JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

JLIS.it

Knowledge Representation of digital Hermeneutics of archival and literary Sources

Marilena Daquino[(a)], **Valentina Pasqual**[(b)], **Francesca Tomasi**[(c)]

a) Università di Bologna, Digital Humanities Advanced Research Centre (/DH.arc), http://orcid.org/0000-0002-1113-7550
b) Università di Bologna, Digital Humanities Advanced Research Centre (/DH.arc), http://orcid.org/0000-0001-5931-5187
c) Università di Bologna, Digital Humanities Advanced Research Centre (/DH.arc), http://orcid.org/0000-0002-6631-8607

Contact: Marilena Daquino, marilena.daquino2@unibo.it; Valentina Pasqual, valentina.pasqual2@unibo.it; Francesca Tomasi, francesca.tomasi@unibo.it

Received: 13 May 2020; **Accepted:** 21 July 2020; **First Published:** 15 September 2020

ABSTRACT

Scholarly analysis of archival, library, and literary sources results in a variety of digital artefacts meant to foster knowledge discovery and new research enquiries. Guidelines and standards to formally represent disciplinary information are available (e.g. XML schemas, ontologies, vocabularies). However, digital artefacts rarely address reusable structured information on the hermeneutical approach adopted by scholars when validating hypotheses. As a consequence, reproducibility and assessment of research results is hampered, and comparing online contradictory information is still a hard task. In this work we show how to leverage Semantic Web technologies in a high-level, portable data model for representing hermeneutical aspects related to cross-disciplinary analysis of archival and literary sources. We showcase three representative scenarios in the Cultural Heritage domain where the model is applied, and we describe benefits and limits of our solution.

KEYWORDS

Digital hermeneutics; Semantic Web; Data model; Archival linked data; Library linked data.

CITATION

Daquino, M., Pasqual, V., Tomasi, F. "Knowledge Representation of digital Hermeneutics of archival and literary Sources." *JLIS.it* 11, 3 (September 2020): 59–76. DOI: 10.4403/jlis.it-12642.

JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

1. Introduction[1]

Hermeneutics has a key role in the development of Humanities. Scholars close read information sources with the purpose of inferring new knowledge, e.g. by using logical approaches such as deduction or induction. Such an epistemological process is accompanied by the validation of hypotheses, which can be assessed on the basis of clues witnessed in primary and secondary sources, scholars' background knowledge, assumptions, facts, and other scholars' statements, e.g. artwork attributions, philologists' readings of a text, descriptions of historical events. Despite several disciplinary methodologies have been proposed over time for validating hypotheses (Pasquali 1952; Maas 1972; Morelli 1883; Ginzburg 1979), these are not always reproducible, especially when based on qualitative assessment methods and humanists' background latent knowledge. As a consequence, the reliability of a statement is often inferred from the authoritativeness of scholars and cultural institutions (i.e. *first-hand* and *second-hand knowledge providers* respectively) responsible for the information (Wilson 1983; Rieh 2002).

Likewise, digital artefacts resulting from Humanists' research activities often lack structured information representing hermeneutical processes. While web applications are powerful presentation tools that can facilitate a deep understanding of research results, data on argumentations around questionable statements (i.e. hypotheses) is usually shallow (e.g. motivations, methods, sources), and relations between contradictory statements are not explicitly described.

For instance, consider the following artwork attribution stated in a cataloguing record of the notable Federico Zeri Photo archive[2] (translated in English for convenience):

> Artwork: Tre Grazie.
>
> Author: Peruzzi Baldassarre.
>
> Reason for attribution: Bibliography
>
> Bibliography: Frizzoni G., Delle pitture di Baldassarre Peruzzi e del giudizio portatone dal sig. Cavalcaselle, in *Il Buonarroti*, 1869, 35; Berenson B., *Italian Pictures of the Renaissance*, 1932, 441; […] Morelli G., *Della Pittura italiana: studii storico-critici; le Gallerie Borghese e Doria-Pamphili in Roma*, 1991, 144
>
> Other attribution: Luini Bernardino, scuola.
>
> Reason for attribution: Christie's auction, 1994.

The attribution is accompanied by relevant background information, such as: motivations supporting the attribution ("Bibliography"), sources (the list of bibliographic references), the date of the attribution (in this case the publication date of the latest bibliographic reference, i.e. 1991), knowledge providers (the list of authors, the Federico Zeri Photo archive), and relations with contradictory attributions ("Luini Bernardino, scuola"). At a first sight, archivists seem to prefer an attribution claimed in scholarly, peer-reviewed, bibliographic works over an attribution claimed by an auction firm, which might be biased by economical interests and therefore being less reliable. Such a

[1] M. Daquino is responsible for sections 2, 3, 5.3; V. Pasqual is responsible for sections 4 and 5.1; F. Tomasi is responsible for sections 1, 5.2; all authors collaborated in writing section 6.

[2] Peruzzi Baldassarre, "Tre Grazie", Catalogo della Fondazione Federico Zeri. http://catalogo.fondazionezeri.unibo.it/entry/work/39794/Peruzzi%20Baldassarre%2C%20Tre%20Grazie. Accessed May 10, 2020.

preference (say, ranking) of sources and methodologies is not explicit in data, despite being the result of established cataloguing methodologies.

Nonetheless, the formal representation of argumentations around questionable statements has several benefits. Machines would be able to reason on contradictory information for comparative purposes, and automatic and semi-automatic methods can be built to validate statements on the basis of qualities such as credibility, authority, and relevance. Such methods could be leveraged to recommend information to users and effectively support their decision-making process. Moreover, such methods can strengthen authoritativeness of cultural institutions among patrons, and can support the former in expensive and time-consuming curatorial activities, such as updating cataloguing data including not up-to-date, and potentially incorrect information with other data providers' information if deemed reliable. Lastly, scholars could benefit from mechanisms to reference questionable statements along with all the necessary information needed to validate the statement (e.g. URIs that identify statements and related hermeneutical aspects, rather than URIs identifying sources).

In this article we propose a high-level, portable data model for representing hermeneutical aspects as structured data. We leverage Semantic Web technologies, which are widely recognised in the Cultural Heritage domain as powerful means for representing complex information in a formal and expressive fashion. The aim is to facilitate the development of hermeneutical-aware technologies that can support scholars' decision-making processes when validating contradictory information in the Humanities.

