RETHINKING CRITICAL EDITIONS OF FRAGMENTARY TEXTS BY ONTOLOGIES

Matteo Romanello, Monica Berti, Federico Boschetti, Alison Babeu, Gregory Crane

The Perseus Project, Tufts University
Eaton 124, Medford MA, 02155, USA
e-mail: matteo.romanello@tufts.edu; federico.boschetti@yahoo.com;
monica.berti@tufts.edu; alison.jones@tufts.edu; gregory.crane@tufts.edu

Abstract

This paper discusses the main issues encountered in the design of domain ontology to represent ancient literary texts that survive only in fragments, i.e. through quotations embedded in other texts. The design approach presented in the paper combines a knowledge domain analysis conducted through semantic spaces with the integration of well established ontologies and the application of ontology design patterns. After briefly describing the specific meaning of "fragment" in a literary context, the paper gives insights into the main conceptual issues of the ontology design process. Lastly, it outlines the overall architecture of protocols, services and data repositories which is required to implement a digital edition of fragments based on the proposed ontology.

Keywords: Digital library; fragmentary texts; domain ontology; ontology design.

1. Introduction

Recently, cyberinfrastructure has been defined as the technological infrastructure needed to properly support the broad development of computing across the disciplines including the humanities. One facet of this humanities cyberinfrastructure is to develop new models and tools, such as the development of standards to represent new digital editions of ancient texts [1]. Standards, protocols and tools now available to scholars constitute the starting point to deal with one of the most challenging problems, the digital representation of collections of fragmentary texts: this requires us to rethink critical editions by shifting from a printed-based representation to a digital one.

By fragmentary texts we mean texts that have been preserved only in fragments, i.e. through quotations by other surviving authors, who quote, paraphrase, summarize, or allude to authors and works that have not survived. Thus fragmentary texts are embedded in surviving works, and collecting fragments means first of all extracting quotations from their contexts. The modern term used to define the source-author of a fragment is "witness", i.e. the author who has quoted the thought and/or the work of another author; the witness can be considered the source of information for a lost work.

In the 19th and 20th centuries many collections of fragmentary authors have been edited, reconstructing works and personalities otherwise lost and forgotten. These collections of fragments contain excerpts from many different sources and can be considered paper representations of hypertexts. New technologies allow philologists to go beyond these collections and the limits of printed editions, constructing editions that are truly hypertextual, including not only excerpts but links to the sources from which the excerpts have been extracted. The work described in the paper is being conducted to provide the Perseus Digital Library with its first collection of fragments [2], taking the subset of Greek historical fragments as an initial testbed[1].

The main goal of this paper is to formalize with ontology – thus making evident and readable for a machine – the semantic contents of modern critical editions that in a printed context are usually expressed by using typographical features. In particular, a digital environment allows us to go well beyond the limits posed by printed editions, expressing in a fuller way the fundamental tools used by scholars to represent some complex relationships among text editions (tables of concordances) and interpretations of texts (critical apparatuses).

The key problem we address here is what semantic contents survive and what we need in order to represent these contents digitally, particularly what is needed in terms of knowledge representation and architecture once we change the medium used to represent critical editions (in this case editions of fragmentary texts).

2. Background

Fragmentary texts are essential to our knowledge of classical (Greek and Latin) literature because they allow us to recover an inestimable cultural heritage. Their

[1] Greek historical fragments are fragments of ancient works written by authors interested in various aspects of ancient Greek history. On the subject see Schepens, G. Jacoby's *FGrHist*: Problems, Methods, Prospects. In Most, G.W., editor. *Collecting Fragments*. Göttingen 1997, pp. 144-172.

importance has also be proven from a quantitative point of view by the results of an analysis we conducted on the data contained in the *Thesaurus Linguae Graecae* (TLG-E), which is currently the reference digital library for Greek literature. For the period between the 8th century B.C and the 3rd century A.D. included, 59% of the authors are preserved only in fragments, 12% are known from both entirely preserved works and fragments, whereas only 29% are known just by entirely preserved works.

The TLG-E includes for each ancient work one canonical edition without critical apparatus. In terms of fragmentary texts in the TLG, we have both editions of fragments and editions of the sources from which the fragments have been extracted: the result is that the text of a fragment is published twice, once in the edition of the fragment and secondly in the edition of the source-author of the fragment, replicating how those texts are published in a printed context (where it is impossible to use a hyperlink to avoid the duplication of a portion of text). In the TLG data, therefore, the text of fragments and their witnesses is duplicated, leading to a certain inconsistency for further quantitative analysis on those data/texts.

The ultimate goal of the work described in this paper is the creation of a digital collection where users can read the sources preserving fragments in multiple editions and with critical apparatuses, and where the hypertextual and hermeneutical nature of fragmentary texts (see Section 5) is more properly represented.

The term "fragment" in the context of literary criticism has a technical meaning, which is slightly different from its meaning in the current use or in the field of computer science. A literary fragment may have multiple sources that have been identified by scholars as sources of information about a lost work, whereas a XML fragment, for instance, is simply a smaller section of one document intended as a whole. Formally we could define a literary fragment as a discontinuous fragment whose discontinuity can exceed the boundaries of a single textual unity.

To sum up, the following are the main issues posed to scholars by fragmentary texts:

- identification of the witness of the fragment (i.e., the source-author who has preserved the fragment) and assessment of his reliability;
- identification of the boundaries of the fragment (i.e., beginning and end of the fragment in the context where it is preserved);
- attribution of the fragment to an author and a work, and collocation inside the narrative (or dramatic) structure of the original work to which the fragment belonged;
- dating of the content of the fragment on the basis of the *realia* (such as historical events and names) eventually mentioned.

