Collecting Fragmentary Authors in a Digital Library

Monica Berti, Matteo Romanello, Alison Babeu, and Gregory Crane
The Perseus Project
Medford, MA, USA
monica.berti@tufts.edu, matteo.romanello@tufts.edu, alison.jones@tufts.edu,
gregory.crane@tufts.edu

ABSTRACT

This paper discusses new work to represent, in a digital library of classical sources, authors whose works themselves are lost and who survive only where surviving authors quote, paraphrase or allude to them. It describes initial works from a digital collection of such fragmentary authors designed not only to capture but to extend the ontologies that traditional scholarship has developed over generations: the aim is representing every nuance of print conventions while using the capabilities of digital libraries to extend our ability to identify fragments, to represent what we have identified, and to render the results of that work intellectually and physically more accessible than was possible in print culture.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*collection, dissemination, standards*

General Terms

Documentation, Performance, Standardization, Languages.

Keywords

Digital Libraries, Fragmentary Authors, Greek Fragmentary Historians, XML, TEI P5 Guidelines.

1. INTRODUCTION

A fragmentary author is an author whose works 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. Greek and Latin literature is rich in fragments covering almost every genre, from epics and poetry to oratory and historiography.

Modern scholars have at their disposal many collections of fragmentary authors built thanks to the great work of philologists from the Renaissance onwards, who have reconstructed works and personalities otherwise lost and forgotten. The importance of fragmentary texts for our knowledge of ancient literature is evident also from a numerical point of view, as it is shown by the data we have drawn from a quantitative analysis on the Thesaurus Linguae Graecae (TLG-E), which is the reference digital library of Greek literature: for the period between the 8th century B.C and the 3rd century A.D. included, 59% of the authors is preserved only in fragments, 12% is known both from entirely preserved works and fragmentary ones, while 29% is represented by surviving works.

New technologies have increasingly offered computerized tools that have been customized for collecting and digitizing ancient sources, leading to the foundation of digital collections of all major classical sources. [5,6] These comprehensive tools allow us to create a new generation of fragmentary corpora that express more scholarly information, are far easier to maintain, and are much more usable than the collections that were possible in print culture[1].

Print collections of fragments contain excerpts from many different sources and are thus paper representations of hypertexts. A single collection may contain excerpts from hundreds of separate editions and serious scholars need to be able to consult current scholarship on any of these cited authors. Now that the source editions from which fragments are drawn are becoming available in digital form, we can construct editions that are truly hypertextual, including not only excerpts but links to the scholarly sources from which those excerpts are drawn. Fragments exist as text quotations embedded in works written by other authors. While duplicating the text of fragmentary quotations in a printed context is often unavoidable, the hypertextual nature of these kinds of relationships among texts can be more properly represented in a digital library. When designing a digital library, therefore, the representation of fragments should seek to avoid the problem of duplicate records. This is particularly important when the texts of a digital library will become data that is analyzed computationally either by algorithms or scholars, such as with statistical or text mining analysis of ancient texts. The way a text corpus is built affects both the kinds of questions that can be asked and the validity of the answers obtained.

Given the great amount and variety of fragmentary texts, we have focused our research on Greek fragmentary historians, because they are in many respects representative for

[1]On fragmentary texts in traditional scholarship, see Most, G. W., *Collecting Fragments.* Göttingen, 1997.

JCDL'09 June 15–19, 2009, Austin, Texas, USA

building a digital collection of fragmentary authors, and also because in the 19th and 20th centuries monumental collections of those authors have been edited, posing fundamental questions on gathering and editing fragments: the emerging digital libraries of classical sources challenge us to rethink these questions and the characteristics of a fragmentary text. The work done so far has allowed us to identify the requirements pertaining to fragmentary texts, and now we are focusing on developing an ontology to represent fragments in a digital library.

2. FRAGMENT AND WITNESS

Collecting fragments means first of all extracting quotations from their context. 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. Digital editions of fragments should consist not of isolated quotations but of pointers to the original contexts from which the editor has excerpted the fragments. While editors should be able to define the precise chunks of text that they feel to be relevant and to be able to annotate these texts in various ways (e.g., distinguishing what they consider to be paraphrase from direct quotation), such fragments should also be dynamically linked to their original contexts and to up-to-date contextualizing information.