The rationale of the article is as follows. In section "Related work" we introduce the background of this work, we provide our definition of digital hermeneutics and acknowledge prior works. In section "Formal representation of digital hermeneutics" we introduce the scope and the strategies we adopted to formally represent hermeneutical aspects. In section "The data model" we present the layered structure of the data model and we introduce its features. In section "Scenarios" we exemplify the usage of the data model in the three representative use cases. Finally, in section "Discussion and conclusion" we discuss benefits and limits of our approach and we address future works.

2. Related work

Hermeneutical practices are described as "Digital hermeneutics" when digital sources and computational methods are involved in the epistemological process. To the best of our knowledge, two definitions of digital hermeneutics exist in literature. On the one hand, digital hermeneutics refers to the critical usage of digital sources in research enquiries. The main concern regards the development of scholars' awareness when leveraging data and web applications (e.g. social network platforms) in their research (Mallery, Hurwitz, and Duffy 1986; Capurro 2000). On the other hand, digital hermeneutics refers to the development of computational methods able to to generate or validate hypotheses given a set of rules, based on quantitative or qualitative methods (Lehnert, Alker, and Schneider 1983; Ramsay 2010; Van Zundert 2016; Maiatsky *et al.* 2018; Romele, Severo, and Furia 2018).

In this work we tackle knowledge representation problems that are relevant to the second definition. Precisely, we are interested in the representation of information compelling to the design of hermeneutics-aware technologies. Research in this area focuses on two main aspects, namely: (1) the formal representation of hermeneutical aspects and (2) the definition of formal assessment methods for validating hypotheses.

The formal representation of hermeneutical aspects addresses the ontological description of aspects that characterise argumentations around questionable information, such as methods, sources, and agents (both human and software) involved in the hermeneutical process, and relations between argumentations (e.g. disagreement, influence). A plethora of technological solutions exist to represent discipline requirements. For instance, TEI critical apparatus Guidelines[3] allow philologists to record different readings on the same text. However, there is no guidance on how to motivate an attribution, how to encode it consistently and shearably, and how to reference contradictory statements (e.g. including URLs, local identifiers, or none), therefore resulting in incomplete, and heterogeneous data. Likewise, Semantic Web ontologies allow us to annotate entities with a variety of context information. However, to date no comprehensive solution exists. For instance, the Open Annotation Data model (Sanderson *et al*. 2013) addresses annotations, responsible agents, and motivations, but does not address relations between annotations. Similarly, CIDOC-CRM (Crofts *et al*. 2011) can be used to describe attribution activities, agents, and motivations, but relations between agents (e.g. influence), sources (e.g. citation), and statements (e.g. disagreement) cannot be described. Other light-weight models focus on narrower representational tasks, e.g. certainty of statements (De Waard and Schneider 2012). The HiCO ontology (Daquino and Tomasi 2015) is an extension of the PROV Ontology (Moreau *et al.* 2015) that allows to record both aspects underlying hermeneutical activities and relations between contradictory information. However, due to the complexity of questionable statements to be annotated, representing hermeneutics in a triple fashion is not always sufficient. In the biomedical domain Named Graphs (Carroll *et al*. 2005) have been applied to formally represent scholars' assertions as layered information extracted from academic publications. In this work we reuse and integrate several existing technologies so as to achieve a comprehensive formal representation of digital hermeneutics.

Secondly, the definition of formal assessment methods for hypotheses validation concerns the design of frameworks of rules based on information quality aspects (e.g. completeness, timeliness, third-party opinions) that enable machines to deduce aspects such as reliability, certainty, and authoritativeness of hypotheses. So far, researchers in Computer science and in Library and Information Science have focused on knowledge representation of argumentations (Bizer and Oldakowski 2004; 2004; Gil and Artz 2007; Schneider, Groza, and Passant 2013) to be leveraged by data quality assessment methods (Naumann and Rolker 2005; Batini *et al*. 2009; Zaveri *et al*. 2016). However, such technologies have never been applied to cataloguing, bibliographic, and Humanities data in order to validate authoritativeness of content information. In this work we show how the formal representation of hermeneutics according to a portable data model can effectively support recommendation tasks in the field of art history.

3. Formal representation of digital hermeneutics

The objective of this work is to define methods and models to formally represent hermeneutical aspects by means of Semantic Web technologies. The ontological representation of hermeneutics

[3] TEI Critical Apparatus Guidelines. https://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html. Accessed May 10, 2020.

focuses on questionable statements that are stated and recorded by somebody in a source (e.g. a cataloguing record). Representational requirements can be summarised as follows:

- **Type of statement**. A classification of the questionable statement, e.g. artwork attribution, philological reading.
- **Sources**. Sources include the document where the statement is recorded, and primary and secondary sources that are used (e.g. cited, analysed, subject of) to support the statement. Sources may be digital artefacts, such as an online cataloguing record, or analog documents, like a book, the back of a photograph. The description addresses factual data useful to identify sources (e.g. identifiers, authors, dates, curating institutions).
- **Agents**. Agents include first and second knowledge providers (e.g. scholars, cultural institutions), and software agents involved in the life-cycle of the questionable statement (e.g. definition, text mining, reconciliation, and linking techniques). The description addresses information necessary to uniquely identify agents and their role in the digital hermeneutical process.
- **Motivations**. A classification of motivations used to justify the endorsement of a hypothesis. The description of motivations addresses sources and agents' statements that may support hypotheses (e.g. preference for peer-reviewed articles, auction catalogues, scholars' verbal communications).
- **Certainty**. Differently from vagueness (e.g. "The author worked in the 90's"), certainty is the quality of a statement that characterises its degree of precision (e.g. "The author wrote her main work in 1994(?)").
- **Relations**. Relations include those between sources (e.g. an article cited in a cataloguing record), between sources and agents (e.g. authors), between statements and sources (e.g. a source provides evidence for the statement), statements and agents (e.g. people that agree on the statement), and between statements (e.g. attributions in agreement). Relations may also be the subject of assertions (e.g. the relation between an artist and the artwork she created).

Outlined concepts can be leveraged by quantitative and qualitative assessment methods to validate the following qualities:

- Statements classification allows to select adequate assessment methods according to the discipline, beliefs, and common practises (e.g. art historians may privilege recent sources, philologists may prefer peer-reviewed sources).
- Bibliographic data allows reasoning on **recentness** of cited documents.
- Motivations allow to rank **reliability** of decisions made by knowledge providers.
- Certainty allows to characterise **credibility** of statements.
- Relations between sources, statements, and agents support the assessment of agents' **authoritativeness** (e.g. by means of citation indexes, or white lists including third-party opinions on scholars and cultural institutions) and software agents' **reliability** (or bias).
- All together such aspects allow to measure **completeness** of context information.

