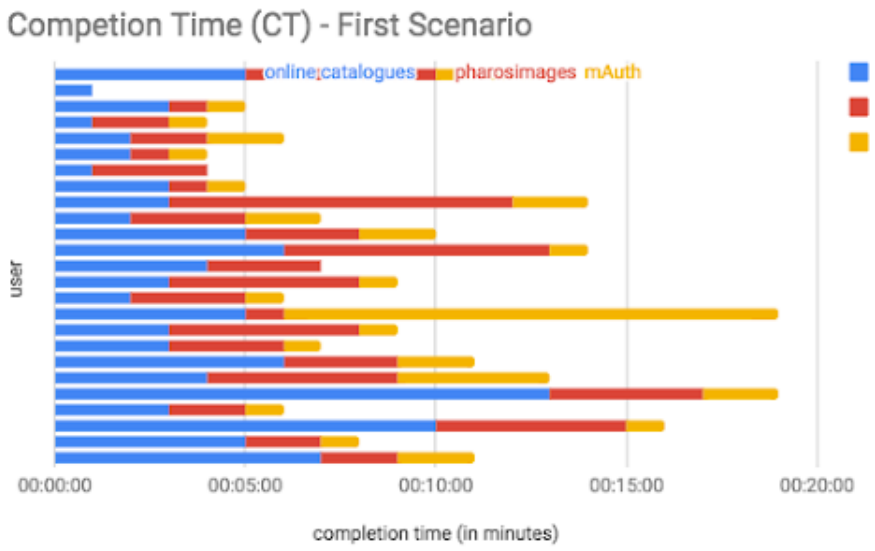


Figure 8.3: Completion time for completing the first scenario in online catalogues, pharosresearch, and mAuth

ing with the three online catalogues, 03:12 minutes for searching with pharosresearch, and 01:50 minutes for searching with mAuth. It is worth to notice that some users had difficulties when using pharosimages, and some were not able to find the artwork at hand. When performing searches using mAuth one user faced WiFi connection problems (according to comments), hence results of the CT measure may be error-prone, and the research took more time.

Results show that the retrieval of the same number of information sources in mAuth requires 55% less time than a traditional research by using online catalogues, and 42% less time than a more sophisticated research in pharosresearch. Results validate the initial hypothesis *H6a (A user can find relevant information relatively faster when they use mAuth rather than pharosresearch and online catalogues and web portals)*.

Total Pages Visited (TPV) measure. Figure 8.4 shows the total number of pages visited by users to complete the first task in the first scenario.

Users were expected to open at least 7 pages in order to get the three web pages describing the artwork "La Schiavona" in three online catalogues. Some users were not able to reach all the requested web pages, hence they answered with a lower number of pages corresponding to the number of pages visited in order to reach one or two web pages out of the three requested (between 2 and 7). We normalised errors to 7, so as to get significant results for the TVC measure. So doing, the TPV measure shows that, on average,

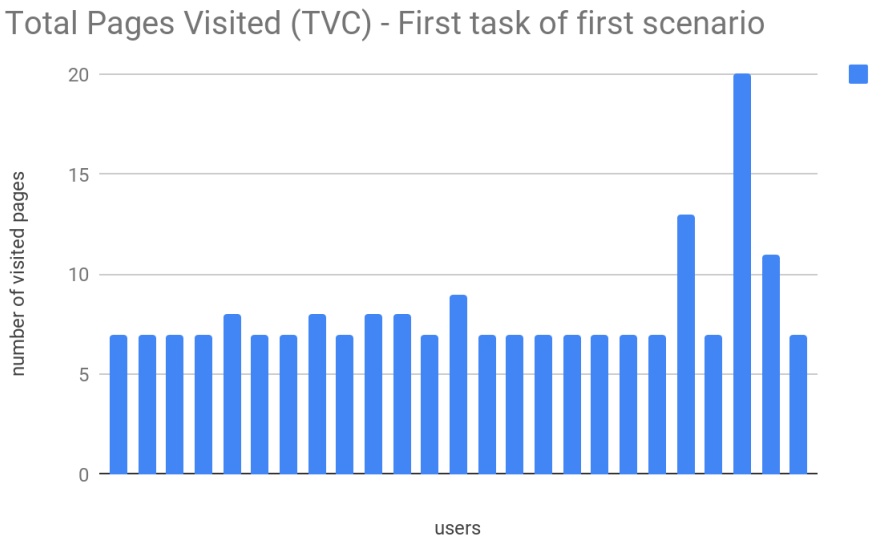


Figure 8.4: Total number of pages visited by users for completing the first scenario in online catalogues, pharosresearch, and mAuth

a user visited 8.16 pages in order to get to desired results.

In pharosresearch users were asked to input the URL of an image retrieved in one of the three aforementioned catalogues and look for the similarity match. We did not ask users to record the number of visited pages in pharosresearch, since results depend on the prior research, and are likely to be error-prone as well. Assuming they looked for the Zeri cataloguing record, which requires to visit 3 pages, and 2 pages to retrieve related results in pharosresearch, we assume a total of 5 pages were visited. However, it is worth to notice that 17 participants out of 25 participants were not able to find results because no matches were found.

Likewise, users were asked to input in mAuth the URL of one of web pages retrieved during the first task and they got results immediately. Assuming they input the Zeri cataloguing record, they visited 5 pages in total to get to the final list of results. Results showed in mAuth include all of the three authorship attributions retrieved in task 1.

In summary, when using mAuth a user is required to visit the same number of pages as in the images aggregator pharosresearch, i.e. 5 pages, and 38,7% less pages than in multiple online catalogues. Results confirm the initial hypothesis *H6b (A user can find relevant information by accessing a less number of pages when they use mAuth rather than pharosresearch and online catalogues and web portals)*.

User satisfaction wrt the information retrieval process (US). In order to evaluate the US measure we first gathered users' feedbacks on the same task performed in the three systems, namely: look for the artwork "La Schiavona" in three online catalogues, pharosresearch, and mAuth. Figure 8.5 shows the results of the comparison of US measure in the three systems. Users were asked to provide a feedback by using the Likert scale (from Strongly agree to Strongly disagree) when answering the question "Was it easy to find sufficient information for validating the most authoritative authorship attribution?".

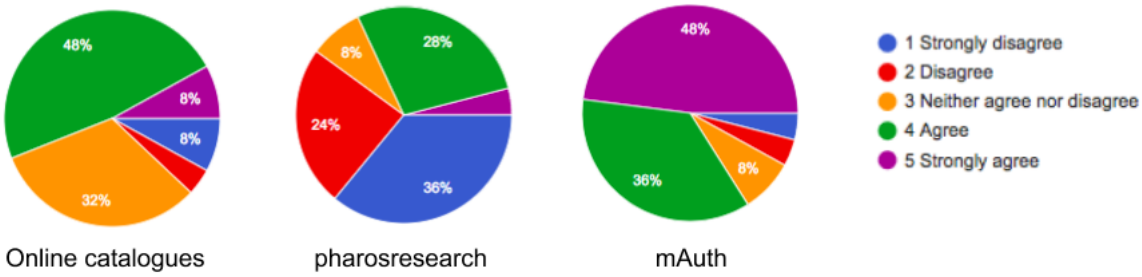


Figure 8.5: User satisfaction with respect the usage of online catalogues, pharosresearch, and mAuth in the first scenario

Results demonstrate that users can generally retrieve the information seeked and assess the veracity of authorship attributions in online catalogues (56% of participants either agree or strongly agree) and in mAuth (84%), while they have more difficulties in pharosresearch (only 32% of participants agree or strongly agree). Such a first comparison shows that mAuth offers a valid alternative to existing services when retrieving attributions related to well-known artworks.

In order to better understand benefits derived by using mAuth, we gather feedbacks also in the second and third scenarios, wherein mAuth only is evaluated. The two scenarios require users to judge contradictory attributions related to less known artworks. In Figure 8.6 are shown users' feedbacks for the two scenarios. Users used the same Likert scale (from Strongly agree to Strongly disagree).

Results demonstrate that the US measure is still high in the third scenario (84%) and lower in the second scenario (56%). Such a discrepancy is due to the different level of complexity of the two scenarios. In the second scenario users are asked to evaluate the goodness of less documented attributions, while in the third scenario more insights are provided. Such results, along with results of the first evaluated scenario, contribute to validate hypothesis *H7a (The user's perception and satisfaction when validating internal grounds of authorship attributions is better when using mAuth rather than pharosresearch and online catalogues and web*

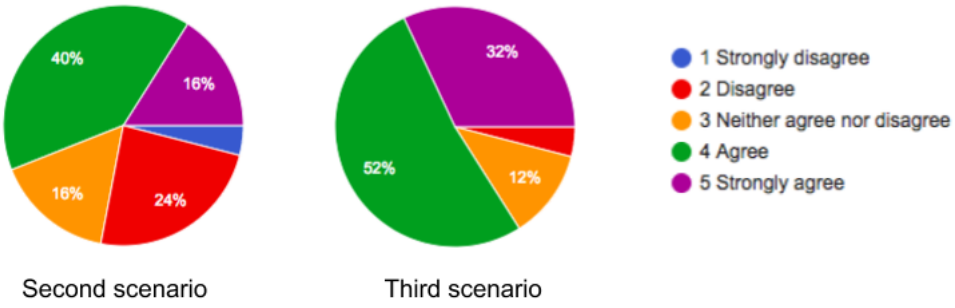


Figure 8.6: User satisfaction with respect to the usage of mAuth only in the second and third scenarios

portals).