The way collections of fragments have been included (and then treated) in a reference digital library for scholars such as the TLG can lead to some data inconsistencies. For example, consider a textual search aimed at finding all occurrences of Ἐριχθόνιος ὁ Ἡφαίστου in the entire corpus of Greek literature provided by the TLG. This search returns five separate results [2], but in fact, Erichthonius is actually only mentioned once by Harpocration inside an entry of the *Lexicon in Decem Oratores Atticos*, a text which is the witness to three fragments of Greek historical authors (Hellanicus, Androtion, and Istrus). The TLG lists separate results for each of these fragmentary historians (two for Hellanicus) as well as one for Harpocration.

From a quantitative point of view, this search gives the wrong impression that the expression searched for appears five separate times in Greek literature. If the aim of the search, however, was the qualitative analysis of contexts where mentions of Ἐριχθόνιος ὁ Ἡφαίστου appear, the duplication of records would become essential, since each passage bears a different quotation context. In other words, in order to have consistent data for different types of analysis it is important to be able to concretely specify the scope of textual search to be performed (e.g., which texts you want to include in your search) and to show the search results in a way that more accurately reflects the corpus of texts. As an example of further analysis to be conducted on a corpus of fragmentary texts it is worth mentioning the eAQUA project[3], which aims at analyzing Greek historical fragmentary texts by exploiting Text Mining techniques. [4]

3. COLLECTING FRAGMENTS

Fragments are texts embedded in works written by other authors. Consider, for example, a fragment of the Greek historian Ion of Chios that is found in a passage of Plutarch's *Lives*:

> *And Ion actually mentions the phrase by which, more than by anything else, Cimon prevailed upon the Athenians, exhorting them "not to suffer Hellas to be crippled, nor their city to be robbed of its yoke-fellow."* (Plut. *Cim.* 16.8, trans. Perrin)

A digital corpus of fragmentary authors (including many quotations such as this) should be characterized by dynamic access to a wide range of primary and secondary sources, providing at least the following fundamental functions:

Quotation as Machine Actionable Link. Ion's fragment should be linked to the full text of Plutarch, *Life of Cimon*, chapter 16, from which the excerpt has been drawn. Some work in this area has been reported by [10].

Alignment of Citation Schemes. This passage should be identified in all other digital editions of Plutarch's *Life of Cimon*. Given that citation schemes may differ, the system should collate multiple editions in order to align multiple citation schemes. For recent work on text collation, see [7,9].

Fragment as Search Query. The excerpted text of Ion should be searchable to find the corresponding passage in all on-line editions, even when they have not been carefully transcribed, but are available only as machine generated OCR texts. In the latter case, the aim is generating links between the fragment and the page image of multiple editions of the same passage. For some interesting work in detecting matching text fragments automatically between different types of documents, please see [11,12].

Dynamic Collation. Critical editions of the same fragment and of its witness should be collated to identify and prioritize differences among them, such as in particular textual variants of the manuscripts and different editors' choices and criteria. [3]

Secondary and Tertiary Sources. A digital corpus of fragmentary authors should provide links to secondary and tertiary sources, identifying passages in papers, monographs, commentaries, and other evidence related to the fragment and/or the context from which the fragment has been excerpted. In addition, document clustering and summarization should be used to partition and classify these passages into meaningful groups and categories identifying common traits, while text mining should derive other significant information from these texts. Some preliminary work in these areas with humanities texts has been presented in [8,13].

Collecting a digital library of fragmentary authors aims also at establishing a scalable, and to the extent possible, automated workflow to deal with multiple editions of the same work. Given that fragmentary texts are one of the most complex and challenging kinds of sources, using them as a test bed allows us to examine the cyberinfrastructure which is being defined to deal with digital editions of classical sources. [1]

The CITE architecture, developed at the Harvard University's Center for Hellenic Studies (CHS), is part of the developing cyberinfrastructure and defines protocols supporting the creation of networked digital libraries. These protocols support the extraction of chunks of text from literary works and the description of relationships among texts and other digital objects, such as raw data and images. The Canonical Texts Services (CTS) protocol allows for the representation of works preserved intact from the past (i.e., not

[2]Hellan., *FGrHist* 4 F 39 = *FGrHist* 323a F 2; Androt., *FHG* I p. 371; Harpocr., Lex. s.v. Παναθήναια; Ist., *FHG* I p. 419.

[3]`http://www.eaqua.net/`

in fragments), since it focuses on the notion of works and editions of a canonical work. [14] Given that fragmentary texts are embedded in texts written by other authors, a CTS repository of witnesses has been created, and the text of the fragments is extracted from it using requests compliant to the protocol. The repository hosts digital editions for every canonical work, witnessing a certain number of fragments.