Selected representational requirements and assessment methods were designed on the basis of interviews with domain experts, empirical analysis, and data validation performed on digital scholarly editions (Daquino and Tomasi 2015; Daquino, Giovannetti, and Tomasi 2019) and art historical photo archives catalogues (Daquino 2020; 2019a) so as to characterise hermeneutical approaches in different disciplines.

Layered knowledge representation. Computational methods act as intermediary tools between sources and readers and can support research methodologies based on qualitative and quantitative methods, e.g. distant reading (Moretti 2013). The mediation of technology in the hermeneutical circle (Stegmüller 1977) results in a representation of the world (i.e. a statement) that must be screened, interpreted, and explained by humans (i.e. an assessment). For instance, quantitative analysis on literary texts may highlight phenomena that cannot be understood without the intervention of scholars that leverage their background knowledge (e.g. facts). Moreover, technology has its own hermeneutical blueprint. Programming allows us to select, manipulate, or discard factual data (i.e. a mining process), and provides a biased perspective of the world.

Such aspects can be represented as separate layers of information that contextualise a questionable statement. Specifically, questionable statements can be represented in a source-driven fashion as facts that are valid in the context of the source where those are recorded. Context information is grouped in separate layers so as to isolate data according to its nature and to define common patterns between layers. Such layers can be summarised as follows:

- Layer 0. Factual data that is part of scholars' background knowledge
- Layer 1. The scope of scholars' questionable statements
- Layer 2. Context information for hypotheses assessment
- Layer 3. Provenance information of the mining processes

In order to provide an efficient data-driven representation of information, avoiding redundancy and overengineering of ontologies, we propose the usage of named graphs to formally represent layers, and the usage of the Nanopublication data model (Groth, Gibson, and Velterop 2010) - prefix **:np** - for relating graphs meaningfully. Figure 1 shows an overview of the Nanopublication data model applied to layers 1-3. Factual data can be linked to questionable statements by means of several properties. Among the others we use the HiCO property **hico:isExtractedFrom** to represent the information source where questionable statements are extracted from.

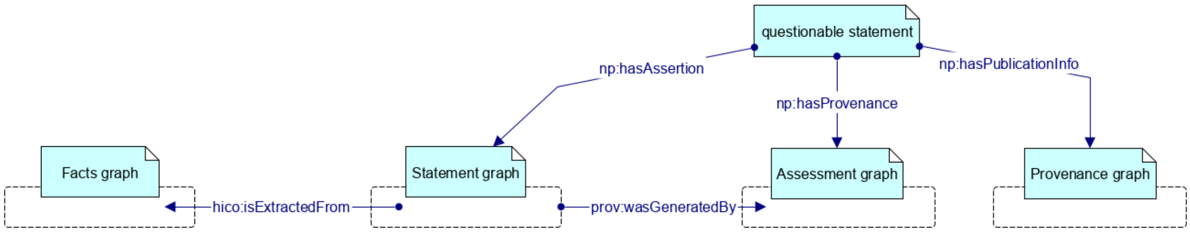


Figure 1. Overview of Nanopublication data model applied to the four layers

Secondly, we reviewed and selected well-known ontologies in the Cultural Heritage domain (Daquino 2019b) to represent the contents of graphs. To foster semantic interoperability, only existing ontologies were reused (with the exception of one property), which can be imported in new ontologies or directly reused as-is according to specific project requirements. However, none of surveyed ontologies covers all the aspects at hand. We therefore designed a modular, orthogonal data model wherein specific ontologies apply to each layer, with little overlap, so as to avoid harmonisation tasks and possible inconsistencies. The selected ontologies (see all prefixes in Fig. 2) are the following:

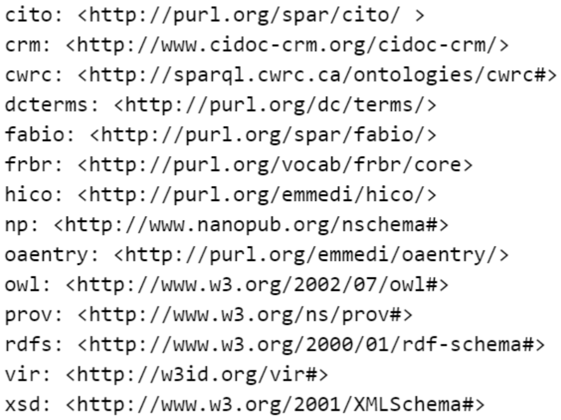
- Layer 0. Factual bibliographic and cataloguing data has a pragmatic nature, that is, to allow quick identification of objects involved in the hermeneutical discourse. We selected the SPAR ontologies (Peroni and Shotton 2018), a well-known set of ontologies for representing the publishing domain, to describe bibliographic resources and serial objects (e.g. photographs), and CIDOC-CRM (Crofts *et al.* 2011) to describe cultural objects that are *unica*.

- Layer 1. The scope of scholars' questionable statements is highly domain-dependent, and there are no one-size-fits-all solutions. We foster the reuse of CIDOC-CRM as much as possible, supported by ancillary ontologies (mostly extensions of the former) to cover specific issues, such as VIR (Carboni and de Luca 2019) for representing iconographic aspects, OAEntry (Daquino *et al. 2017)* for representing relations between artefacts, and CRMtex (Felicetti and Murano 2017) for representing textual features.
- Layer 2. Context information for hypothesis assessment includes features that characterise the validity of the statement, such as classifications, motivations, responsible agents, and sources. HiCO (Daquino and Tomasi 2015), an extension of the PROV Ontology covers all of such aspects, along with terms from CWRC[4] to qualify the degree of certainty.
- Layer 3. Provenance information of the mining process addresses procedures and responsible agents involved in the automatic or semi-automatic generation of information that populate prior layers (e.g. knowledge extraction, data reengineering, reconciliation). We selected the PROV (Moreau *et al.* 2015) ontology to describe such aspects.