3. Approach

An ontology is the most suitable solution to represent critical editions of ancient texts for two main reasons: first, we want to be able to link different kinds of resources (page images as PDF, texts as (X)HTML or XML) that have in common the possibility of being referred to via URIs, which is one of the principles of the Semantic Web; second, information contained in critical editions constitutes a layer of interpretations and a description of relations about texts that is important to keep clearly distinct from the texts themselves. Indeed, the use of stand-off metadata encoded within ontology allows us to express an open-ended number of interpretations, whereas a markup-based solution would not make this possible due to obvious reasons of overlapping hierarchies. Using such a formalism affects the way we can access data, since it will be possible to apply logical reasoning on a knowledge base of ontological data and to use this data to provide semantic information retrieval, as has recently been demonstrated by GoPubMed [3] in the field of medicine.

The approach adopted here aims at reusing existing ontologies rather than at proposing a completely new one: the goal is defining an ontology by subclassing or specializing classes and properties derived from stable and widely adopted existing ontologies, combining them together so that they can be as expressive as possible. In particular, we pursued the goal of compatibility between our ontology and the CIDOC Conceptual Reference Model (CIDOC-CRM) for the sake of interoperability over the long-term. Indeed, in the field of humanities, the CIDOC-CRM has emerged as a bridging solution to make interoperable, for example, different digital collections of archeological data [4, 5].

Finally, in order to give the designed ontology a more solid structure and to reduce the arbitrariness of the knowledge representation it expresses, we have conducted the preliminary knowledge domain analysis by using ontology learning techniques, and then we have refined those results by applying upper level ontologies and ontology design patterns. Regarding the forward compatibility of our ontology with future possible developments, we believe that as long as the ontology design has been based on evidence that emerged from the application of ontology learning techniques to a corpus of texts, it should be possible to extend the ontology as necessary with new methods or for new texts.

4. Knowledge Domain Analysis

Preliminary knowledge domain analysis for the ontology design was based on the exploration of semantic spaces for a corpus of 170 research articles.

According to recent paradigms in ontology learning [6], corpus analysis helps to identify the most relevant terms to describe the concepts involved in the ontology, to cluster them, and to provide evidence about their relations.

We applied a supervised strategy, where the evidences of the automatic procedures are filtered by the agreement of three scholars. The articles were selected by a philologist, specialized in the domain of fragmentary historical literature, from journals of classical philology downloaded from the JSTOR archive. All articles are in English, related to Latin and Greek literature, about different literary genres (e.g., epic, tragic, comic, and historical).

Text was extracted from the original pdf files and processed with Infomap [7], which applies techniques of Latent Semantic Analysis (LSA). Text was preprocessed with TreeTagger [8] for lemmatization and part of speech tagging. The first seed term, "fragment", was used to find the most relevant associations in the top fifty word list provided by the Infomap *associate* tool filtered by part of speech "nn" (noun). By the agreement of experts, terms related to philological issues (such as "reading" and "quotation"), terms related to subjective evaluation and uncertainty (such as "supposition"), and terms concerning the whole/part and spatial relations (such as "block" and "line", or "beginning" and "end") were selected and classified for the second generation of seeds. Within these three categories, new associations are selected adding to the first seed, "fragment", the new relevant word(s), for example "supposition". In this way, new seeds are generated for the next generation, stopping the iterations either when the list of terms associated are all relevant or no relevant new terms are provided in the next generation. At the end of this process, we have lists of terms strongly related (according to the expert agreement) to a specific category associated to the original "fragment" term. For example, for the category of subjective evaluation and uncertainty, we have a final list of terms that contain "possibility", "exception", "debate", "preference", "consideration", "assumption", "caution", "authenticity", "purpose", "strife", "interpretation", "supposition", and then "certainty" and "evidence", both antonyms of "uncertainty".

Finally, terms are clustered with the k-means algorithm, after the reduction of the original semantic space dimensions to two dimensions, in order to represent them in a bi-dimensional graph. The result is shown in Fig. 1.