Up to now the Perseus project has produced new digital editions of a set of authors who preserved fragments in their works: Athenaeus' *Deipnosophistae* (ed. Kaibel), Harpocration's *Lexicon in Decem Oratores Atticos* (ed. Dindorf), and Photius' *Lexicon* (eds. Porson and Theodoridis). These texts have been chosen for their literary importance and for the fact that they include an impressive amount of fragments (of various genres). Moreover, digitizing Harpocration's and Photius' *lexica* – which contain approximatively 15,000 lemmatized entries – is an important contribution toward increasing the rather small number of existing digital editions of lexicographic texts.

The choices made during the creation of TEI P5 editions of this set of authors have been in a certain measure determined by the overall workflow which is being tested on fragmentary texts. The adopted approach consists of leveraging one deep structured edition of a text to improve other existing editions that are currently available, like OCR transcriptions from page images. Specifically, the TEI P5 editions that have been produced are being used as best approximations to ground truth to correct the OCR output of several editions of the same text that have been made electronically available thanks to mass digitization projects. The encoding rationale is both to preserve and keep distinct the physical and logical layers of the texts. In order to use the text of the produced editions as supporting material in the phases of OCR training and post-processing, however, it has been necessary to accurately encode some information about the physical appearance of the text. Encoding physical features of a text, such as page and line divisions, for instance, has allowed us to determine the correct text corresponding to a given page image during the training process, and thus it has been possible to avoid typing it manually. Furthermore, information about line divisions and numbering will be even more useful when conjectures from the critical apparatus will be linked to the text passage referred to.

4. REPRESENTING FRAGMENTS

A digital representation of the characteristics of a text consists not just of a mere reproductive and mechanical process, but constitutes an interpretative act. Accordingly, encoding fragments is first of all the result of interpreting them, such as creating metadata and meta-information about them. Conceiving a digital edition of fragments implies finding new digital paradigms and solutions to express information about texts that are already present in printed critical editions and encoded by using editorial and presentational features. Working on a digital edition, traditional tools and resources used by scholars such as canonical references, tables of concordances and indexes need to be converted into machine actionable contents. One current goal of the research described in this paper is to develop an ontology appropriate for representing textual features of the fragments and interpretations of these features, providing at least the data described in the following paragraphs.

Numbering and ordering fragments may vary – even substantially – from one critical edition to another. The result is that the same fragment can have different numbers according to different editions. The order chosen by the editor to arrange the fragments in a given sequence is often intentional since it assumes a hypothetical reconstruction of the lost original narrative sequence. The correspondence between fragments having different numbers in different editions, which is usually registered in the table of concordances of a printed edition, needs to be translated into machine actionable content. Given that tables of concordances align multiple bibliographical references to the same textual object (e.g., the fragment), they can be digitally represented by using the Reference Index protocol and services from the above mentioned CITE architecture.[4] In fact, Reference Indexes encode index entries in a machine actionable way, expressing them as mappings between permanent references and digital objects (or even between such references and raw data). Further, the RefIndex protocol can be suitably used to align also multiple citation schemes of different editions of the same work which was identified above as a key feature for the new digital collection being produced.

Fragments are classifiable according to multiple criteria ranging from internal to external factors, such as contents, authors, works, literary genres, and subjects. Furthermore, a fragment can be a *testimonium* (i.e., a fragment consisting of evidence on the author's life and works, providing biographical and bibliographical information about a fragmentary author) and a *fragmentum* (i.e., a fragment consisting of a quotation, paraphrasis, or summary of an identified or unidentified work written by the fragmentary author). An ontology for representing fragments should provide a taxonomy to classify a text fragment at least as *testimonium* or *fragmentum*, and to specify different kinds of fragments (i.e., historical fragments, poetical fragments, philosophical fragments, etc.).

The ontology should also provide a mechanism to mark the beginning and the end of a fragment in the text of its witness according to different editions. The length of a fragment depends on editor's criteria, because ancient writers never employ quotations marks, and therefore identifying precisely the extent of a fragment may be a difficult task, even when the text of a lost work is cited verbatim. For all these reasons, marking up fragment boundaries directly in the text of witnesses is not a suitable solution. Instead, different interpretations about the beginning and the end of fragments are translated into a collection of pointers encoded as standoff markup. The digital analogues of canonical references[5] which are being used in this project are CTS URNs, since they support pointing to text chunks inside different hierarchical levels of the text structure (i.e., when using CTS URNs it is possible to cite a paragraph, a chapter or a book inside a digital edition of a work). The granularity of this kind of identifier as defined by the CTS protocol allows one to identify and then retrieve specific chunks of text over the Web in a machine actionable way, where the smallest textual unit to be addressed is the character. [15] To sum up, given the CTS repository containing different TEI XML editions of fragment witnesses (see section 3), the text of each fragment – as established by the different editions – is at the

[4] `http://katoptron.holycross.edu/cocoon/ tdig-inc/techpub/indexing`

[5] Canonical references (e.g., Plut. *Cim.* 16.8) are traditionally adopted by Classicists to refer to ancient texts.

same time linked to its context and extracted for visualization by using a collection of pointers to the texts.