4. The data model

The aim of the data model is to guide ontology designers through the description of hermeneutical aspects by providing them real-world, documented, and tested examples. We outline here the scope of each layer, including competency questions, a diagram of classes and properties, and a brief explanation. Consider the following natural language scenario as a guiding example through the description: *A scholar compares the hue of inks used in a manuscript. She notices that a different ink in a few paragraphs is used to highlight quotations. However, since the manuscript is damaged because of humidity, she is not confident with the statement. She cites a scholar's work in agreement to support her statement. She records her statements in a blog post.*

[4] Canadian Writing Research Collaboratory, http://sparql.cwrc.ca/. Accessed May 10, 2020.



```
cito: <http://purl.org/spar/cito/ >
crm: <http://www.cidoc-crm.org/cidoc-crm/>
cwrc: <http://sparql.cwrc.ca/ontologies/cwrc#>
dcterms: <http://purl.org/dc/terms/>
fabio: <http://purl.org/spar/fabio/>
frbr: <http://purl.org/vocab/frbr/core>
hico: <http://purl.org/emmedi/hico/>
np: <http://www.nanopub.org/nschema#>
oaentry: <http://purl.org/emmedi/oaentry/>
owl: <http://www.w3.org/2002/07/owl#>
prov: <http://www.w3.org/ns/prov#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
vir: <http://w3id.org/vir#>
xsd: <http://www.w3.org/2001/XMLSchema#>
```

Figure 2. Ontology prefixes used

Layer 0. Information in this layer answers the following questions: What are the artefacts that are part of the hermeneutic discourse? What is the logical organisation of the components of the artefacts? The layer includes bibliographic metadata (e.g. the edition of a work), the physical and logical description of artefacts (e.g. the folios of a manuscript), and explicit, factual, relations between artefacts, such as citations or quotations. Statements of layer 0 can be organised in several named graphs, which can be in turn subject to assertions included in layer 1.
Layer 0 includes the bibliographic information of the manuscript and the cited work, and a representation of the structure of the manuscript so as to uniquely identify the paragraph subject of the statement. Figure 3 illustrates such aspects.

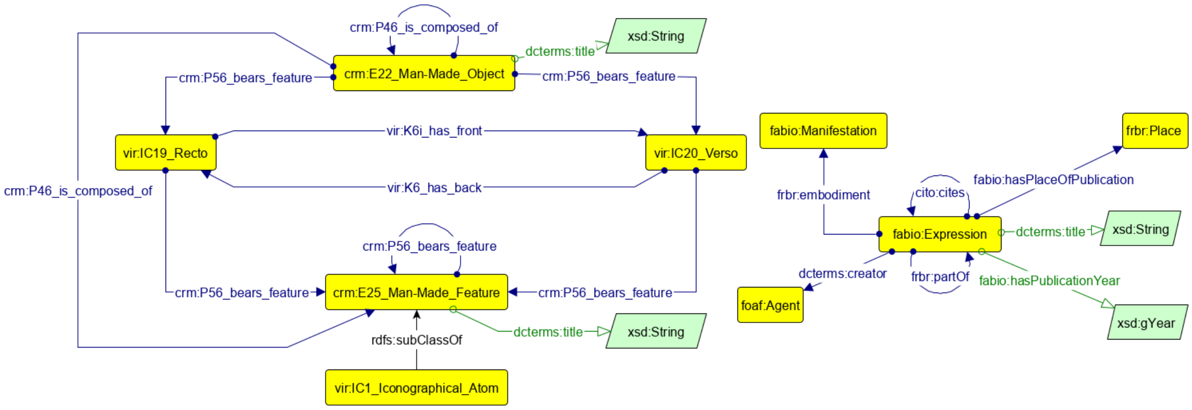


Figure 3. Data model for describing factual data

In detail, a cultural object (e.g. the manuscript) is an individual of the class **crm:E22_Man-Made_Object**, composed of textual and visual contents (e.g paragraphs, maps, illustrations) which are individuals of **crm:E25_Man-Made_Feature** or **vir:IC1_Iconographical_Atom**, as well as to physical features (e.g. recto and verso, instances of **vir:IC19_Recto** and **vir:IC20_Verso**). A document (e.g. the cited work) is represented as an individual of the class **fabio:Expression** when referring to its contents (e.g. the text of the quotation) and as an individual of the class **fabio:Manifestation** when

JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

referring to a specific tangible version or edition (e.g. a text included in a specific edition of the document). Bibliographic metadata include – among the others – title (**dcterms:title**), authors (**dcterms:creator**), place of publication (**fabio:hasPlaceOfPublication**), and date (**fabio:hasPublicationYear**). Relations between artefacts are represented by using subproperties of cito:cites to specify the type of reference.

Layer 1. Information in this layer answers the following question: what an agent argues about? It may include any kind of questionable statement, and competing statements conveying different information on the same topic or artefact can coexist in separated graphs. Following the prior example, Layer 1 includes a representation of the scholar's statement about the usage of a different ink. Since such a layer is highly domain-dependent we do not provide a data model, while we exemplify three applications in section Scenarios.

Layer 2. Information in this layer answers the following questions: What type of claim is it? Who claims that? When was it claimed? What is the primary source of the statement? What is the degree of certainty? It includes provenance and context information of a statement belonging to layer 1, such as dates, sources, and criteria supporting them, and the degree of certainty. Following the prior example, this layer includes links between entities part of Layer 0 and Layer 1, and context information, such as information on the scholar, dates, sources, and certainty of the statement. Figure 4 illustrates aspects belonging to layer 2.

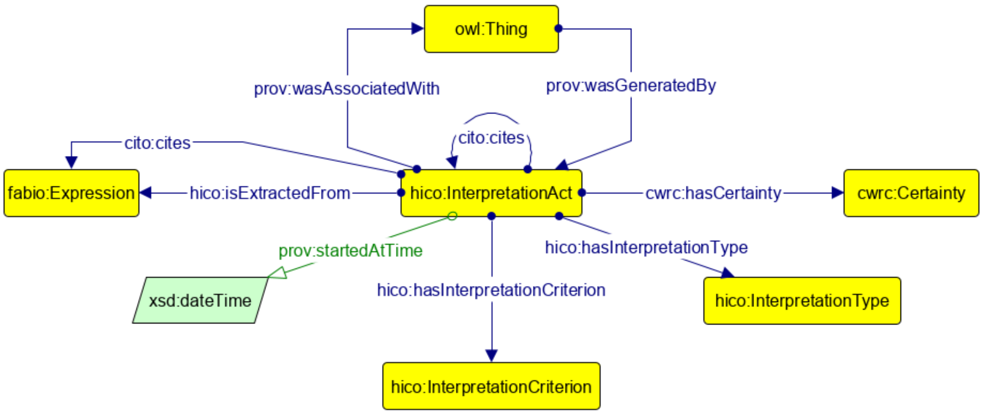


Figure 4. Data model for describing context information for hypotheses validation

An individual of the class **hico:InterpretationAct** is used to characterise the assertion – e.g. *the usage of a different ink to highlight a quotation*, here represented as an individual of **owl:Thing** for the sake of simplicity. The interpretation act is associated to a type (**hico:hasInterpretationType**, e.g. paleographic), its authors (**prov:wasAssociatedWith**), motivations (**hico:hasInterpretationCriterion**, e.g. comparison), a datetime (**prov:startedAtTime**), a degree of certainty (**cwrc:hasCertainty**), the source where the statement has been extracted (**fabio:Expression**, e.g. the blog post), and the sources used to support the statement (subproperties of **cito:cites**, e.g. the cited edition). Likewise, relations between statements can be represented by using subproperties of **cito:cites** to link individuals of the class **hico:InterpretationAct**.

JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

Layer 3. Information in this layer answers questions like: Who is responsible for the machine-readable version of the statement? When was it extracted? It represents the meta-context of a statement that has been automatically or semi-automatically generated.

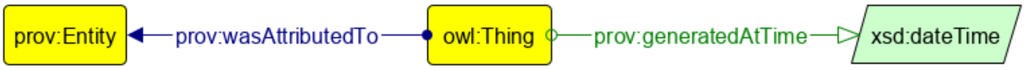


Figure 5. Data model for describing provenance information of technological processes

The property **prov:wasAttributedTo** is used to address the entity (a person or a software agent) responsible for the mining and publication of data, and the property **prov:generatedAtTime** is used to record the publication time. For instance, information about the paleographic statement recorded in the blog post was automatically extracted by a web scraper and transformed into RDF according to the current data model by a reengineering software.

5. Scenarios

We exemplify the usage of the data model in two representative scenarios in the Cultural Heritage domain, namely:

- The representation of multi-perspective analysis of an archival source recorded in a cycle of scholars' lectures. Data is extracted from the cycle of lectures called "Scrivere, rappresentare, conoscere nel rinascimento. Pellegrino Prisciani, un intellettuale eclettico tra la corte e il mondo" that was held by respectively a philologist, a palaeographer, and an art historian, who analysed Pellegrino Prisciani's *Historiae Ferrarie* illuminated manuscripts.[5]
- The representation of inter and intra-textual relations (citations, comments, and translations) between bibliographic resources mentioned in a literary source. Assertions are mined from the TEI/XML encoded edition of Paolo Bufalini's notebook and are included in a semantic scholarly edition.[6]

Lastly, we exemplify the potential of the data model applied into a framework for harvesting, comparing, and ranking contradictory information automatically extracted from cataloguing data. The framework, called mAuth, is composed of an API and web application for assessing and recommending artwork attributions recorded in archival and photographic documentation.

5.1 The lectures on Pellegrino Prisciani's *Historiae Ferrarie*

The objectives of this project are (a) to identify differences and commonalities between different hermeneutical approaches, (b) showcase real-world examples of statements belonging to different scholarly approaches, (c) validate the portability of the data model across domains, and (d) support the production of the data model documentation.

[5] Historiae Ferrariae: ASMO, Manoscritti, nn. 129–133: Pellegrino Prisciani, voll. I, IV, VII, VIII, IX.
[6] The digital edition is available at http://projects.dharc.unibo.it/bufalini-notebook/, DOI: 10.6092/unibo/amsacta/6415. The dataset and the code are available at https://github.com/marilenadaquino/bufalinis-notebook. Accessed May 10, 2020.

A bottom-up approach was used to select relevant, representative, adequate, and coherent sources. Document analysis was performed by skimming (superficial examination), reading (thorough examination), and interpreting contents which may be fuzzy, incomplete, and contradictory (Bowen 2009, 33). The aim of the analysis is to highlight the underlying coherence of the cognitive structures to be represented in the formal model (Myers and Avison 2002, 10). We designed several competency questions addressing aspects emerged from sources so as to design and refine the data model. An exemplar RDF dataset was created for testing the data model. The documentation, hereafter called MIMA (Multi-disciplinary Interpretations model on Manuscript Apparatus)[7] allows to represent the logical and physical structure of an illuminated manuscript, its contents, and scholars' comments on fragments of the manuscript. Figure 6 illustrates classes and properties of MIMA for representing scholars' assertions on the manuscript at hand (Layer 1).