User satisfaction wrt the ranking of results (USR). In order to validate hypothesis *H7b (The user's perception and satisfaction when validating a sorted list of attributions in mAuth is high)*, we collected two specific feedbacks related to the usage of mAuth in the three scenarios.

Figure 8.7 shows results of the RSS score, which reflects the user's satisfaction with respect to the ranking of the attributions, and the PAS score, which provides a feedback on the user's perception on the most authoritative attribution highlighted in the list of results. Together the two scores address the USR measure.

In particular, to evaluate the RSS measure, users were asked to answer the question "Do you agree with the ranking of results (i.e. the score attributed to each provided attribution and the order in the list)?". To evaluate the PAS measure, users answered the question "Do you agree with the suggested attribution?", which is meant to evaluate the goodness of the attribution with the highest rank.

In the first scenario all the attributions agree on the same artist and the most ranked attribution is provided by the Zeri Foundation. Results show that the RSS measure returns positive values in the 72% of the cases, and the PAS measure in the 88% of the cases.

As already mentioned, the second scenario is the most complex one. The two retrieved attributions are in disagreement and both cite scholars' attributions as motivation. In this case, scholars' citation indexes are served along with results, but these do not affect the ranking model. The most ranked attribution is the most recent one and it is supported by two criteria, namely the Zeri archival classification and a scholar (Everett Fahy) with low citation indexes. The less ranked attribution is less recent and cites as motivation

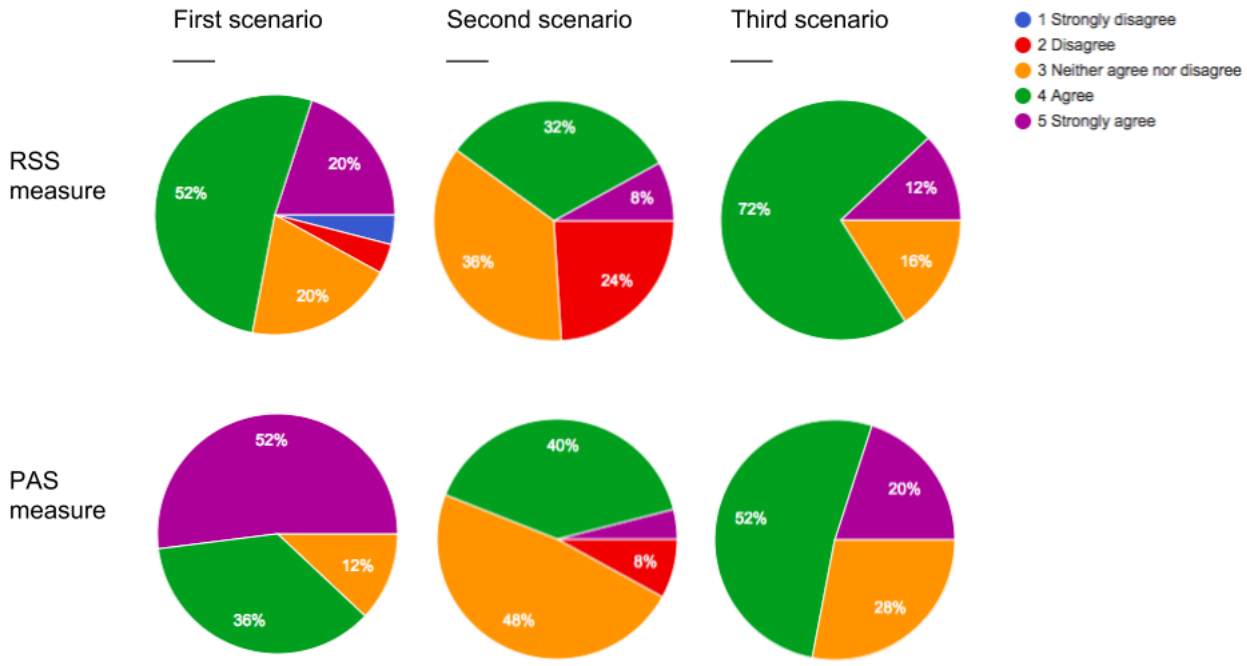


Figure 8.7: User satisfaction with respect to the ranking of attribution (RSS) and to the highlighted attribution (PAS) in mAuth, in the three scenarios

only a scholar (Bernard Berenson), that nonetheless has far higher citation indexes. The RSS measure in this scenario is 40% and the PAS measure is 44%. It is worth to notice that 36% of participants were not able to judge with respect to the ranking, and 48% of participants were not able to judge with respect to the most ranked attribution. The negative feedback on the ranking is estimated as 24%, and the negative feedback on the most authoritative attribution is 8%.

In the third scenario four attributions are retrieved. More insights and documentation are offered to the user in order to judge the goodness of the most ranked attribution. Two domain experts are in agreement, attributions are recent, well documented, and cite a wide bibliography. In this case, the RSS value is positive for the 84% of participants and the PAS value is positive for the 72% of participants. No negative feedbacks are recorded.

8.4 HiCO Ontology evaluation

The second and third scenarios are used for the evaluation of the HiCO Ontology. The aim is to show benefits and limits of our approach. In particular, HiCO terms presented in Chapter V correspond to the

internal grounds of a questionable piece of information that has to be evaluated by the user. HiCO terms were used by three out of six trusted providers, are taken into account in the conceptual framework, and are further elaborated in the ranking model. The web application serves to users a comparative analysis of retrieved attributions and shows values annotated by using predicates addressed in HiCO. Such terms and predicates address the following features:

- The date of the questionable statement.
- The criterion adopted to motivate the questionable statement.
- The primary source of information, e.g. a scholar, a museum, an auction firm.
- The secondary source recording the questionable statement, e.g. a cataloguing record.
- The agreement or disagreement with other questionable statements.

Users' satisfaction when using mAuth (i.e. 84% in the first and third scenarios, 56% in the second scenario) and users' feedback on ranked results (respectively 72%, 40% and 84% in the three scenarios) demonstrate that data retrieved are sufficient to assess the veracity of statements in most of the cases, and that the consequent ranking is useful.

In the second scenario we recorded a lower positive feedback, due to (1) the lack of sufficient information for validating the goodness of retrieved authorship attributions, since only two contradictory statements are found, and to (2) the need of bespoke measures in order to assess scholars' cognitive authoritativeness. However, such aspects are not in the scope of the ontology and do not affect its evaluation.

In summary, the HiCO ontology can be deemed a valid means for art historical research activities since it offers (1) a terminological basis for an ontology-based data integration, (2) a framework of terms for evaluating textual authoritativeness of questionable statements, and (3) resulting users' satisfaction is high in two out of three common scenarios in the domain at hand.

8.5 Discussion

The user-centered evaluation shows the benefits derived by using mAuth and the HiCO ontology in art historical research activities. It reveals that Semantic Web technologies can effectively support scholars' tasks, such as gathering information, analyse internal grounds of information sources, compare sources, and efficiently support the decision-making process. Moreover, feedbacks provided by participants with a heterogeneous background show that the soundness of the framework is confirmed in other fields of the Humanities, which could be explored by applying the ranking model to other types of sources and information.

The limits of such an approach to rank questionable statements are highlighted in the second scenario. When sources are not well-documented or these rely on domain experts' authoritativeness only, textual authoritativeness is not sufficient to validate the goodness of contradictory statements.

However, negative results recorded in the second scenario provide some insights on the impact that citation metrics have on users' expectations and judgement. A preliminary assessment was conducted by asking users to judge whether citation indexes positively affect the ranking of results. Figure 8.8 shows users' feedback on the perceived relevance of citation metrics in the second and third scenarios.