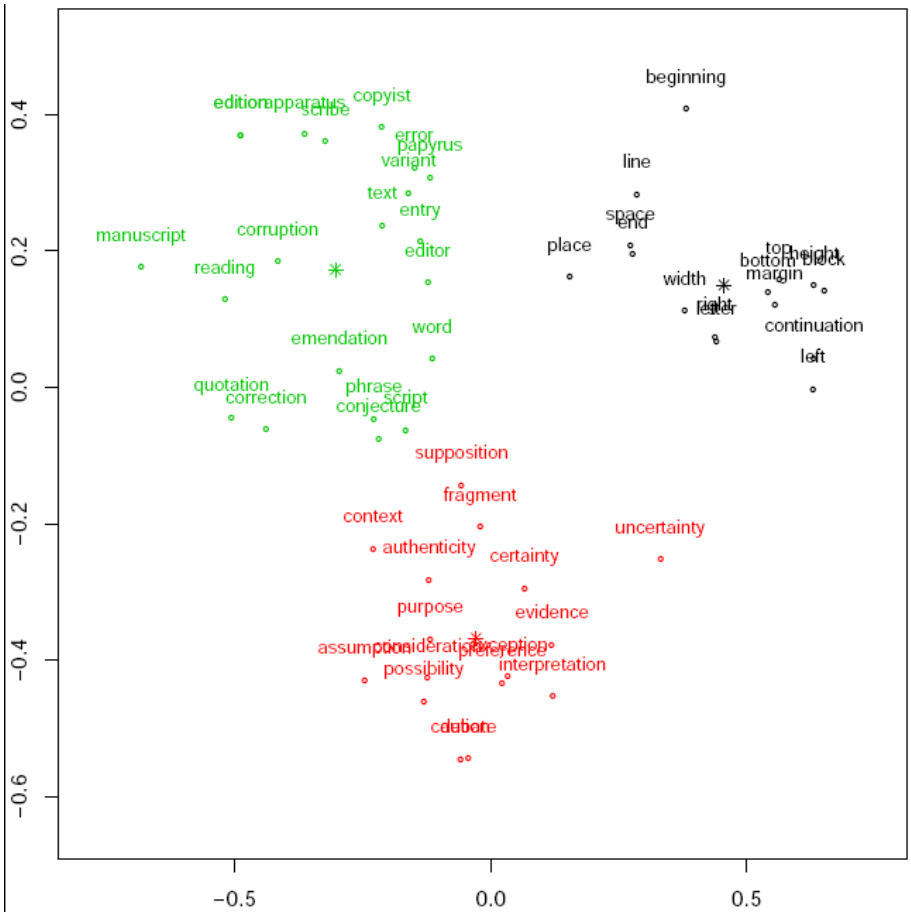

Figure 1: Cluster chart of relevant terms related to “fragment”.

5. Ontology Overview

The proposed ontology[2] is based on the theoretical assumption that fragments do not actually exist outside of scholars’ interpretations. From an ontological perspective, this assumption places fragmentary texts closer to interpretations than to texts. Therefore, a new class called textual-interpretation was created as a subclass

[2] When classes or properties of the ontology are mentioned in the text, they appear respectively in a sans serif font and in italics. When classes or properties of the ontology are mentioned in the text, they appear respectively in a sans serif font and in italics.

of the interpretation class defined by PhiloSurfical, a domain ontology aimed at representing philosophy and philosophy-related concepts. The derivation of fragment from textual interpretation reflects the deep hermeneutical nature of every philological activity, from the individuation of fragments inside surviving texts to the formulation of variants and conjectures aimed at restoring the original text of surviving works.

Moreover, in our ontology we assume that every textual interpretation is supported by a publication where a scholar provides evidences in support of the argued interpretation. In regards to the scientific domain, [9] identified agents, claims and justifications are the main concepts required to provide scholarly discourse with a computable structure. Indeed, in the ScholOnto ontology, scholarly discourse is represented as made up of claims that are submitted by agents (not just scholars but also software agents), and are backed up by justifications expressed within documents of different kinds. At an abstract level, this representation is valid also for the philological domain, where interpretations (fragments, conjectures, etc.) are expressed in and supported by scholarly publications (critical editions, commentaries, papers, etc.). An already existing ontology that is suitable for encoding the bibliographic metadata of modern publications is the Bibliographic Ontology (BIBO) [10]. BIBO's main benefit is that it allows any existing bibliographic legacy schema to be converted into its format, which can help overcome the lack of interoperability due to the large number of available bibliographic formats.

For the representation of surviving works we leveraged the already existing Functional Requirements for Bibliographic Records (FRBR) [11]. The overall structure is currently being reviewed on the basis of CIDOC-CRM principles [12]. Once completed, this process should lead to a FRBR object oriented specification that may replace – even in our ontology – the current OWL implementation. The classes derived from the FRBR ontology present some slight modifications required to fit the needs of properly representing ancient works. In particular, the class text passage was provided with properties referring to the topology of text passages, namely the set of relationships (like inclusion, proximity, etc.) that can be induced when comparing at least two text passages. This topology – even if not yet formalized within an ontology – was defined in the framework of the Canonical Text Services protocol (CTS) [13], which is one of the main components of the architecture identified to implement the ontology (see Section 6 for more details).

Doubt and uncertainty, as it was confirmed by the preliminary knowledge domain analysis, are an essential part of the scholarly discourse about fragments. According to the classification proposed by [14], the uncertainty implied by the philological discourse pertains essentially to the categories of vagueness (i.e., the work X was probably written around Y date but we do not have enough evidences

to prove it) and belief-function (i.e., W attributes the work X to author Z, where Y attributes it to Q). In the printed reference edition of historical fragmentary texts edited by Jacoby[3], an uncertain attribution is pointed out by using question marks (one or more indicating different degrees of uncertainty), and then it is explained with more details in the commentary. Once the uncertainty implied by scholars' interpretation is made explicit, it becomes possible to take it into account when displaying information to users, or when retrieving information from the knowledge base. For instance, scholars could benefit from a functionality allowing them to look for just those fragments whose date is uncertain or that were uncertainly attributed to different authors or works by different scholars. At this moment, we take into account in our ontology the work done by the W3C Uncertainty Reasoning for the World Wide Web XG Group [15] to represent ontologically doubt and uncertainty.

Given these initial remarks about the hermeneutical nature of fragments and the other ontologies involved, we next describe the main ontology concepts and the rationale behind the choices made during the process of ontology design.

5.1. Fragment and Witness

Fragments and canonical texts pertain to two different ontological categories, since the properties and the axioms that are true for the former are not always true for the latter. Empirically, this fact can be observed if we try to apply the FRBR model to fragmentary texts. The basic assumption of this model is that for each work some manifestations and expressions exist. Homer's *Iliad* has both a notional concept (the concept of "Iliad") and as many expressions as the existing modern editions of this text. On the contrary, for a work that has only survived through fragments, this generalization is not valid. If we consider, for example, the lost work *Atthis* written by Hellanicus, we can refer to the notional concept for this work but there are no editions of the *Atthis* that we can properly regard as expression exemplars in the meaning defined by FRBR. Indeed, fragments are generally published in critical editions where they are grouped by literary genre, chronological order, or by the authors they are attributed to.