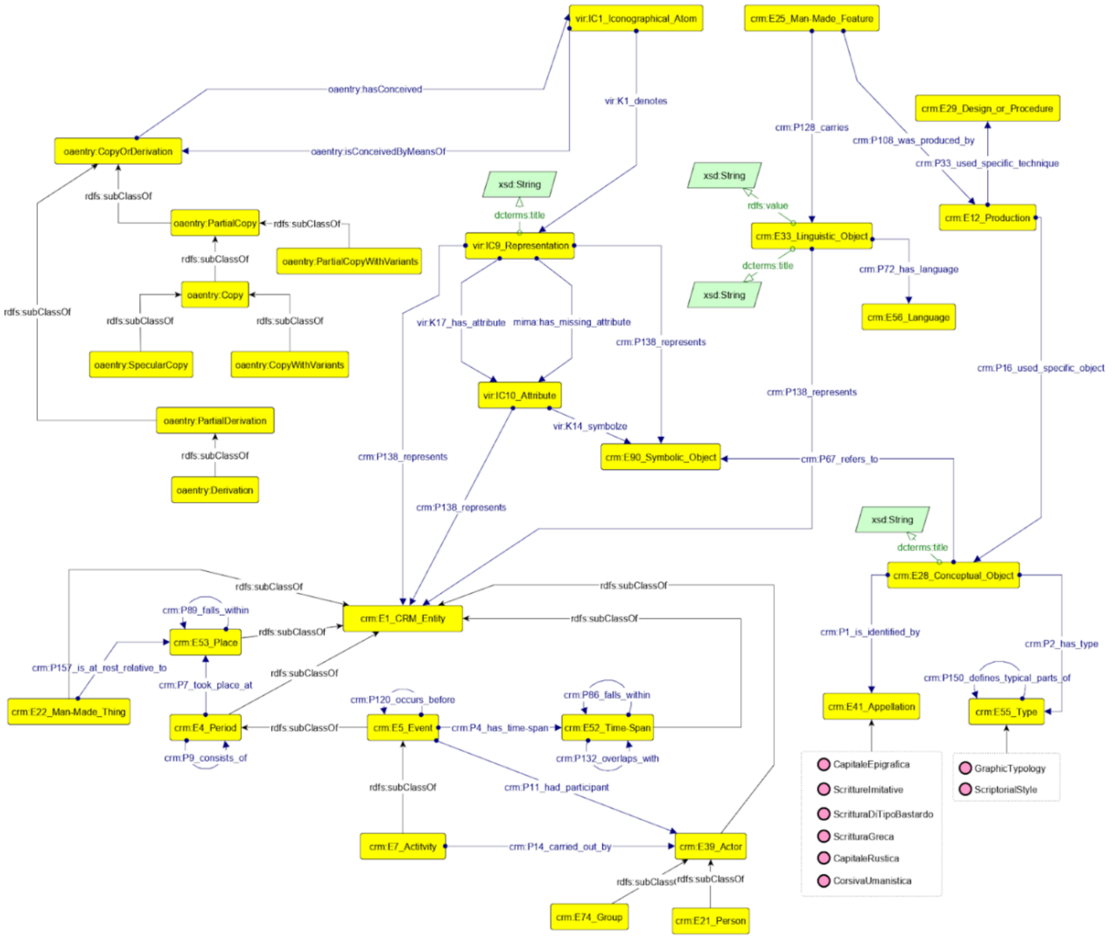


Figure 6. MIMA properties and classes to represent scholars' assertions

[7] The lectures, the dataset, and the documentation of the data model are available at https://github.com/mima-data-model/mima-documentation/. Accessed May 10, 2020.

JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

A common scenario of multi-disciplinary analysis is the attribution of a conceptual meaning (respectively **vir:IC9_Representation** in case of images, and **crm:E33_Linguistic_Object** in case of text fragments) to physical features of the manuscript (instances of **vir:IC1_Iconographical_Atom** and **crm:E25_Man-Made_Feature** classes).
For instance, consider the following scenario in natural language: *The usage of inscriptional capitals is a means to convey the monumentality of the work "Historiae Ferrariae"*.
A scholar analyses the writing process (**crm:E12_Production**), characterised by a specific writing system (e.g Latin alphabet, **crm:E29_Design_or_Procedure**), a scriptorial style (e.g. the *corsiva umanistica* **crm:E28_Conceptual_Object crm:P2_has_type crm:E55_Type**), and other features (e.g. bindings on the right side, **crm:P150_defines_typical_parts_of**) and attributes to a graphic typology a symbolic meaning (**crm:E28_Conceptual_Object crm:P67_refers_to crm:E90_Symbolic_Object**). Figure 7 shows a graphical representation of the example described according to the data model.

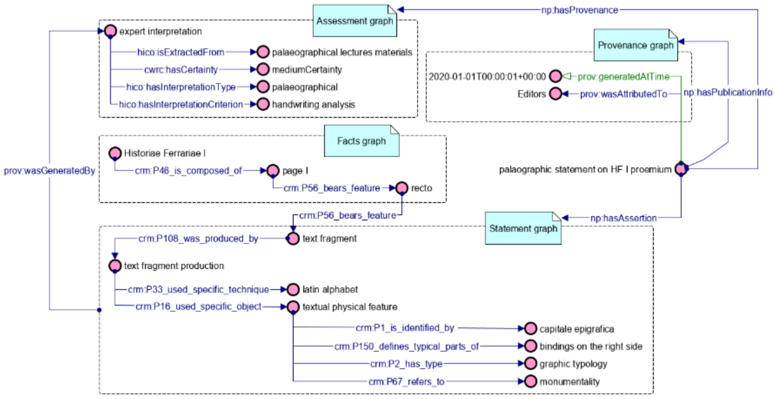


Figure 7. Graphical representation of a statement on a physical feature of the manuscript

The data model allows one to (1) clearly distinguish factual information from questionable contents, (2) to represent all the aspects that characterise the latter as separate annotations, including both information on hermeneutical approach and digital methods applied to the original source, and (3) to reason on such annotations for validating the statement itself when compared to contradictory statements.

5.2 The scholarly digital edition of *Paolo Bufalini's notebook*

The semantic digital edition of Paolo Bufalini's notebook aims at making explicit the author's view on the relations between works that populate his extensive bibliographic collection. The original notebook includes quotations, translations, and comments on classic Latin works and modern contemporary authors, and highlights a significant number of *inter-textual* and *intra-textual* relations between such works. The edition of the notebook was originally encoded in TEI/XML and secondly reengineered in an RDF dataset now leveraged by the web application dedicated to the exploration of the notebook (Daquino, Giovannetti, and Tomasi 2019).

JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

The core of the dataset represents authors' (including Bufalini and cited authors) and editors' statements on the relations between works, people, and between works and people. Figure 8 provides an overview of main classes and properties that are included in Layer 1.

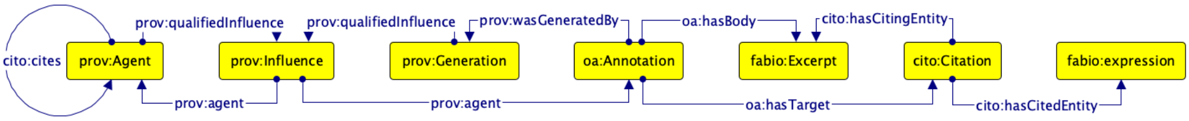


Figure 8. Main properties and classes representing Paolo Bufalini's assertions

For instance, consider the following scenario in natural language: *In an excerpt (comm. n.019) Paolo Bufalini quotes an excerpt of A. Romagnoli's "Opere" (bibl. ref. n.55) in order to support his statement "F. Nietzsche appreciates A. Schopenhauer"*. Figure 9 shows a graphical representation of the statement described according to the data model.

Excerpts of the notebook are represented as individuals of **fabio:Excerpt**, editors' and author's comments as individuals of the class **oa:Annotation**, people are represented as individuals of the class **prov:Agent**, relations between excerpts are defined by means of **cito:Citation**, and influences between excerpts and between people are defined as influences (**prov:Influence**). Specifically, relations between people (e.g. Bufalini comments on Schopenauer's influence on Nietzsche) are represented as an influence of one person over another. Relations between excerpts (e.g. Bufalini quotes an excerpt of Romagnoli's work "Opere") can be represented as a citation when the reference is explicit, or as an influence when less explicit, between an excerpt and the generation (**prov:Generation**) of another excerpt. Relations between people and excerpts can be similarly described. Moreover relations can be further specified by using subproperties of cito:cites (e.g. **cito:agreesWith**, **cito:citesAsRelated**).