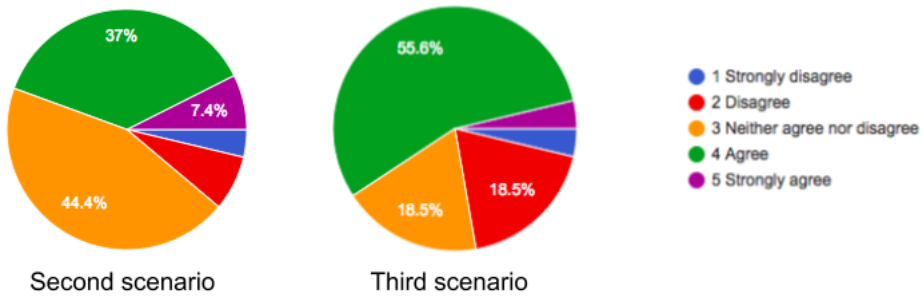


Figure 8.8: Users' perception of citation metrics in the second and third scenarios

In the second scenario, data gathered are not sufficient to estimate users' satisfaction of the usage of citation indexes, since we cannot clearly address why the 44.4% of participants was not able to judge. Moreover, the question misled users into thinking that citation metrics somehow affected the ranking model. If such metrics were taken into account in the ranking model, the order of the two attributions would have been swapped (since Bernard Berenson's indexes are far higher than Everett Fahy's indexes). Comments to the

questionnaire revealed the disappointment of some participants, who perceived the actual ranking as an error.

In the third scenario, where the ranking of results is confirmed by the citation indexes of cited scholars, feedbacks are clearer: 59.3% of users agree on the benefits of such indexes. However, it is not sufficient to claim that the ranking model should take into account such metrics.

In order to evaluate limits of textual authoritativeness and support the assumption that cognitive authoritativeness is the key element when evidences are not sufficient, we collected users' feedback on the dimensions they deem relevant in order to rank authorship attributions in each scenario. At the end of each task participants were asked to answer the following question "Which criteria would you deem relevant to rank results?". Results of the questionnaire are shown in Figures 8.9, 8.10, and 8.11, corresponding to the three scenarios.

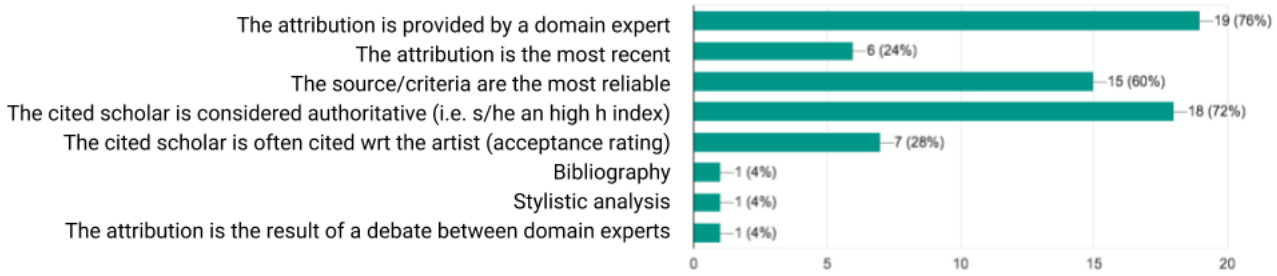


Figure 8.9: Users' feedback on criteria deemed relevant for ranking results retrieved in the first scenario

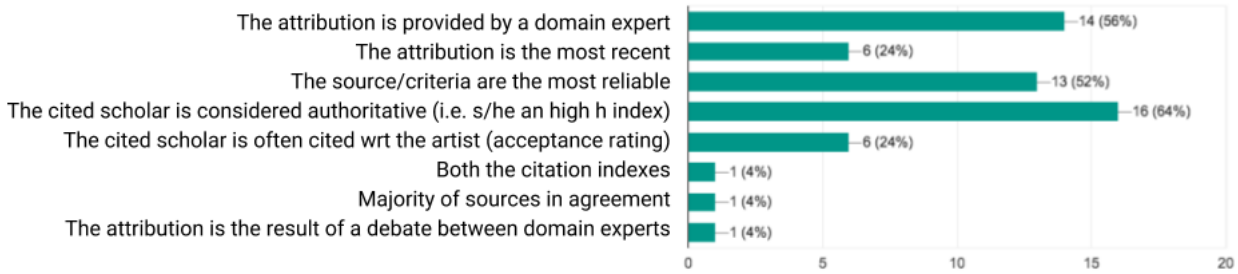


Figure 8.10: Users' feedback on criteria deemed relevant for ranking results retrieved in the second scenario

Five dimensions, described by the first five bars in each of the above graphs, were suggested to users in a multiple choice list, and the user had to check the box if s/he agreed with the relevance of the dimension at hand. Some users added few other dimensions, which can be anyway grouped with the aforementioned

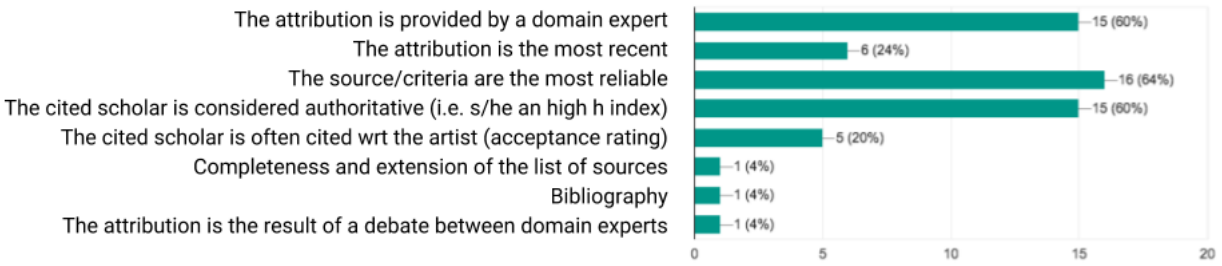


Figure 8.11: Users' feedback on criteria deemed relevant for ranking results retrieved in the third scenario

five ones.

It is worth to notice that in all of the three cases, the majority of users agreed on (1) the importance of the provenance of the attribution, i.e. "the attribution is provided by a domain expert", on (2) the importance of being deemed authoritative in the community, i.e. "the cited scholar is considered authoritative (s/he has an high h-index)", and (3) the relevance of criteria supporting the attribution, i.e. "the source or the criteria underpinning the attribution are the most reliable". Less users agreed on the importance of the date of the attribution, i.e. "the attribution is the most recent one", and on the acceptance rating of scholars, i.e. "the cited scholar is often cited with regard to the artist (acceptance rating)".

Such a situation is only partially reflected in the current ranking model, namely:

- The data provider is labelled as domain expert or not in a list of trusted providers and related attributions are ranked accordingly.
- The citation indexes are served along with results, but these do not affect the ranking model. The h-index and the acceptance rating mentioned in the questionnaire correspond to indexes described in Chapter VI.
- The criteria underpinning the attribution are already the most weighted dimensions in the current ranking model.
- The date of the attribution is currently low weighted than other dimensions in the ranking model, and it reflects users' expectations.

In summary, users' perception on the current ranking model and the conceptual framework is positive

when information on cited scholars' cognitive authoritativeness is not fundamental for the sake of the judgement, or when it confirms the actual ranking. Improvements in the ranking model will have to be taken into account when textual authoritativeness is not sufficient. However, as already mentioned, providing reliable and comprehensive citation indexes in the Arts and Humanities is challenging, due to the infancy of research on historical citation networks, and it will deserve attention in future works.

Chapter 9

Conclusion

This thesis addresses three research problems related to ontology-based data representation, assessment and consumption in the context of art historical photo archives. In particular, we aim at facilitating knowledge discovery in the Arts and Photography domain in Cultural Heritage, so as to support connoisseurship related activities, such as gathering authorship attributions in the Web and support the decision-making process. At the same time, we aim at providing effective means for data quality improvement in art historical photo archives that serve such information.

Three main challenges are faced in this work, namely: (RP1) The formal representation of questionable information in the Photography and Arts domains by leveraging well-grounded formal languages and technologies; (RP2) The formalisation of the dimensions characterising the methodology of art historical data providers when publishing questionable information; (RP3) Support users' decision-making process when assessing reliability of authorship attributions. Seven hypotheses were formulated and were discussed throughout this work. We systematically review them here below along with related contributions, in the light of the evidences presented in previous chapters. Finally insights on the expected impact and limitations of the current research are presented and future research plans are outlined.

9.1 Hypotheses and contributions

The main contribution of this thesis is the conceptual framework of IQ measures addressing textual authoritativeness of authorship attributions recorded by art historical data providers. The framework encompasses a number of contributions that aim at tackling the aforementioned research problems. The framework and our seven hypotheses were evaluated by using the design-science method proposed by Hevner [Hevner et al., 2004], described in Chapter III.

The first two hypotheses we formulated regard issues related to knowledge representation in the Arts and Photography domain in Cultural Heritage.

H1. We can reuse existing ontologies for representing information included in cataloguing records produced by art historical photo archives. This hypothesis is formulated on the basis of the rich literature related on ontologies for the Cultural Heritage (CH) domain presented in Chapter II.