Since in a digital collection we need to refer to fragments as discrete objects, a comprehensive catalog of unique fragment identifiers is being built. On the other hand, the Canon of Greek Literary Works developed at CHS already provides unique identifiers for the so called "canonical" authors and works. Some preliminary work in creating catalog and authority records for fragmentary authors has already been conducted by the Perseus Digital Library. [2] Each catalog entry will have associated metadata about fragmentary authors and their works. In terms of personal names, the set of tags developed by TEI P5 (Names, Dates, People, and Places: chapter 13) could be useful for covering a wide range of information on fragmentary authors, such as their names, literary and geographical epithets (e.g., Plato Comicus, Hecataeus Abderita, or Hecataeus Milesius), and chronology. A fragment may also bear the title of the work from which it has been extracted. This kind of information should be encoded, even if attributing a fragment to a work and managing titles of ancient works can be challenging: in most cases, witnesses do not cite the title of the work from which they have drawn the fragment; moreover, in ancient sources the title of a work may be attested with more or less significant variants. For all these reasons, we are working also on the automatic extraction of information from the paratext of modern editions, and in particular from the *index auctorum*, where editors usually give information about ancient authors and works cited in the text.

5. CONCLUSIONS

Building a digital corpus of fragmentary authors can contribute in various ways to the making of a full, dynamic, and hypertextual digital library of Classical sources. The first aim of this enterprise is collecting and reconstructing an invaluable cultural heritage, preserving authors and works otherwise lost, while gathering editions and scholarly commentaries often scattered in different libraries, which may be remote and difficult to access. Secondly, working with fragments means moving incessantly across primary and secondary sources, connecting and interpreting them both synchronically and diachronically, according to many analytical approaches and perspectives, ranging from critical evaluations to literary classifications and historical reconstructions. Thirdly, extracting a corpus of fragmentary authors within a digital collection of Classics may constitute a good practice for managing ancient sources in a digital environment, and refining techniques for their representation. Conversely, envisioning a collection of fragmentary texts in a digital library means working in a wider context, going well beyond the specificity of the field and the limitations imposed by most of the traditional printed editions.

6. ACKNOWLEDGMENTS

This work was supported by grants from the National Endowment for the Humanities, the Mellon Foundation and the Joint Information Systems Committee.

7. 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.

[2] A. Babeu. Building a "FRBR-Inspired" catalog: The perseus digital library experience. Technical report, 2007.

[3] F. Boschetti. 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.

[4] M. Buechler, G. Heyer, and S. Grunder. eAQUA - bringing modern text mining approaches to two thousand years old ancient texts. 2008.

[5] G. Crane. Classics and the computer: An end of the history. In J. U. Susan Schreibman, Ray Siemens, editor, *A Companion to Digital Humanities*, pages 46–55, Oxford, 2004. Blackwell.

[6] G. Crane. What do you do with a million books? *D-Lib Magazine*, 12(3), Mar. 2006.

[7] R. V. den Branden. A modest proposal. analysis of specific needs with reference to collation in electronic editions. In *Digital Humanities 2008*, Oulu, Finland, June 2008.

[8] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 213–222, Lisbon, Portugal, 2007. ACM.

[9] S. Feng and R. Manmatha. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 118, 109. ACM Press, 2006.

[10] O. Kolak and B. N. Schilit. Generating links by mining quotations. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 117–126, Pittsburgh, PA, USA, 2008. ACM.

[11] J. Lee. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 479, 472. Association for Computational Linguistics, 2007.

[12] D. Metzler, S. Dumais, and C. Meek. *Similarity Measures for Short Segments of Text*, pages 16–27. 2007.

[13] D. Mimno and A. Mccallum. Organizing the OCA: learning faceted subjects from a library of digital books. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 385, 376. ACM Press, 2007.

[14] M. Romanello. A semantic linking system for canonical references to electronic corpora. In P. Zemanek, editor, *Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages, Prague 16-17 November 2007*, Prague. Special special issue of Yearbook Chatreššar.

[15] N. Smith. Citation in classical studies. *Digital Humanities Quarterly*, 3(1), 2009.