Fragments are always scholarly reconstructions and interpretations of the content and structure of lost works. Scholars base their hypotheses about fragments on the interpretation of passages of surviving works that bear evidence for lost authors and works. When scholars publish fragments they usually indicate for each fragment the surviving works that bear witness for it. Texts of this kind are called

[3] *Die Fragmente der Grieschischen Historiker*, v. Jacoby, F. I-III. Berlin - Leiden 1923-1958.

witnesses, but from an ontological point of view "to be witness for a fragment" is the role played by specific text passages in relation to a fragment, rather than a kind of text.

In our ontology, the text passage class acts as a bridge between fragments and surviving works. A fragment is a scholarly interpretation which has already been published and can have one or more sources, namely text passages of surviving works that bear witness to the lost work to which a fragment is attributed. A text passage always refers to a specific edition: in the ontology this is expressed by the fact that a text passage is a subclass of FRBR's ExpressionFragment, and therefore the scope of a passage is always its reference edition. When dealing with multiple editions of texts, and even more so when dealing with variants and conjectures, it is of primary importance to refer every text passage to an existing edition, no matter if printed or digital. Since each editor can establish a different text for a given passage, text passages used without reference to an edition – unless it is implicit – are not precise enough.

The last facet of text passages to be considered is their granularity. By granularity we mean the precision with which we can point to the word span of a text passage. The top right cluster in Fig. 1 clearly shows how philologists use certain terms as a coordinate system to refer precisely to texts. One of the most frequently discussed problems concerning fragments is determining where a fragment starts and where it ends, or in other words, figuring out what words contained in a given source text passage pertain precisely to a fragment. In a digital context, we need pointers that are granular enough to allow us to address single words and even single characters of a text passage, such as CTS URNs in the CTS protocol. Since not every kind of resource available on the Web, however, is provided with a likely pointing mechanism, we use highly precise unique identifiers only for those resources already available within CTS services.

5.2. Attribution, Classification and Ordering

An ontology devised to properly represent fragments needs to handle one of the most frequent scenarios found in the scholarly discussion about fragments: scholars may disagree – as it often happens – about the attribution of a fragment to an author or to the work to which it originally pertained to (because in many cases the title of the work is not cited in the quotation of the fragment).

Taxonomies and classifications used by scholars to organize fragments may overlap and change. For instance, in the reference edition of Presocratic philosoph-

ical fragments edited by Diels and Kranz[4], the fragments are divided into three main categories: 1) *testimonia;* 2) *ipsissima verba;* 3) *imitations*. Jacoby in his edition of historical fragments, however, simply distinguishes between fragments bearing evidence about the author's life, called *testimonia*, and content fragments of the lost works, called properly *fragmenta*. Applying the "classification pattern" [16] to this problem, it is possible to correctly represent this complex reality of fragment classifications, by linking an attribution to a fragment via the *has-attribution* property. What is important here is to keep all the possible classifications clearly distinct from the actual ontological classes, in order to improve and ensure the ontology's applicability to different genres of fragments.

As far as concerns fragmentary texts, we deal substantially with the following kinds of attributions: 1) chronological attributions (i.e., dating the fragment and its content), represented by the class **DATE ATTRIBUTION**; 2) attribution to an ancient lost work, or even to a work section, represented by **WORK ATTRIBUTION**; 3) attribution to an author which corresponds to **AUTHOR ATTRIBUTION**. The superclass **ATTRIBUTION** was introduced for reasons concerning the reification of statements and particularly in order to be able to associate an uncertainty of some degree to attribution statements. Dealing with fragments means also dealing with the attempt of scholars to date fragments on the basis of the events described or alluded to by the fragments themselves. The Historical Event Markup Language (HEML) provides a suitable RDF model to encode chronological concepts. Since it has been proven that HEML can be integrated with both CIDOC-CRM and with the CTS protocol [17], it will be possible to include it in our architecture and rely on it to encode dates and events.

Furthermore, even the order chosen by the editor to arrange the fragments in the printed edition is meaningful since it subsumes a hypothetical reconstruction of the lost original narrative sequence. The property *precedes*, along with its inverse property *follows*, has been introduced to record and make evident this implicit interpretation about the original structure and development of a fragmentary work. Indeed in the case of fragments from dramatic plays (like tragedies and comedies), as well as in the case of fragments from historical works, different choices about positioning a fragment in the overall structure of the text can noticeably change its meaning.

5.3. Variants and Conjectures

Critical apparatus is the term by which philologists usually refer to the page sec-

[4] Diels, H. - Kranz, W. *Die Fragmente der Vorsokratiker.* I-III. Berlin 1951-1952[6].

tion of a critical edition where variant readings and conjectures are recorded and presented to the reader. However, the concept of critical apparatus is not simply applicable to the organization of information in a printed medium. If we formalize its semantics, what we observe is the survival of the concepts of variant readings and conjectures. The main novelty of representing them with the proposed ontology is the possibility of representing and accessing textual interpretations beyond the limits of printed books and disciplinary fields.

Since the text of fragments is essentially the text of their witnesses with the addition of scholars' textual interpretations, the reading and conjectures recorded in the apparatus of an edition of fragments actually refer to the text of those witnesses. Once we are able to overcome the physical limits of printed editions by joining together variants and conjectures referring to the same texts, it also becomes possible to look at the texts from a new and broader perspective, with possible consequences for our knowledge and comprehension of them.