However, the heterogeneity of descriptive approaches adopted by stakeholders affects the definition of a shared, comprehensive data model capable to represent all the information needed to support connoisseurship activities.

To overcome such heterogeneous representations we (1) reviewed the most used ontologies in the CH and (2) we designed two bespoke ontologies, called FEntry Ontology and OAEntry Ontology, where we gathered existing and new terms. The two ontologies collect terms mainly from CIDOC-CRM and the SPAR Ontologies for describing photographs, artworks depicted, bibliographic references, and archival documents. The ontologies are the result of the mapping to RDF of two of the most comprehensive cataloguing standards, as shown in Chapter I, namely ICCD-F and ICCD-OA.

The two ontologies were evaluated by relying on a data-driven ontology development and evaluation methodology, called SAMOD. In Chapter IV we described how we transformed a subset of the Zeri photo archive into RDF according to the aforementioned ontologies, which shows that we reached a comprehensive representation of the domain by applying the two ontologies on a representative use case. In Chapter V we showed how such ontologies are obtained by comparison with two golden standards in the domain, namely CIDOC-CRM and the FRBR conceptual model. In Chapter VIII we showed a sample of the iterative mapping process from ICCD-OA and ICCD-F standards to RDF, so as to demonstrate

how ontologies can be deemed consistent.

H2. The interpretative process that generates questionable information can be effectively represented by using Semantic Web technologies, such as ontologies. This hypothesis is formulated on the basis of the limits in the state of the art in the formal representation of questionable information, as shown in Chapter II.

The contribution of this thesis is the development of a task ontology, called HiCO Ontology, that takes into account the main representational approaches in the CH, such as provenance-aware and interpretation-driven descriptive approaches, and reuses existing ontologies for describing questionable information. In particular HiCO is developed as an extension of the PROV Ontology and some terms are aligned to CIDOC-CRM.

The HiCO Ontology is described in Chapter V, and its evaluation is illustrated in Chapter VIII. In particular, we evaluated the conceptual framework underpinning an online application, called mAuth, leveraging the HiCO Ontology in an ontology-based ranking model. Results of the user-centered evaluation show that in two cases out of three the information is sufficient and useful to assess the veracity of a statement, while in one case more insights would be needed for the task. However, in the latter case negative results depend on the low number of sources available rather than the type of information recorded and knowledge organisation aspects.

We formulated three hypotheses related to the definition of dimensions characterizing the methodology of art historical providers.

H3. Analytical data and domain experts' feedback can be used to formalize the criteria underpinning the methodology of art historical data providers when publishing authorship attributions. This hypothesis is motivated by the absence of detailed rules on how cultural heritage institutions publish information in secondary sources. As detailed in Chapter II and Chapter VI, cataloguing rules provide controlled vocabularies for describing motivations supporting questionable attributions, but no rules on how to rank contradictory sources are given.

The contribution of this thesis is the data analysis of the methodologies of art historical photo archives when publishing authorship attributions associated to artworks depicted in photographs. The analysis of

the dataset of the representative use case (the Zeri photo archive), reviewed by domain experts, and the comparative analysis of this with other two archives, namely Villa I Tatti and the Frick Art Reference Library, resulted in the definition of a rating of criteria to be used in the ranking model.

We evaluated the rating of criteria in Chapter VIII along with the ranking model. Indeed, the rating of criteria is the most weighted parameter in the ranking model. When assessing users' satisfaction with respect to the ranking model, we are implicitly evaluating the rating too. In two cases out of three the ranking is positively perceived. In the third case, i.e. the second scenario of the user-centered evaluation, the rating is affected by the subjective perception of cited scholars' cognitive authoritativeness, that may be deemed more relevant than the reliability of criteria.

H4. The evaluation of textual authoritativeness of sources recording authorship attributions can be based on a documentary, evidence-based approach. This hypothesis is based on the distinction between textual authoritativeness and cognitive authoritativeness illustrated in Chapter VI. The review of dimensions characterising information quality illustrated in Chapter II showed a significant gap when applied to cultural heritage data, and the absence of case studies on art historical data.

The main contribution of this thesis is the design of a conceptual framework for assessing textual authoritativeness of secondary sources recording authorship attributions. The framework takes into account a number of well-known measures for assessing information quality in the web, pruned in order to fit art historical data. This includes features such as: authoritativeness of data providers, rating of criteria motivating the attribution, timeliness of the attribution, number of agreements.

The conceptual framework is evaluated in Chapter VIII by means of the user-centered evaluation. The survey shows that textual authoritativeness is sufficient and useful for validating the most authoritative attribution in two out of three common scenarios in art historical research tasks, namely:

- Retrieval of attributions in agreement related to well-known artworks, including a domain expert only and a number of less scholarly sources.

- Retrieval of contradictory, well-documented attributions related to less known artworks, all provided by domain experts, and characterised by a significant discrepancy in information quality.

Only in the second scenario, where two contradictory attributions rely on cited scholars' authoritativeness only, textual authoritativeness is not sufficient for the assessment. In such cases, more insights on cognitive authoritativeness are needed.

A second contribution, based on the design of the conceptual framework and on the survey of policies in art historical photo archives, is the definition of a number of policies for information quality improvement. Moreover, a low-cost data integration process based on Batini C. et al. [Batini et al., 2009] is defined.

Five out of seven steps outlined in Chapter VI have been developed by means of mAuth crawler. Data is available for being integrated in online catalogues by means of the mAuth API. The last two steps of the workflow, namely error detection and cost optimisation, will be object of a dedicated trial with the Zeri photo archive, detailed in the next section (Impact of research).

H5. Measuring scholars' authoritativeness in the arts field can be achieved by developing bespoke metrics. This hypothesis is motivated by the lack of available citation indexes for most of the art historians that worked in the last two centuries. Moreover, the heterogeneity of types of citations in the Arts and Humanities require bespoke metrics to be developed.

The contribution of this thesis is the development of two bespoke metrics for evaluating scholars' authoritativeness in two contexts, namely: the artist-related index, which represents the perception of scholars' authoritativeness in the community, and the acceptance rating, which represents the perception of scholars' authoritativeness with respect to a given artist.

However, the two metrics are not completely evaluated in this work, since they are in a too early stage. Only three data sources have been harvested in order to obtain the aforementioned scores, and further work should be done in the Arts and Humanities literature, which is out of scope in this work. Nonetheless, the relevance and the potential of such metrics is demonstrated by users' feedback on the criteria they deem relevant for ranking results, showed in Chapter VII. Between 60% and 72% of users consider h-index similar metrics fundamental for assessing the validity of a questionable statement.

Finally, we formulated two hypotheses related to the assessment of reliability of statements in connoisseurship related activities by means of Semantic Web technologies.

H6. Linked Open Data and Semantic Web technologies can support and satisfy common require-

ments of research activities in the Arts and Humanities. This hypothesis is based on the assumption that Semantic Web technologies are suitable for achieving common objectives in the Arts and Humanities, such as gathering sources of information and provide insights on their internal grounds.

The contribution of this thesis is the development of the mAuth semantic crawler, which harvests authorship attributions from six trusted Linked Data providers and serves structured information on the internal grounds of information sources.

The mAuth crawler is evaluated in Chapter VIII and is compared to two competing systems, namely: three online catalogues and an image aggregator called pharosresearch. The evaluation shows that retrieving information with mAuth requires 55% less time than a traditional research by using online catalogues, and 42% less time than a more sophisticated research in pharosresearch. Secondly, a user is required to visit the same number of pages as in the images aggregator pharosresearch, i.e. 5 pages, and 38,7% less pages than in multiple online catalogues.

H7. Automatic and curated methods can support the decision-making process in connoisseurship activities. This hypothesis is based on the aforementioned assumption that Semantic Web technologies are suitable for achieving common objectives in the Arts and Humanities, and can satisfy users' sophisticated information needs.

The contribution of this work is the implementation of the mAuth recommender system, described in Chapter VII. mAuth leverages the conceptual framework described in Chapter VI in a ranking model that sorts authorship attributions according to a number of features. Since the goodness of authorship attributions can only be assessed on the basis of a subjective analysis, an evaluation of the recommender system can only be achieved by relying on users' perception and satisfaction. The user-centered evaluation in Chapter VIII shows that 84% of users are satisfied when using mAuth in two out of three common research tasks, and 56% in a more complex scenario. Moreover, 72% and 84% of users are satisfied of the ranking of results respectively in the first and third scenario; only the 44% of users are satisfied of the ranking of results in the second scenario. Similarly, 88% and 72% of users are satisfied of the most ranked attribution in the first and third scenario respectively; only 48% of users are satisfied in the second scenario. As above explained, the second scenario represents the impact of cognitive authoritativeness in the assessment of questionable information, which is not currently taken into account in the ranking

model due to the limits in providing reliable citation metrics for the Arts and Humanities field. However, the promising results encourage further research on how to tune the ranking model so as to accomplish more complex tasks, detailed in the last section of this chapter.