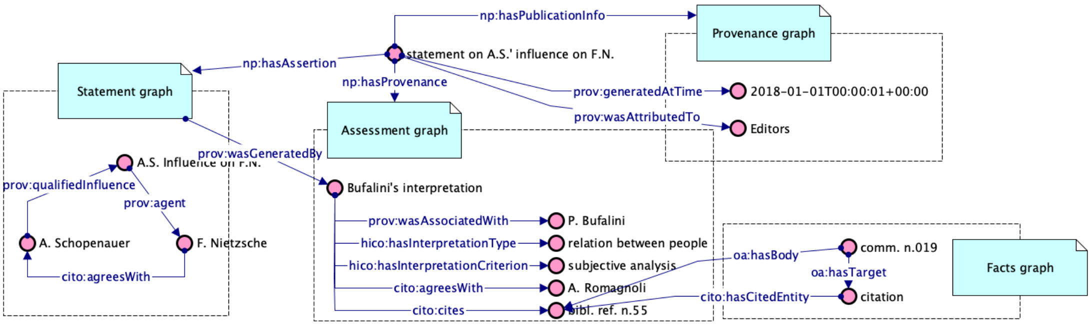


Figure 9. Graphical representation of a statement in Paolo Bufalini's notebook

JLIS.it 11, 3 (September 2020)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12642

Again, the data model allows us to distinguish assertions on real-world entities from the source of information, hermeneutical aspects relevant to such assertions, and to provide mechanisms for uniquely identifying complex statements by means of a citable URI.[8]

5.3 *MAuth*, a recommending system of artwork attributions

mAuth - Mining Authoritativeness in art history[9] is a semantic crawler and recommender system of attributions recorded in online catalogues of art historical photo archives and general purpose websites (e.g. Wikidata).[10] It is based on a framework of assessment methods that support users' decision-making process when validating contradictory artwork attributions (Daquino 2019b). In detail, given the URL of a cataloguing record describing an artwork, the application fetches data about the artwork at hand from several online Linked data sources, extracts information about motivations and sources, reengineers data, and returns an ordered list of attributions accompanied by information to support the assessment.

For instance, following the example presented in the introduction, a research in mAuth on the artwork "Tre Grazie" returns four competing attributions. For each retrieved attribution, information returned includes:

- Sources: Bibliographic data of the artwork (title, author, and date) and bibliographic resources that confirm the attribution.
- Motivations: a short description of the motivation or sources justifying the preferred attribution at hand (e.g. scholars' verbal communication, bibliography)
- Agents: the cultural institution issuing the cataloguing record and scholars that claimed the attribution.
- Relations: citations of bibliographic resources, people's claims, and relations with competing attributions.

Attributions are sorted according to a rating system. Several assessment methods were used to reason on argumentations and obtain a ranking score. The final score is obtained by summing partial scores that represent the following features: relevance, based on the number of sources agreeing on the attribution at hand; reputation, based on white lists of domain-experts (i.e. whether providers are cultural institutions or general purpose websites); reliability, based on a ranking of motivations justifying the attribution; and timeliness, i.e. the recentness of sources recording the attribution. Moreover, two informative scores are presented to users to support their evaluation of first-hand attribution providers' reputation, namely: h-index and acceptance rating. The former is representative of the number of scholars' citations among data sources, while the latter is a percentage that refers to the number of times attributions claimed by the scholar have been accepted.

The system has been evaluated by means of a user study (Daquino 2020) performed by over 30 domain experts (including art historians, cataloguers, and art history teachers) that confirmed the usefulness of such tools. Few drawbacks still affect the reliability of results, such as the lack of extensive citation

[8] See for instance the URI identifying the statement in the example https://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-comm-019-np.

[9] Available at http://purl.org/emmedi/mauth/search.

[10] See http://wikidata.org/.

indexes for humanists and the definition of metrics that take into account the variety of citation forms that characterise scholarly networks in the Humanities (e.g. citing a cataloguing record or a verbal communication).

6. Discussion and conclusion

The proposed data model allows one to formalise aspects that characterise the hermeneutical approach used by scholars when validating their hypotheses on a questionable statement. The usage of Semantic Web technologies allowed us to build a conceptual framework of concepts, properties, and terms that can be used for reasoning on consistent data and enable validation tasks. The usage of existing vocabularies and technologies is a necessary, although not sufficient, condition to achieve portable, shareable software solutions across disciplines, avoiding case-by-case designed strategies. Moreover, the proposed data model is meant to provide future ontology designers of a comprehensive decision tool for designing scalable Linked Open Data applications in the Humanities.

The advantages of having a portable data model for representing hermeneutics are several. The increasing interest in the Open Science community to foster research reproducibility by means of FAIR data (Wilkinson *et al.* 2016) is compelling for researchers in the Humanities. Humanities open data must comply with requirements of quality, accuracy, and shareability. Current models and methods in Digital Humanities are notable but scattered attempts to develop guidelines and methodologies that are rarely able to interoperate across domains, and straightforward data reuse in different contexts is hampered. Having a data model that allows us to identify, cite, and reason on argumentations around questionable statements is a necessary step in this direction.

The limit of the data model is its full portability across domains. As a matter of fact, the Humanistic discourse is broad, heterogeneous, and multi-faceted. Data to be included in Layer 1 cannot be clearly, uniquely identified, since the variety of topics cannot be reduced into replicable patterns. Likewise, assessment methods for reasoning on data included in Layer 2 are highly domain-dependent. While knowledge extraction of aspects relevant to hypotheses validation can be effectively automatised, the definition of assessment methods requires an initial phase of knowledge acquisition (e.g. interview with domain experts, data exploration).

In future works we aim at filling the gap in two directions, namely: (1) to define portable representation patterns peculiar of the Humanities discourse, so as to provide guidance on the definition of contents that populate Layer 1, and (2) to examine how assessment methods can be generalised and tuned according to the discipline or a school of thoughts at hand.

References

Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. "Methodologies for Data Quality Assessment and Improvement". *ACM Computing Surveys* 41 (3):1–52. https://doi.org/10.1145/1541880.1541883.

Bizer, Christian, and Radoslaw Oldakowski. 2004. "Using Context-and Content-Based Trust Policies on the Semantic Web". In *Proceedings of the 13th International World Wide Web Conference on*

Alternate Track Papers Posters, 228–229. New York: Association for Computing Machinery. https://doi.org/10.1145/1013367.1013409.

Bowen, Glenn A. 2009. "Document Analysis as a Qualitative Research Method". *Qualitative Research Journal* 9 (2):27–40. https://doi.org/10.3316/QRJ0902027.

Capurro, Rafael. 2000. "Hermeneutics and the Phenomenon of Information". *Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology* 19:79–85.

Carboni, Nicola, and Livio de Luca. 2019. "An Ontological Approach to the Description of Visual and Iconographical Representations". *Heritage* 2 (2):1191–1210. https://doi.org/10.3390/heritage2020078.