For instance, many scholars working on Athenaeus would also like, when looking at the text of his works, to be able to take into account the variants and conjectures recorded by those scholars who also edited fragments for which Athenaeus bears evidence. As text passages always refer to a given edition, variants and conjectures also need to be referred to a specific edition on which they can be mapped, in order to be correctly interpreted. Further problems related to variants and conjectures in a digital environment include their automatic extraction from critical apparatuses and how to map them to the text passages referred to [18].

5.4. First Ontology Population

In this section we describe a first attempt to populate the ontology by leveraging one of the aforementioned printed reference tools, tables of concordances. This approach – aimed at leveraging the formal structure of printed reference materials – is generalizable to fragments other than historical ones and can be applied to other materials provided that they are represented in a consistent and structured format, such as tables of chronological dates or indices of names [19].

Tables of concordances contain "hidden semantics", such as a list of equivalence statements about entities, resulting in triples like "X is the same as Y". In this case, the entities are fragments in different reference editions. A typical concordance entry for historical fragments would be "FGrHist 323a F 2 = FHG I 371"[5]. This

[5] FGrHist is the conventional abbreviation for Jacoby's work, while FHG is the conventional abbreviation for Müller's work (*Fragmenta Historicorum Graecorum.* I-V. Coll. Müller, K. - Müller, T.

concordance means that the same fragment was published by Jacoby as fragment 2 of author 323a (= Hellanicus) and by Müller at page 371 of volume one of his collection (where Hellanicus' fragments are published). It is worthwhile to note here how this statement does not actually mean that the text of the two fragments established by the respective editors is the same. Indeed, each editor may have printed the fragment text accepting different scholars' interpretations (i.e., conjectures) or variant readings attested by the manuscripts.

Specifically, we converted the tables of concordances of Jacoby's edition recording equivalences between fragments as numbered in Jacoby's and Muller's edition. By combining automatic parsing and a few manual adjustments to the OCR of those tables, it was possible to extract from them the following information. Firstly, each fragment was assigned a unique identifier, since we needed to refer to them as discrete objects in order to express ontological statements about them. The labels used in printed editions to refer to fragments are encoded as instances of the canonical reference class, which is a subclass of CIDOC-CRM's appellation. Practically, this means that a fragment can have multiple labels associated to it, which is important because scholars often differ in both the abbreviations and the format of canonical references they use to refer to ancient texts. Secondly, we encoded equivalence statements between fragments by using OWL's *same-as* property. Lastly, since the tables of concordances are organized by author name it was possible to encode the attributions of fragments to ancient authors according to Jacoby's interpretations (e.g., fragment N is attributed to author Z).

6. Representing Fragments by Ontologies: Proof of Concept

In this section we give a proof of concept of how the above illustrated concepts harmonize with each other into an ontological representation of fragmentary texts. For the sake of clarity, some minor details were not taken into account in the diagrams illustrating the relationships between classes, instances and properties (Figs. 2, 3). Only classes and properties borrowed from other ontologies are prefixed by namespaces that refer mostly to well known or above mentioned ontologies.

We consider as an example the fragment 10 of Istrus the Callimachean in Jacoby's edition, whose Domain Namespace Id (DNID) [20] can be used as URI identifying the resoruce.

Parisiis 1841-1884.), which is a collection of Greek historical fragments published in the 19th century.

The interpretative act underlying the individuation of this fragment – evident in the hierarchy of the class **FRAGMENT** – is backed by a **BOOK**, namely the critical edition of fragments published by Jacoby himself. The resource representing this edition is the URI pointing to the Library of Congress' record. In the section referred to by the **CANONICAL REFERENCE** "FgrHist 334 F 10", the editor provided evidences supporting the following formal statement: the passage 556f of Athenaeus' *Deipnosophistae* contains (according to Jacoby) a quotation of a lost work by Istrus the Callimachean. This statement is expressed by two properties: *has-source* and *has-attribution*. The former indicates that according to Jacoby a precise passage of the *Deipnosophistae* is considered the witness of a fragment, whereas the latter states that according to the same editor the fragment should be attributed to a person whose Latin **APPELLATION** (in this case the name) was Istrus. Provided that we have no absolutely certain evidences and that an attribution is essentially an interpretation and thus a belief-function, the same fragment may have more than one author attribution.