9.2 Impact of research

Besides the above described contributions, mAuth shows that the potential of methods developed can be exploited in a number of similar situations and application fields. As described in Chapter VI, technologies like mAuth can effectively lower barriers in expensive and time-consuming tasks related to the cataloguing process undertaken by cultural heritage institutions. For instance, the image matching process performed on the Zeri photo archive showed that around 940 cataloguing records of artworks overlap with Villa I Tatti - Berenson Library records - corresponding to the 5% of records provided by the Zeri Foundation and the 8% of records provided by Villa I Tatti for pursuing this research.

To this extent, not only information on authorship attributions could be integrated in online catalogues, but also other missing or partial information. For instance, the data collection provided by Villa I Tatti, called “homeless”, could be integrated with updated information on the provenance of artworks - that is here completely missing - which is in turn accurately annotated in the Zeri photo archive. In fact, mAuth crawler is highly customisable, and allows users to change the following parameters by changing related configuration files:

- The linksets where to look into for equivalence links
- The sources to be accessed for retrieving the information at hand
- The methods for accessing data sources
- The type of information to be sought in information sources

Users are also encouraged to tune the query that retrieves information into the so generated knowledge base, i.e. the mAuth attributions graph, and to modify the ranking model according to their preferences.

Likewise, less scholarly information sources, such as DBpedia and Wikidata may want to update their information on artworks by referencing trustworthy sources of information. The mAuth API accepts as input the persistent URI of one of the six trusted providers addressed in this study, including the aforementioned DBpedia and Wikidata URIs.

Lastly, the methodology assessment in art historical photo archives, detailed in Chapter VI, can be a valid basis for similar researches in other fields of the Arts and Humanities. The analytical and comparative approaches adopted to validate domain experts' assertions and their archival policies, i.e. the ten steps for assessing how archives publish questionable information, demonstrated to be a highly curated and reliable process for the analysis of stakeholders' methodologies, which can fit for other research fields, such as philological and historical analysis of text sources.

9.3 Limitations

Limitations of this work mainly concern the boundaries of automatic methods for the assessment of authoritativeness of questionable information.

First, primary sources, i.e. expertises, archival documents, and bibliographic resources created by connoisseurs are not analysed. We base the conceptual framework on second-hand knowledge providers, that is, cultural institutions that cite the former sources in secondary sources, i.e. cataloguing records. We assume that data providers' authoritativeness can be inherited by their statements. However, we currently lack of any quality control mechanism on types of cited sources. Cultural institutions' choice of sources may be biased, e.g. citation of an archive creator's attribution regardless a detailed review of literature, citation of attributions that are biased by market interests. As a consequence, we are not able to predict the goodness of a provider's intention when validating the goodness of the assertion itself.

Secondly, evaluated sources include online sources created by cultural institutions by means of a non-standardised cataloguing process. Therefore sources may be affected by information quality issues, due to lack of resources, time, and expertise. Moreover, museums, galleries and other types art historical data providers are not included at this stage of the research since these do not record any information on motivations supporting attributions and they would not provide any relevant insight on how to improve

the ranking model itself. In fact, if these were included, they would have been penalised in the final ranking. To this extent, we acknowledge the need to explore new metrics for measuring uncertainty of providers' authoritativeness and balance the final ranking model accordingly.

Lastly, the user-centered evaluation showed that limits of an evidence-based approach are evident when available information depends on cognitive authoritativeness only. In such cases, users' subjectivity when evaluating contradictory information is still high and hard to be evaluated. We acknowledge there is the need of more sophisticated metrics for evaluating scholars' cognitive authoritativeness - other than traditional citation indexes. As shown in our preliminary evaluation of citation indexes for measuring scholars' credibility, we need flexible metrics to be applied in narrow contexts (i.e. a limited number of available datasets) and that take into account peculiarities of citations in the Humanities, such as references to notes on photographs, archival documents and verbal communications. To this extent, our proposed metrics envision a new research line to be explored in future work.

9.4 Future Work

This work provides a number of contributions in the field of (1) information retrieval for connoisseurship activities, in (2) data quality improvement in art historical photo archives, and in (3) supporting users' decision-making process when validating authorship attributions. However, throughout the description of the work done, we have highlighted a number of limits, open issues and new potential research directions. Specifically, we foresee two research directions will be the focus of future works, namely: Enhancing information retrieval in the field of connoisseurship by increasing the number of data sources to be harvested. The creation of citation data and an in depth analysis of citation networks in the field of Arts and Humanities.

In particular, future work aims at leveraging soon-to-be-published linked datasets of PHAROS partners,[1] which already include the three analysed photo archives. Secondly, the Linked Art[2] project will publish a bucket of relevant data sources belonging to museums and galleries. Despite museums and galleries do not generally record detailed information on the documentation supporting authorship attributions, these

[1] `http://pharosartresearch.org/institutions`
[2] `https://linked.art/`

are nonetheless authoritative data providers that can contribute to validate statements in photo archives catalogues, and that may in turn benefit of mAuth findings.

By increasing the number of data sources, we expect we will have to extend some aspects of this research, namely the list of criteria and the ranking model. This will allow us to generalise results on the basis of a higher number of stakeholders and achieve iteratively a shared conceptual framework for defining textual authoritativeness in the art historical research field.

Such an extension of the number of data sources will provide also the grounds for extending services provided by mAuth. In particular, the extraction of bibliographic references and other types of scholars' citations will allow us to improve citation metrics developed for the sake of the proof of concept, and to provide links to full-text sources, so as to allow users to analyse the primary sources that corroborate a questionable statement.

In particular, we aim at focusing on the creation of curated citation data that encompass different types of citations that include detailed information on the function of the citation itself, such as agreement, disagreement, citation of evidences, etc. We aim at developing a bespoke knowledge base that gathers citation data and provides a number of services, such as the computation of citation indexes on-the-fly, so as to be reused in the context of mAuth and other applications.

To this extent, we have started reconciling art historians to the Duke University Dictionary of Art Historians[3] records, which provide an extensive bibliography for plenty of art historians. We plan to extend this work and gather more data sources that can contribute to shape art historians' authoritativeness.

Moreover, we recently developed BCite [Daquino et al., 2018], a bibliographic correction service that provides users a user-friendly tool for cleaning bibliographic data of an input article and creates at the same time RDF citation data according to the OpenCitations model [Peroni and Shotton, 2018a]. By continuing our collaboration in the field of open citation data we hope to (1) provide new grounds on the definition of cognitive authoritativeness in the Arts field, (2) further tune the ranking model we developed, and (3) provide a satisfying tool for performing more complex tasks related to connoisseurship activities.

Lastly, we foresee a data integration trial at the Federico Zeri photo archive, so as to leverage findings of this

[3] `http://arthistorians.info/`

research in a real scenario. Specifically, we will enrich the Federico Zeri Foundation online cataloguing records with the competing attributions retrieved by means of mAuth. Attributions will be fetched by leveraging the mAuth API and will be integrated in records by means of client-side scripts that will show the history of attributions on demand. This approach implies that no efforts are required on the Zeri database side, while it will facilitate data cleansing activities that will be performed on the mAuth knowledge base directly.

Bibliography

[Aitchison, 1977] Aitchison, J. (1977). Unesco thesaurus: A structured list of descriptors for indexing and retrieving literature in the fields of education, science, social science, culture and communication. comp. Technical report.

[Alexander and Meehleib, 2001] Alexander, A. and Meehleib, T. (2001). The thesaurus for graphic materials: Its history, use, and future. *Cataloging & Classification Quarterly*, 31(3-4):189–212.

[Alexander, 2008] Alexander, K. (2008). RDF in JSON: a specification for serialising RDF in JSON. *SFSW 2008*.

[Alexander et al., 2009] Alexander, K., Cyganiak, R., Hausenblas, M., and Zhao, J. (2009). Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In *Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09*. Citeseer.

[Antoniou and Van Harmelen, 2004] Antoniou, G. and Van Harmelen, F. (2004). Web ontology language: Owl. In *Handbook on ontologies*, pages 67–92. Springer.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

[Baca and Harpring, 2009] Baca, M. and Harpring, P. (2009). Categories for the description of works of art (CDWA).

[Barthe, 2000] Barthe, C. (2000). De l'échantillon au corpus, du type à la personne. *Journal des anthropologues. Association française des anthropologues*, (80-81):71–90.

[Batini et al., 2009] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):16.

[Belovari, 2013] Belovari, S. (2013). Professional minutia and their consequences: provenance, context, original identification, and anthropology at the Field Museum of Natural History, Chicago, Illinois. *Archival Science*, 13(2-3):143–193.