Carroll, Jeremy J., Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. "Named Graphs". *Journal of Web Semantics* 3 (4):247–67. https://doi.org/10.1016/j.websem.2005.09.001.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. 2011. "Definition of the CIDOC Conceptual Reference Model". Version 5.0.4. ICOM/CIDOC CRM Special Interest Group. http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html.

Daquino, Marilena. 2019a. "Art Historical Photo Archives and Semantic Web : Problems, Resources and Research Lines". *JLIS.It* 10 (2):37–47. https://doi.org/10.4403/jlis.it-12533.

———. 2019b. *Mining Authoritativeness in Art Historical Photo Archives*. Studies on the Semantic Web 40. Amsterdam: IOSPress.

———. 2020. "A Computational Analysis of Art Historical Linked Data for Assessing Authoritativeness of Attributions". *Journal of the Association for Information Science and Technology* 71 (7):757–69. https://doi.org/10.1002/asi.24301.

Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi. 2019. "Linked Data per Le Edizioni Scientifiche Digitali. Il Workflow Di Pubblicazione Dell'edizione Semantica Del Quaderno Di Appunti Di Paolo Bufalini". *Umanistica Digitale* 3 (7). https://doi.org/10.6092/issn.2532-8816/9091.

Daquino, Marilena, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 2017. "Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data". *Journal on Computing and Cultural Heritage* 10 (4):1–21. https://doi.org/10.1145/3051487.

Daquino, Marilena, and Francesca Tomasi. 2015. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects". In *Metadata and Semantics Research*, edited by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, 544:424–36. Cham: Springer International Publishing.

De Waard, Anita, and Jodi A. Schneider. 2012. "Formalising Uncertainty: An Ontology of Reasoning, Certainty and Attribution (ORCA)". In *Joint Workshop on Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine*, 930:8–15.

Felicetti, Achille, and Francesca Murano. 2017. "Scripta Manent: A CIDOC CRM Semiotic Reading of Ancient Texts". *International Journal on Digital Libraries* 18 (4):263–270. https://doi.org/10.1007/s00799-016-0189-z.

Gil, Yolanda, and Donovan Artz. 2007. "Towards Content Trust of Web Resources". *Web Semantics: Science, Services and Agents on the World Wide Web* 5 (4):227–39. https://doi.org/10.1016/j.websem.2007.09.005.

Ginzburg, Carlo. 1979. "Clues: Roots of a Scientific Paradigm". *Theory and Society* 7 (3):273–88.

Groth, Paul, Andrew Gibson, and Jan Velterop. 2010. "The Anatomy of a Nanopublication". *Information Services & Use* 30 (1–2):51–56. https://doi.org/10.3233/ISU-2010-0613.

Lehnert, Wendy C., Hayward R. Alker, and Daniel K. Schneider. 1983. "The Heroic Jesus: The Affective Plot Structure of Toynbee's Christus Patiens". In *Proceedings of the Sixth International Conference on Computers and the Humanities*, edited by Sarah K. Burton and Douglas D. Short, 358–367. Rockville: Computer Science Press.

Maas, Paul. 1972. *Critica Del Testo*. Firenze: Le Monnier.

Maiatsky, Michail, Alexey Boyarsky, Natalia Boyarskaya, Ekaterina Velmezova, and Michael Piotrowski. 2018. "VICOGLOSSIA: Annotatable and Commentable Library as a Bridge between Reader and Scholar (a Proof of Concept Study: Early Soviet Philological Culture)". *Umanistica Digitale*, no. 2:161–84. https://doi.org/10.6092/issn.2532-8816/7253.

Mallery, John C., Roger Hurwitz, and Gavan Duffy. 1986. "Hermeneutics: From Textual Explication to Computer Understanding?" A.I. memo 871. MIT artificial intelligence laboratory.

Moreau, Luc, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles. 2015. "The Rationale of PROV". *Web Semantics: Science, Services and Agents on the World Wide Web* 35:235–57. https://doi.org/10.1016/j.websem.2015.04.001.

Morelli, Giovanni. 1883. *Italian Masters in German Galleries : A Critical Essay on the Italian Pictures in the Galleries of Munich, Dresden, Berlin*. Translated by Louise M. Richter. London: G. Bell and Sons.

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Myers, Michael D., and David Avison. 2002. "An Introduction to Qualitative Research in Information Systems". In *Qualitative Research in Information Systems: A Reader*, 3–12. SAGE Publications.

Naumann, Felix, and Claudia Rolker. 2005. "Assessment Methods for Information Quality Criteria". In *Proceedings of 5th International Conference on Information Quality*, 148–162.

Pasquali, Giorgio. 1952. *Storia Della Tradizione e Critica Del Testo*. Firenze: Le Monnier.

Peroni, Silvio, and David Shotton. 2018. "The SPAR Ontologies". In *The Semantic Web – ISWC 2018*, 119–36. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_8.

Ramsay, Stephen. 2010. "The Hermeneutics of Screwing Around; or What You Do with a Million Books". In *Playing with Technology in History Conference*. Niagara-on-the-Lake.

Rieh, Soo Young. 2002. "Judgment of Information Quality and Cognitive Authority in the Web". *Journal of the American Society for Information Science and Technology* 53 (2):145–61. https://doi.org/10.1002/asi.10017.

Romele, Alberto, Marta Severo, and Paolo Furia. 2018. "Digital Hermeneutics: From Interpreting with Machines to Interpretational Machines". *AI & SOCIETY*, 1–14. https://doi.org/10.1007/s00146-018-0856-2.

Sanderson, Robert, Paolo Ciccarese, Herbert Van de Sompel, Shannon Bradshaw, Dan Brickley, Leyla Jael Garcia Castro, and Timothy Clark. 2013. "Web Annotation Data Model". W3C community draft. https://www.w3.org/TR/annotation-model/.

Schneider, Jodi, Tudor Groza, and Alexandre Passant. 2013. "A Review of Argumentation for the Social Semantic Web". *Semant. Web* 4 (2):159–218.

Stegmüller, Wolfgang. 1977. "The So-Called Circle of Understanding". In *Collected Papers on Epistemology, Philosophy of Science and History of Philosophy*, 1–25. Dordrecht: Springer Netherlands.

Van Zundert, Joris J. 2016. "Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics". In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 331–47. Oxford: Blackwell Publishing.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, *et al.* 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Wilson, Patrick. 1983. *Second-Hand Knowledge: An Inquiry into Cognitive Authority*. Westport: Greenwood Press.

Zaveri, Amrapali, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2016. "Quality Assessment for Linked Data: A Survey". *Semantic Web* 7 (1):63–93.