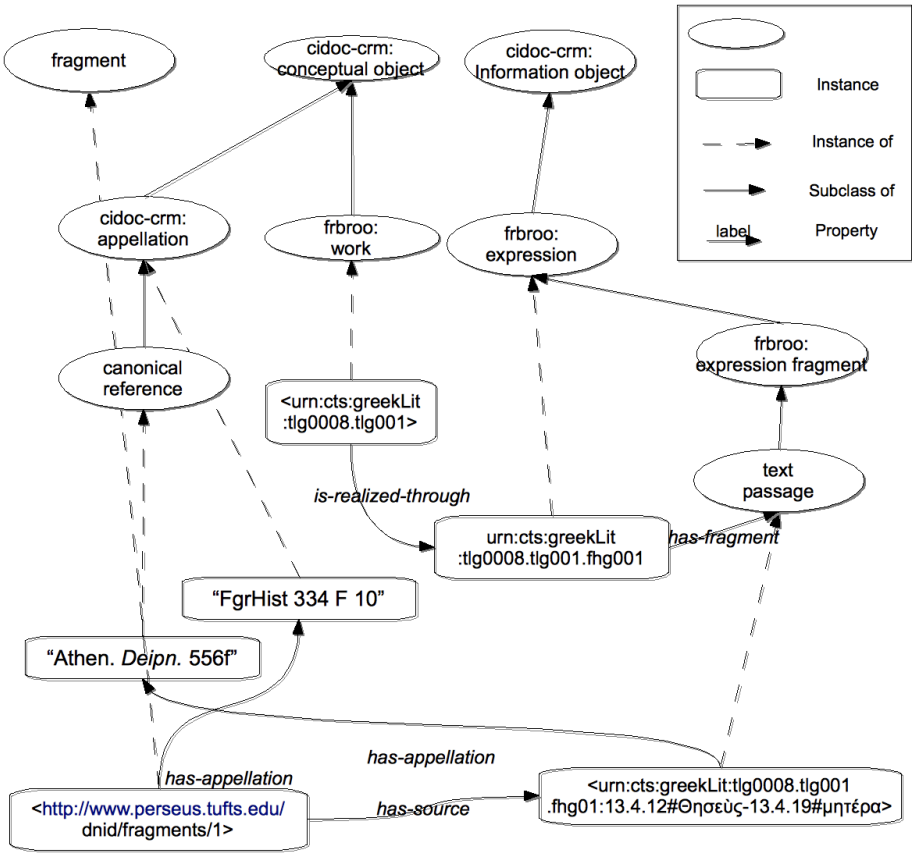

Figure 2: Model for the fragment-witness relationship.

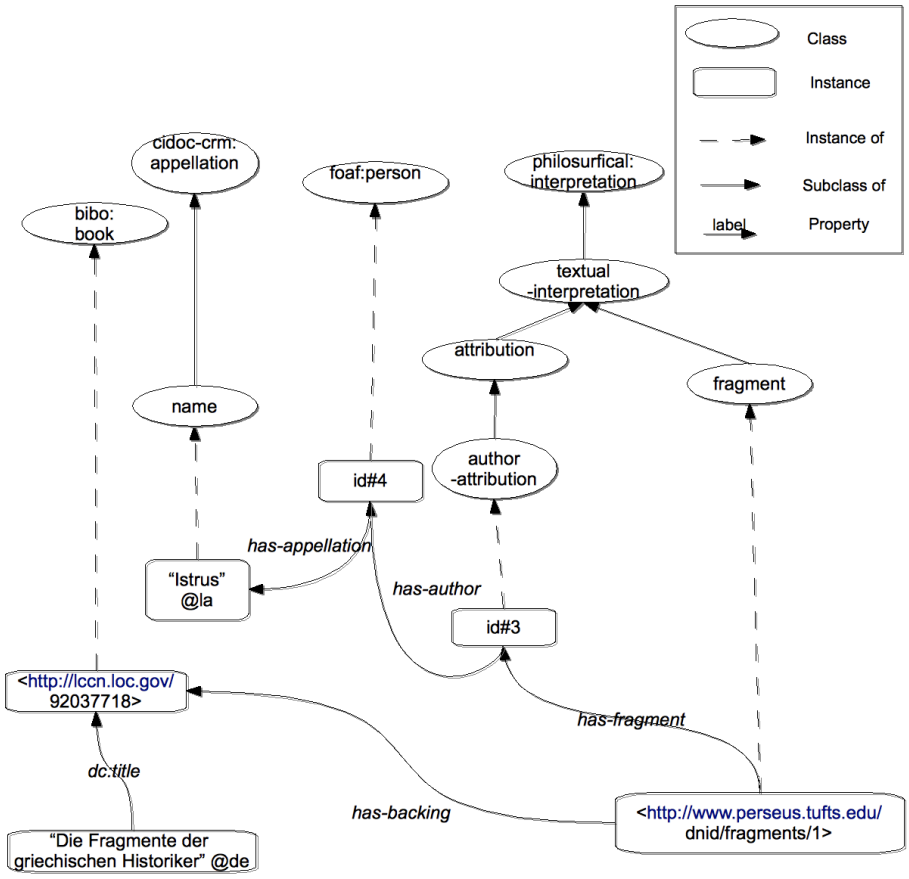

Figure 3: Model for backing and attribution of a fragment.

The passage referred to by the **CANONICAL REFERENCE** "Athen. *Deipn.* 556f" refers implicitly to the edition of the text published by G. Kaibel[6]. Since a digital version derived from that edition is accessible via CTS webservice, we can refer to notional concept, editions and passages of this edition by using the correspondent CTS URNs. In particular, the CTS URN for the text passage in the example allows us to refer with high precision and granularity to the exact words drawn from Athenaeus that Jacoby attributed to Istrus. The capability of addressing this aspect is crucial since, as illustrated above, the beginning and end of fragments is a matter of debate among scholars.

[6] Kaibel, G. *Athenaei Naucratitae Deipnosophistarum libri xv*. I-III. Leipzig 1887-1890.

7. Architecture

Our work is firmly based upon the CITE architecture which is being developed at Harvard's Centre for Hellenic Studies (CHS). CITE and in particular the CTS protocol can be used to provide the layer of services upon which fourth generation digital libraries can be built [21]. Indeed CTS specifies a network protocol to access XML encoded texts with high granularity allowing us to point to any of the hierarchical levels defined for a text (books, sections, paragraphs, etc.). In terms of the representation of fragmentary texts, CTS services are suitable for hosting collections of texts that constitute the witnesses for fragmentary texts[7]. Since the CTS is built upon FRBR with slight modifications and FRBR is conceived as a formal ontology and currently being adapted to CIDOC-CRM, it was possible to easily integrate them inside the proposed ontology (as shown in section 5.1) and then in the overall architecture [22].