[Bennett et al., 2003] Bennett, R., Lavoie, B. F., and O'neill, E. T. (2003). The concept of a work in WorldCat: an application of FRBR. *Library Collections, Acquisitions, and Technical Services*, 27(1):45–59.

[Bergman, 2009] Bergman, M. (2009). Advantages and Myths of RDF. *AI3*.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked Data. http://www.w3.org/DesignIssues.

[Berners-Lee et al., 1992] Berners-Lee, T., Cailliau, R., Groff, J.-F., and Pollermann, B. (1992). World-wide web: The information universe. *Internet Research*, 2(1):52–58.

[Berners-Lee et al., 1996] Berners-Lee, T., Fielding, R., and Frystyk, H. (1996). Hypertext transfer protocol–HTTP/1.0. Technical report.

[Berners-Lee et al., 2004] Berners-Lee, T., Fielding, R., and Masinter, L. (2004). Uniform resource identifier (uri): Generic syntax. Technical report.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, 284(5):34–43.

[Bizer and Cyganiak, 2009] Bizer, C. and Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(1):1–10.

[Bizer et al., 2011] Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global.

[Borst, 1997] Borst, W. N. (1997). Construction of engineering ontologies for knowledge sharing and reuse. Centre for Telematics and Information Technology (CTIT).

[Brank et al., 2005] Brank, J., Grobelnik, M., and Mladenić, D. (2005). A survey of ontology evaluation techniques. In *Proc. of 8th Int. multi-conf. Information Society*.

[Bray et al., 1999] Bray, T., Hollander, D., and Layman, A. (1999). Namespaces in xml. http://www. w3. org/TR/1999/REC-xml-names-19990114.

[Bray et al., 1997] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (1997). Extensible markup language (XML). *World Wide Web Journal*, 2(4):27–66.

[Brewster et al., 2004] Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data driven ontology evaluation. In *International Conference on Language Resources and Evaluation (LREC 2004)*.

[Brickley, 2000] Brickley, D. (2000). Resource Description Framework (RDF) schema specification 1.0. http://www.w3.org/TR/rdf-schema.

[Brilliant, 1988] Brilliant, R. (1988). How an art historian connects art objects and information. Graduate School of Library and Information Science. University of Illinois.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

[Caraffa, 2011] Caraffa, C. (2011). *Photo archives and the photographic memory of art history*. Deutscher Kunstverlag Berlin.

[Carlisle, 2003] Carlisle, P. (2003). UK archival thesaurus (UKAT): Construction and editing methodology (Version 6.0). https://ukat.aim25.com/downloads/methodologyV6.pdf.

[Chan, 1995] Chan, L. M. (1995). *Library of congress subject headings: principles and application*. ERIC.

[Cheng et al., 2010] Cheng, J., Hu, X., and Heidorn, P. B. (2010). New measures for the evaluation of interactive information retrieval systems: Normalized task completion time and normalized user effectiveness. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, volume 47, page 72. American Society for Information Science.

[Cimiano, 2006] Cimiano, P. (2006). *Ontologies*. Springer.

[Coburn et al., 2010] Coburn, E., Light, R., McKenna, G., Stein, R., and Vitzthum, A. (2010). LIDO-lightweight information describing objects version 1.0. ICOM International Committee of Museums. http://lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf.

[Committee, 2013] Committee, B. S. (2013). Rare Book and Manuscript Section of the Association of College and Research Libraries. Descriptive cataloging of rare materials:(Graphics). http://rbms.info/dcrm/dcrmg/.

[Cooke, 1999] Cooke, A. (1999). *Authoritative Guide to Evaluating Information on the Internet. Neal-Schuman NetGuide Series.* ERIC.

[Cowles, 2014] Cowles, E. (2014). VRA Core Schemas and Documentation. http://www.loc.gov/standards/vracore/schemas.html.

[Coyle and Hillmann, 2007] Coyle, K. and Hillmann, D. (2007). Resource Description and Access (RDA): Cataloging rules for the 20th century. *D-Lib*, 13(1/2).

[Daquino, 2018a] Daquino, M. (2018a). mAuth - Corpus analysis. DOI: 10.6084/m9.figshare.7411262.v1.

[Daquino, 2018b] Daquino, M. (2018b). mAuth - Results of the User Study. figshare. DOI: 10.6084/m9.figshare.7409384.v2.

[Daquino et al., 2019] Daquino, M., Carriero, V., and Tomasi, F. (2019). Convergenze semantiche tra musei, archivi e biblioteche. Ontologie per le relazioni interpersonali. *JLIS*, 10(1).

[Daquino et al., 2016] Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., and Vitali, F. (2016). Zeri Photo Archive RDF Dataset. Centro Risorse per la Ricerca (CRR-MM), Università di Bologna.

[Daquino et al., 2017] Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., and Vitali, F. (2017). Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(4):21.

[Daquino et al., 2018] Daquino, M., Tiddi, I., Peroni, S., and Shotton, D. (2018). Creating Open Citation Data with BCite. In *CEUR Workshop Proceedings*, volume 2184.

[Daquino and Tomasi, 2015] Daquino, M. and Tomasi, F. (2015). Historical Context Ontology (HiCO): a conceptual model for describing context information of cultural heritage objects. In *Research Conference on Metadata and Semantics Research*, pages 424–436. Springer.

[De Boer et al., 2012] De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., and Schreiber, G. (2012). Supporting linked data production for cultural heritage institutes: the Amsterdam museum case study. In *Extended Semantic Web Conference*, pages 733–747. Springer.

[Deliot, 2014] Deliot, C. (2014). Publishing the British national bibliography as linked open data. *Catalogue & Index*, 174:13–18.

[Delmas-Glass, 2016] Delmas-Glass, E. (2016). The YCBA Historic Frame Collection: using Semantic Web Technology to contribute to the scholarship of British Art. *ARTis ON*, (2):9–23.

[Di Noia et al., 2016] Di Noia, T., Ragone, A., Maurino, A., Mongiello, M., Marzocca, M. P., Cultrera, G., and Bruno, M. P. (2016). Linking data in digital libraries: the case of Puglia Digital Library. In *WHiSe@ESWC*, pages 27–38.

[Dijkshoorn et al., 2018] Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., ter Weele, W., and Wielemaker, J. (2018). The Rijksmuseum collection as linked data. *Semantic Web*, 9(2):221–230.

[Doerr et al., 2008] Doerr, M., Bekiari, C., LeBoeuf, P., and nationale de France, B. (2008). FRBRoo, a conceptual model for performing arts. In *2008 Annual Conference of CIDOC, Athens*, pages 15–18.

[Doerr et al., 2003] Doerr, M., Hunter, J., and Lagoze, C. (2003). Towards a core ontology for information integration. *Journal of Digital information*, 4(1).

[Dublin Core Metadata et al., 2012] Dublin Core Metadata, I. et al. (2012). Dublin core metadata element set, version 1.1. Dublin Core Metadata Initiative. http://dublincore.org/documents/dces.

[Duranti, 1998] Duranti, L. (1998). *Diplomatics: new uses for an old science*. Scarecrow Press.

[Edwards and Hart, 2004] Edwards, E. and Hart, J. (2004). *Photographs objects histories: on the materiality of images*. Routledge.

[Edwards and Morton, 2015] Edwards, E. and Morton, C. (2015). *Photographs, museums, collections: between art and Information*. Bloomsbury Publishing.

[Eisner, 1996] Eisner, E. (1996). Evaluating the teaching of art. *Evaluating and assessing the visual arts in education: International perspectives*, pages 75–94.

[Erxleben et al., 2014] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., and Vrandečić, D. (2014). Introducing Wikidata to the linked data web. In *International Semantic Web Conference*, pages 50–65. Springer.

[Fallside and Walmsley, 2004] Fallside, D. C. and Walmsley, P. (2004). Xml schema part 0: primer second edition. W3C recommendation. https://www.w3.org/TR/xmlschema-0.

[Farahat et al., 2007] Farahat, A. O., Chen, F. R., Mathis, C. R., and Nunberg, G. D. (2007). Systems and methods for authoritativeness grading, estimation and sorting of documents in large heterogeneous document collections. Google Patents. US Patent 7,188,117.

[Fernández-López and Gómez-Pérez, 2002] Fernández-López, M. and Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, 17(2):129–156.

[Fernández-López et al., 1997] Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, pages 33–40.

[Franceschini et al., 2014] Franceschini, C. et al. (2014). Classifying content. Photographic collections and theories of thematic ordering. *Visual Resources. An international journal of documentation*, 30(2).

[Freedberg, 2006] Freedberg, D. (2006). Why connoisseurship matters. *Munuscula Amicorum: Contributions on Rubens and his colleagues in honour of Hans Vlieghe*, 1:29–43.

[Fresa, 2014] Fresa, A. (2014). PHOTOCONSORTIUM: International Consortium for Promoting and Valorising Photographic Heritage. *Uncommon Culture*, 5(9/10):73–78.

[Gangemi and Presutti, 2009] Gangemi, A. and Presutti, V. (2009). Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer.

[Gartner, 2002] Gartner, R. (2002). METS: Metadata Encoding and Transmission Standard. *JISC Techwatch report TSW*, pages 2–5.

[Gartner, 2003] Gartner, R. (2003). MODS: Metadata object description schema. *JISC Techwatch report TSW*, pages 3–6.

[Gidley, 2005] Gidley, M. (2005). Paula Richardson Fleming, Native American Photography at the Smithsonian: The Shindler Catalogue (Washington, DC and London: Smithsonian Books, 2003. *Journal of American Studies*, 39(1):120–121.

[Ginzburg, 1979] Ginzburg, C. (1979). Roots of a scientific paradigm. *Theory and Society*, 7(3):273–88.

[Gombrich, 1985] Gombrich, E. H. (1985). *Meditations on a hobby horse and other essays on the theory of art*. Phaidon Oxford.

[Gonano et al., 2014] Gonano, C. M., Mambelli, F., Peroni, S., Tomasi, F., and Vitali, F. (2014). Zeri e LODE: Extracting the Zeri photo archive to linked open data: formalizing the conceptual model. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 289–298. IEEE Press.

[Gosden et al., 2007] Gosden, C., Larson, F., and Petch, A. (2007). *Knowing things: exploring the collections at the Pitt Rivers Museum, 1884-1945*. Oxford University Press.

[Grob et al., 2011] Grob, B., Baltussen, L., Heijmans, L., Kits, R., Lemmens, P., Schreurs, E., Timmermans, N., et al. (2011). Why reinvent the wheel over and over again? how an offline platform stimulates online innovation. *Archives & Museum Informatics*.

[Gruber, 1993] Gruber, T. (1993). What is an Ontology. http://www-ksl.stanford.edu/kst/whatis-an-ontology. html.

[Guarino et al., 2009] Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.

[Guha et al., 2016] Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51.

[Harpring, 1997] Harpring, P. (1997). The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. In *Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums, Archives and Museum Informatics*.

[Harpring, 2010] Harpring, P. (2010). Development of the Getty vocabularies: AAT, TGN, ULAN, and CONA. *Art Documentation: Journal of the Art Libraries Society of North America*, 29(1):67–72.

[Harpring, 2013] Harpring, P. (2013). Getty Vocabularies and linked data. http://files.archivists.org/conference/sandiego2012/401-Harpring.pdf.

[Harpring et al., 2006] Harpring, P., Lanzi, E., Whiteside, A., McRae, L., et al. (2006). *Cataloging cultural objects: A guide to describing cultural works and their images*. American Library Association.

[Haslhofer and Isaac, 2011] Haslhofer, B. and Isaac, A. (2011). data.europeana.eu - The Europeana Linked Open Data Pilot. In *DCMI International Conference on Dublin Core and Metadata Applications*. http://eprints.cs.univie.ac.at/2919/.

[Hausenblas and Cygankiak, 2016] Hausenblas, M. and Cygankiak, R. (2016). Linked Data Life Cycles. http://www.slideshare.net/mediasemanticweb/linked-data-life-cycles.

[Hayes, 2004] Hayes, P. (2004). RDF semantics, W3C recommendation. http://www.w3.org/TR/rdf-mt/.

[Heritage, 2012] Heritage, E. (2012). *MIDAS Heritage: The UK Historic Environment Data Standard*. English Heritage.

[Hevner et al., 2004] Hevner, A., March, S., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Q*, 28(1):75–105.

[Hillmann et al., 2010] Hillmann, D., Coyle, K., Phipps, J., and Dunsire, G. (2010). Rda vocabularies: process, outcome, use. *D-Lib magazine*, 16(1/2):6.

[Horrocks et al., 2004] Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al. (2004). SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission*, 21:79.

[Hyland and Wood, 2011] Hyland, B. and Wood, D. (2011). The joy of data - a cookbook for publishing linked government data on the web. In *Linking government data*, pages 3–26. Springer.

[ICA, 1999] ICA (1999). General International Standard Archival Description ISAD(G). International Council on Archives Committee on Descriptive Standards.

[IFLA, 1974] IFLA (1974). International Standard Bibliographic Description for Monographic Publications (ISBD).

[Jussim, 1989] Jussim, E. (1989). *The eternal moment: Essays on the photographic image*. Aperture.

[Kelly et al., 2009] Kelly, D. et al. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224.

[Kleinberg, 1998] Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*. Citeseer.

[Klic et al., 2018] Klic, L., Nelson, J. K., Pattuelli, M. C., and Provo, A. (2018). Florentine Renaissance Drawings: A Linked Catalog for the Semantic Web. *Art Documentation: Journal of the Art Libraries Society of North America*, 37(1):33–43.

[Klijn and de Lusenet, 2003] Klijn, E. and de Lusenet, Y. (2003). Sepiades.

[Knight and Burn, 2005] Knight, S. and Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science*, 8.

[Knoblock et al., 2017] Knoblock, C. A., Szekely, P., Fink, E., Degler, D., Newbury, D., Sanderson, R., Blanch, K., Snyder, S., Chheda, N., Jain, N., et al. (2017). Lessons learned in building linked data for the American art collaborative. In *International Semantic Web Conference*, pages 263–279. Springer.

[Kroeger, 2013] Kroeger, A. (2013). The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC Future. *Cataloging & classification quarterly*, 51(8):873–890.

[Larson and Janakiraman, 2011] Larson, R. R. and Janakiraman, K. (2011). Connecting archival collections: the social networks and archival context project. In *International Conference on Theory and Practice of Digital Libraries*, pages 3–14. Springer.

[Le Boeuf et al., 2015] Le Boeuf, P., Doerr, M., Ore, C. E., and Stead, S. (2015). Definition of the CIDOC conceptual reference model. *ICOM/CIDOC Documentation Standards Group and CIDOC CRM Special Interest Group, version 6.1*.

[Leary, 1985] Leary, W. H. (1985). *The Archival Appraisal of Photographs: A RAMP Study with Guidelines*. ERIC.

[Lebo et al., 2013] Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). Prov-o: The prov ontology. *W3C recommendation*.

[Lee et al., 2002] Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & management*, 40(2):133–146.

[Lei et al., 2007] Lei, Y., Nikolov, A., Uren, V., and Motta, E. (2007). Detecting Quality Problems in Semantic Metadata without the Presence of a Gold Standard. In *EON*, pages 51–60.

[Lien and Edwards, 2014] Lien, S. and Edwards, E. (2014). *Uncertain images: Museums and the work of photographs*. Ashgate Publishing, Ltd.

[Lozano-Tello and Gómez-Pérez, 2004] Lozano-Tello, A. and Gómez-Pérez, A. (2004). Ontometric: A method to choose the appropriate ontology. *Journal of Database Management (JDM)*, 15(2):1–18.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 251–263. Springer.

[Maginnis, 1990] Maginnis, H. B. (1990). The role of perceptual learning in connoisseurship: Morelli, Berenson, and beyond. *Art history*, 13(1):104–117.

[Mambelli, 2014] Mambelli, F. (2014). Una risorsa online per la storia dell'arte: il database della fototeca Zeri. *Digital Humanities: progetti italiani ed esperienze di convergenza multidisciplinare. Quaderni Digilab, Università di Roma La Sapienza*.

[Mambelli, 2018] Mambelli, F. (2018). Bene, documento, fonte. la fotografia negli archivi fotografici degli storici dell'arte. *INTRECCI d'arte*, 7(4):91–101.

[Marden et al., 2013] Marden, J., Li-Madeo, C., Whysel, N., and Edelstein, J. (2013). Linked open data for cultural heritage: evolution of an information technology. In *Proceedings of the 31st ACM international conference on Design of communication*, pages 107–112. ACM.

[Marien, 2006] Marien, M. W. (2006). *Photography: A cultural history*. Laurence King Publishing.

[Mazzini and Ricci, 2011] Mazzini, S. and Ricci, F. (2011). EAC-CPF Ontology and Linked Archival Data. In *SDA*, pages 72–81.

[McKenna and Patsatzi, 2007] McKenna, G. and Patsatzi, E. (2007). *SPECTRUM: The UK museum documentation standard*. Museum Documentation Association.

[Mitchell et al., 2011] Mitchell, G. R., Church, S., Bartosh, T., Godana, G. D., Stohr, R., Jones, S., and Knowlton, A. (2011). Measuring scholarly metrics. Oldfather Press. https://digitalcommons.unl.edu/commstudiespapers/25/.

[Moreau et al., 2015] Moreau, L., Groth, P., Cheney, J., Lebo, T., and Miles, S. (2015). The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:235–257.