We do not want to limit the resources used, however, to only those available through CTS services and referable via CTS URNs (the set of identifiers used by the CTS protocol). Given the number of critical editions made available by Google Books and given the existence of other digital libraries and digital editions of texts not currently exposing a CTS-compliant interface, we want to be able to refer to any resource published on the Web provided that it at least has a URI. Collections such as the Perseus Digital Library or the Suda on Line already provide digital editions of works that are also sources of fragments. Critical editions of fragments are currently available as page images on Google Books as well [23,24], even if there are still issues with the OCR of ancient Greek script and even if those editions are out of date (i.e., not the current reference editions).

The proposed ontology is currently being implemented using OWL for the sake of integration with the external ontologies implied. The produced RDF triples representing fragmentary texts will be stored in a knowledge base practically implemented as an RDF store. This knowledge base is expected to have a SPARQL end point to allow the triples contained in it to be used also by other user communities to describe Web resources according to the Linked Data model [25]. Practically this means that whenever possible this knowledge base will contain statements of equivalence between already existing URIs, for instance a Wikipedia page about an ancient author and its CTS URN.

Other ways to further populate our knowledge base of semantic data are:

- RDFization of the Perseus' FRBR catalogue [26], providing a huge amount of

[7] The texts produced up to now are accessible through the CTS protocol at <http://cts3fhg.appspot.com/>.

catalogue records and links to existing resources for modern editions of Greek and Latin works;
- crawling of CTS repositories and conversion to RDF of XML web service responses;
- use of RDF records about resources contained in other digital libraries[8].

Regarding user access to the created knowledge base, one suitable solution may be to create an interface that allows users to browse, create ontologically encoded semantic data and read ancient texts all in the same environment. This solution is currently being pursued by Philospace, a desktop application providing access to semantic resources produced in the framework of the Discovery project [27]. Philospace relies upon DBIN [28], an application written in JAVA aimed at enabling the creation of Semantic Web communities that allows developers to create specialized domain applications through a plugin mechanism called "brainlets". As further development of this work, we plan to create a DBIN brainlet for the domain of Classical Philology, a semantic environment allowing scholars to browse and create semantic annotations about ancient texts. Provided that a CTS client is easily pluggable into DBIN, a suitable feature of this brainlet will be the capability of displaying to the reader text passages when available through CTS services.

8. Related Work

The greatest efforts in applying ontologies for scholarly purposes are being conducted in the field of Philosophy where representing ideas and interpretations is a task of primary importance. PhiloSurfical [29] provides an ontological formalization to represent philosophical ideas and interpretations, along with a tool to browse them. The above mentioned Discovery project is heavily exploiting ontologies in order to allow end users to express annotations and interpretations on a semantic digital library of texts [30]. The ontologies for texts created in the framework of this project, however, are tailored specifically to modern texts and editions that from a philological point of view, differ substantially from the ancient ones.

As far as concerns the field of classical studies, Semantic Web related technologies have recently received renewed interest. Discussion groups such as *Graph of Ancient World Data* [31] clearly show the interest of this community in Semantic Web related technologies, in particular the communities of archaeologists [32].

[8] An RDF description of the resources the Perseus Digital Library contains along with all the source code can be downloaded by users and developers at <http://sourceforge.net/projects/perseus-hopper>.

One of the main reasons for this is the increasing need for interoperability, in order to access the amount of data that different projects have produced and distributed with different formats up to now. In the same direction, the Text Encoding Initiative (TEI) has created a special interest group [33] for these topics, focusing on mappings between the TEI encoding scheme and the CIDOC-CRM data model [34].

9. Conclusion

This paper examined the main issues encountered during the design of an ontology to represent fragmentary texts and provided a theoretical and architectural foundation for the digital representation of fragment editions. As a result of formalizing the reality of fragmentary texts through ontologies, their hermeneutical nature as scholars' interpretative acts emerged. Furthermore, the results obtained during the knowledge domain analysis through applying a supervised method on a small corpus of texts written by philologists were encouraging. In particular, they demonstrated the importance of basing the ontology design on evidences that spontaneously emerge from a text corpus.

Acknowledgements

Grants from the Andrew W. Mellon Foundation ("The Cybereditions Project") and the NEH in conjunction with the IMLS ("Scalable Named Entity Identification in Classical Studies") provided support for this work.

Notes and References

[1] American Council of Learned Societies. Our Cultural Commonwealth: The final report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences. 2006. Available at http://www.acls.org/cyberinfrastructure/ (March 2009).

[2] BERTI, M. et al. Collecting Fragmentary Authors in a Digital Library. To appear in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, Austin, Texas: ACM Digital Library, 2009 (preprint). Available at http://www.perseus.tufts.edu/~ababeu/JCDL09_sp.pdf (April 2009).

[3] DOMS, A. and SCHROEDER M. GoPubMed: exploring PubMed with the Gene Ontology. Nucl. Acids Res. 33, July 2005, pp. 783-786.

[4] KUMMER, R. Integrating Data from The Perseus Project and Arachne using the CIDOC CRM. 2007. Available at http://cidoc.ics.forth.gr/workshops/heraklion_

october_2006/kummer_presentation.pdf (March 2009).

[5] D'ANDREA, A. and NICCOLUCCI F. Mapping, Embedding and Extending: Pathways to Semantic Interoperability, the Case of Numismatic Collections. In *Fifth European Semantic Web Conference Workshop: SIEDL 2008-Semantic Interoperability in the European Digital Library*, pp. 63-76. Available at http://image.ntua.gr/swamm2006/SIEDLproceedings.pdf#page=69 (April 2009).

[6] BUITELAAR, P. *Ontology Learning from Text: Methods, Evaluation and Applications.* [Amsterdam; Washington DC]: IOS Press, 2005.

[7] Infomap NLP Software. Available at http://infomap-nlp.sourceforge.net/ (April 2009).