[Morelli and Richter, 1883] Morelli, G. and Richter, L. M. S. (1883). *Italian masters in German galleries*. G. Bell and Sons.

[Moriarty, 2000] Moriarty, C. (2000). A Backroom Service: The Photographic Library of the Council of Industrial Design, 1945-1965. *Journal of design history*, 13(1):39–57.

[Naumann and Rolker, 2005] Naumann, F. and Rolker, C. (2005). Assessment methods for information quality criteria. In *In Proceedings of the International Conference on Information Quality (IQ*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik.

[Noy et al., 2001] Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05.

[Pałubicki et al., 1978] Pałubicki, J., Couprie, L., and van de Waal, H. (1978). Iconoclass an iconographic classification system. *Studia Źródłoznawcze/Commentationes*, 23.

[Park, 2009] Park, J.-R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & classification quarterly*, 47(3-4):213–228.

[Parker, 1982] Parker, E. B. (1982). Graphic materials: rules for describing original items and historical collections. Library of Congress.

[Pattuelli et al., 2013] Pattuelli, M. C., Miller, M., Lange, L., Fitzell, S., and Li-Madeo, C. (2013). Crafting linked open data for cultural heritage: Mapping and curation tools for the linked jazz project. *Code4Lib Journal*, 21(4).

[Pemberton et al., 2000] Pemberton, S. et al. (2000). XHTML 1.0 the extensible hypertext markup language. W3C Recommendations. https://www.w3.org/TR/xhtml1.

[Peroni, 2016] Peroni, S. (2016). A simplified agile methodology for ontology development. In *OWL: Experiences and Directions–Reasoner Evaluation*, pages 55–69. Springer.

[Peroni and Shotton, 2012] Peroni, S. and Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:33–43.

[Peroni and Shotton, 2018a] Peroni, S. and Shotton, D. (2018a). The OpenCitations Data Model. figshare. DOI:10.6084/m9.figshare.3443876.v5.

[Peroni and Shotton, 2018b] Peroni, S. and Shotton, D. (2018b). The SPAR ontologies. In *International Semantic Web Conference*, pages 119–136. Springer.

[Peroni et al., 2012] Peroni, S., Shotton, D., and Vitali, F. (2012). Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 9–16. ACM.

[Petersen, 1990] Petersen, T. (1990). Developing a new thesaurus for art and architecture. Graduate School of Library and Information Science. University of Illinois.

[Pitti, 1999] Pitti, D. V. (1999). Encoded archival description: An introduction and overview. *D-Lib Magazine*, 5(11).

[Pitti, 2004] Pitti, D. V. (2004). Creator description: encoded archival context. *Cataloging & classification quarterly*, 38(3-4):201–226.

[Porzel and Malaka, 2004] Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*, pages 1–6. Citeseer.

[Poveda Villalón et al., 2012] Poveda Villalón, M., Suárez-Figueroa, M. C., and Gómez-Pérez, A. (2012). The landscape of ontology reuse in linked data. In *Ontology Engineering in a Data-driven World (OEDW 2012)*.

[Prud et al., 2006] Prud, E., Seaborne, A., et al. (2006). SPARQL query language for RDF. W3C Recommendation.

[Reichheld and Markey, 2011] Reichheld, F. F. and Markey, R. (2011). *The ultimate question 2.0: How net promoter companies thrive in a customer-driven world*. Harvard Business Press.

[Reist et al., 2015] Reist, I., Farneth, D., Stein, R. S., and Weda, R. (2015). An introduction to PHAROS: aggregating free access to 31 million digitized images and counting. Speech at CIDOC 2015. http://network.icom.museum/fileadmin/user$_u pload/minisites/cidoc/BoardMeetings/CIDOC_PHAROS_F$
$Stein - Weda_1.pdf$.

[Ricci, 2014] Ricci, F. (2014). Il progetto italiano reload al lodlam summit 2013 linked open data in libraries archives and museums. *DigItalia*, 2:173–181.

[Richardson, 1719] Richardson, J. (1719). *Two Discourses: I. An Essay on the Whole Art of Criticism, as it Relates to Painting... II. An Argument in Behalf of the Science of a Connoisseur; Wherein is Shewn the Dignity, Certainty, Pleasure, and Advantage of it*.

[Rieh, 2002] Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American society for information science and technology*, 53(2):145–161.

[Ronzino et al., 2011] Ronzino, P., Amico, N., and Niccolucci, F. (2011). Assessment and comparison of metadata schemas for architectural heritage. In *International CIPA Symposium*.

[Ruddock, 2011] Ruddock, B. (2011). Linked Data and the LOCAH project. *Business Information Review*, 28(2):105–111.

[Sanders, 2017] Sanders, S. (2017). Linked Library Data: It's Happening. http://www.exlibrisgroup.com/linked-library-data-its-happening/.

[Saur, 1998] Saur, K. (1998). IFLA Study Group on the functional requirements for bibliographic records. Functional requirements for bibliographic records: final report. UBCIM Publications-New Series.

[Scannapieco, 2006] Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications*. Springer.

[Schlak, 2008] Schlak, T. (2008). Framing photographs, denying archives: the difficulty of focusing on archival photographs. *Archival Science*, 8(2):85–101.

[Schreur, 2018] Schreur, P. E. (2018). Linked data for production (ld4p): a multi-institutional approach to technical services transformation. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 429–430. International World Wide Web Conferences Steering Committee.

[Schultz, 2015] Schultz, D. (2015). Photo Archives in the History of Art History: Investigating the Collection in the History of Art Department. History of Art at Oxford University. https://oxfordarthist.wordpress.com/2015/12/09/photo-archives-in-the-history-of-art-history-investigating-the-collection-in-the-history-of-art-department/.

[Schwartz, 1988] Schwartz, G. (1988). Connoisseurship: The penalty of ahistoricism. *Museum Management and Curatorship*, 7(3):261–268.

[Schwartz, 1995] Schwartz, J. M. (1995). "We make our tools and our tools make us": Lessons from Photographs for the Practice, Politics, and Poetics of Diplomatics. *Archivaria*, 40.

[Schwartz, 2002] Schwartz, J. M. (2002). Coming to terms with photographs: Descriptive standards, linguistic "othering," and the margins of archivy. *Archivaria*, 54.

[Schwartz and Cook, 2002] Schwartz, J. M. and Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival science*, 2(1-2):1–19.

[Sennett, 1986] Sennett, R. S. (1986). Earl Photographic Catalogues: An Untapped Resource. *Visual Resources*, 3(2):75–96.

[Simon et al., 2013] Simon, A., Wenz, R., Michel, V., and Di Mascio, A. (2013). Publishing bibliographic records on the web of data: opportunities for the BnF (French National Library). In *Extended Semantic Web Conference*, pages 563–577. Springer.

[Studer et al., 1998] Studer, R., Benjamins, V. R., Fensel, D., et al. (1998). Knowledge engineering: principles and methods. *Data and knowledge engineering*, 25(1):161–198.

[Su, 1992] Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management*, 28(4):503–516.

[Summers et al., 2008] Summers, E., Isaac, A., Redding, C., and Krech, D. (2008). LCSH, SKOS and linked data. In *Proc. Int?l Conf. on Dublin Core and Metadata Applications*. Humboldt-Universität zu Berlin.

[Thibodeau, 1995] Thibodeau, S. (1995). Archival context as archival authority record: the ISAAR (CPF). *Archivaria*, 40.

[Tillett, 2005] Tillett, B. (2005). What is FRBR? a conceptual model for the bibliographic universe. *The Australian Library Journal*, 54(1):24–30.

[Tunesi, 2014] Tunesi, A. (2014). Stefano bardini's photographic archive: a visual historical document. University of Leeds.

[Ugander et al., 2011] Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the Facebook social graph. *arXiv*. arXiv:1111.4503.

[Uschold and King, 1995] Uschold, M. and King, M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing*. Citeseer.

[Van Hooland and Verborgh, 2014] Van Hooland, S. and Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. Facet publishing.

[Van Steen, 2014] Van Steen, N. (2014). Metadata management in europeana photography. *SCIRES-IT-SCIentific RESearch and Information Technology*, 4(2):127–132.

[Villazón-Terrazas et al., 2011] Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., and Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In *Linking government data*, pages 27–49. Springer.

[Wache et al., 2001] Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information-a survey of existing approaches. In *IJCAI-01 workshop: ontologies and information sharing*, volume 2001, pages 108–117. Citeseer.

[Walton, 2013] Walton, D. (2013). *Argumentation schemes for presumptive reasoning*. Routledge.

[Wick and Vatant, 2012] Wick, M. and Vatant, B. (2012). The geonames geographical database. http://geonames. org.

[Wilson, 1983] Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Greenwood Press.

[Zaveri et al., 2016] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.

[Zeri, 1995] Zeri, F. (1995). *Confesso che ho sbagliato: ricordi autobiografici*. Longanesi.