[8] TreeTagger. Available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (April 2009).

[9] BUCKINGHAM SHUM, S. et al. ScholOnto: an Ontology-based Digital Library Server for Research Documents and Discourse. International Journal on Digital Libraries 3, 2000, pp. 237-248.

[10] Bibliographic Ontology Specification. Available at http://bibliontology.com/ (April 2009).

[11] Functional Requirements for Bibliographic Records (FRBR). Available at http://www.ifla.org/VII/s13/frbr/ (April 2009).

[12] The CIDOC CRM - Introduction to FRBRoo. Available at http://cidoc.ics.forth.gr/frbr_inro.html (April 2009).

[13] The Canonical Text Services Protocol. Available at http://chs75.harvard.edu/projects/diginc/techpub/cts (April 2009).

[14] JOUSSELME, A.L., MAUPIN, P. and BOSSE E. Uncertainty in a Situation Analysis Perspective. In *Proceedings of the Sixth International Conference on Information Fusion*, Vol. 2, pp. 1207-1214, 2003. Available at http://www.ieeexplore.ieee.org/iel5/8886/28065/01255342.pdf.

[15] W3C Uncertainty Reasoning for the World Wide Web XG. Available at http://www.w3.org/2005/Incubator/urw3/wiki/FrontPage (April 2009).

[16] PRESUTTI, V. and GANGEMI A. Content Ontology Design Patterns as Practical Building Blocks for Web Ontologies. In *Conceptual Modeling - ER 2008*, pp. 128-141, 2008. Available at http://dx.doi.org/10.1007/978-3-540-87877-3_11 (April 2009).

[17] ROBERTSON, B. Exploring Historical RDF with Heml. Digital Humanities Quarterly *3*, Changing the Center of Gravity: Transforming Classical Studies Through CyberinfrastructureWinter 2009. Available at http://www.digitalhumanities.org/dhq/vol/003/1/000026.html (March 2009).

[18] BOSCHETTI, F. Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses onto Reference Text. In *Proceedings of the Corpus Linguistics Conference (CL2007)*, 2007. Available at http://

ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf (January 2009).

[19] ROMANELLO, M. et al. When Printed Hypertexts Go Digital: an Index-driven Approach to the Automatic Markup of Text Quotations. Poster to appear in *Hypertext 2009: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, Torino, Italy: ACM Digital Library, 2009. Available at http://www.perseus.tufts.edu/~ababeu/ht159-romanello.pdf (April 2009).

[20] Domain Name (Space) Identifiers (DNID). Available at http://www.dnid-community.org/ (April 2009).

[21] STEWART, G; et al. A New Generation of Textual Corpora: Mining Corpora from Very Large Collections. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pp. 356-365, ACM, 2007. Available at http://dx.doi.org/10.1145/1255175.1255247 (April 2009).

[22] SMITH, N. Citation in Classical Studies. Digital Humanities Quarterly *3*, Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure, Winter 2009. Available at http://www.digitalhumanities.org/dhq/vol/003/1/000028.html (March 2009).

[23] Die Fragmente der Vorsokratiker. Available at http://books.google.com/books?id=xQYrAAAAMAAJ (April 2009).

[24] Fragmenta Comicorum Græcorum. Available at http://books.google.com/books?id=SJQCAAAAQAAJ (April 2009).

[25] BIZER, C. et al. Linked Data on the Web (LDOW2008). In *Proceeding of the 17th international conference on World Wide Web*, pp. 1265-1266, [Beijing, China]: ACM, 2008. Available at http://portal.acm.org/citation.cfm?doid=1367497.1367760 (April 2009).

[26] BABEU, A. Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience. 2007. Available at http://www.perseus.tufts.edu/~ababeu/Perseus FRBRExperiment.pdf (April 2009).

[27] NUCCI, M. et al. Semantic Web Powered Distributed Digital Library System. In *ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing held in Toronto, Canada 25-27 June 2008* / Edited by: Leslie Chan and Susanna Mornati, pp. 130-139, 2008. Available at http://elpub.scix.net/cgi-bin/works/Show?130_elpub2008 (April 2009).

[28] TUMMARELLO, G. et al. Enabling Semantic Web Communities with DBin: An Overview. In *The Semantic Web - ISWC 2006*, pp. 943-950, 2006. Available at http://dx.doi.org/10.1007/11926078_69 (April 2009).

[29] PASIN, M. et al. Capturing Knowledge about Philosophy. In *Proceedings of the 4th international conference on Knowledge capture*, 47-54, [Whistler, BC, Canada]: ACM, 2007. Available at http://portal.acm.org/citation.cfm?id=1298406.1298416 (January 2009).

[30] NUCCI, M. et al. Talia: A Framework for Philosophy Scholars. In *Proceedings of Semantic Web Applications and Perspective*, [Bari, Italy] 2007. Available at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-314/39.pdf (January 2009).

[31] Graph of Ancient World Data. Available at http://groups.google.com/group/gawd/ (April 2009).

[32] BINDING, C. et al. Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM. In *Research and Advanced Technology for Digital Libraries*, pp. 280-290, 2008. Available at http://dx.doi.org/10.1007/978-3-540-87599-4_30 (April 2009).

[33] TEI Ontology SIG WIKI. Available at http://wiki.tei-c.org/index.php/Main_Page (April 2009).

[34] EIDE, O. and CHRISTIAN E. TEI, CIDOC-CRM and a Possible Interface Between the Two. Digital Humanities, pp. 62-64, 2006.

June 2009
Printed on demand
by *"Nuova Cultura"*
www.nuovacultura.it

Book orders: ordini@nuovacultura.it