The Digital Critical Edition of Fragments Theoretical Problems and Technical Solutions

Matteo Romanello

The process of creating a digital critical edition of fragmentary texts is also a valuable opportunity to look at the nature of such texts under a new light. This paper[1] contains some reflections about how the nature of fragments can be more appropriately represented in a digital critical edition and what challenges this task poses.

1. Introduction

Our knowledge of Greek literature relies heavily on fragments. More than a half of ancient Greek authors are preserved only in fragments, whereas less than a third are known by means of works that were entirely preserved[2]. These figures remind us of the key role played by fragments with regard to our knowledge and understanding of that literature. While the examples of literary fragments described in this paper are all drawn from Greek literature, the proposed solution to represent fragments in a digital environment might apply to similar texts in different domains.

By "fragment" is meant here a portion of a literary work that has survived only in a quotation by a different author or in another work[3]. Such literary fragments are distinct from material fragments such as papyrus scraps or *ostraka* (i.e. vase fragments)[4]. The philological term to refer to a work as evidence of a lost work is "witness": for instance Athen. *Deipn.* 556f is the witness of the fragment FGrHist 334 F 10[5].

The paper presents some considerations about the nature of fragmentary texts that arose during the process of creating a digital collection of ancient fragments as described in Berti et al. (2009). The project aimed to provide the Perseus Digital Library[6] with a collection of fragmentary texts starting with the fragments of ancient Greek historians. In this paper some theoretical problems concerning fragmentary texts are presented and discussed along with their possible technical solution.

[1] This work is part of the PhiloGrid project, <http://www.perseus.tufts.edu/hopper/grants#philogrid>.

[2] For the precise figures see Berti et al. (2009: 1).

[3] The goal of this paper is not to define or redefine the concept of fragment nor to define what a digital edition is. On the contrary its aims is to discuss some aspects concerning the nature of fragments that should be carefully considered when creating a digital edition.

[4] However, it is not hard for a philologist to find examples where these boundaries are blurred if not absent.

[5] 'FGrHist' is common abbreviation for the edition by Jacoby (1958).

[6] http://www.perseus.tufts.edu/hopper/

2. Related Work

The idea of digital critical edition assumed in this paper is well described in Bodard-Garcés (2009) as Open Source Critical Edition (OSCE). Providing multiple editions of the same text and a digital analogue of the print critical apparatus are some of the features that should characterise "post-incunabular digital libraries", as they were called by G. Crane (2006), and are still missing from existing digital collections of ancient texts, including fragmentary texts.

As far as concerns the fragments of Greek historians in particular there is an edition published online by Brill and entitled *New Jacoby*[7]. The *New Jacoby* is a publication aiming to complete and update the monumental *Fragmente der Griechischen Historiker (The Fragments of the Greek Historians)* by F. Jacoby (1958). Although it is accessible online via annual subscription, this is not sufficient to make it a digital edition according to our definition. Indeed, it is one of those cases of digital edition that do not go beyond "the simulation of paper", to borrow T. Nelson's (2007) words. The *Thesaurus Linguae Graecae* (TLG)[8], the reference electronic corpus of Greek literature, includes in its collection texts by fragmentary authors. However, not only it is not *open* but as an edition it can be hardly considered *critical* given its lack of any form of critical apparatus.

3. The Nature of Fragments

To better understand how Classics scholars perceive fragmentary texts a preliminary experiment was carried out by means of a computational technique called semantic space analysis[9]. This technique is based on the hypothesis for word distribution defined by Rubenstein and Goodenough (1965) according to which words with similar meaning occur in similar contexts.

We took the full text of 170 research articles written in English and selected from the JSTOR archive. They are all related to fragments from Latin and Greek literature and belonging to several genres (e.g., epic, tragic, comic, and historical). The full text of those articles was then processed using a piece of software called Infomap[10] to obtain the data necessary to compute the semantic spaces for a set of words that were considered relevant from a philological perspective. The identified semantic spaces were then plotted on a bi-dimensional graph where they are represented as cluster of words (Figure 1). The obtained graph shows how words cluster together on the basis of their semantic relation.

[7] http://www.brill.nl/brillsnewjacoby

[8] http://www.tlg.uci.edu/

[9] The experiment is extensively described in Romanello et al. (2009: 5-6). Moreover, Boschetti (2010: 57-82) provides an example of semantic space analysis applied to Greek literature as corpus.

[10] http://infomap-nlp.sourceforge.net/

Looking at the chart produced, it is possible to identify three main semantic clusters of words corresponding to as many different semantic spaces and distinguished in the chart by means of distinct colours:

1. **position:** the cluster in the top right corner (in black) contains those words that provide the coordinates to identify passages within the text (e.g. "top, bottom, left, right, beginning, end, line, margin"). Those words play a crucial role because they allow scholars to refer to the text passage that are being discussed. Moreover, they show how much the determination of the actual boundaries of fragments within the citing text is an open problem.

2. **text criticism:** this cluster, situated in the top left corner and displayed in green, contains the technical terms of philology (e.g. "editor, apparatus, scribe, copyist, manuscript, reading, emendation, conjecture").

3. **interpretation:** the words contained in this third cluster (displayed in red) are essentially concerned with the act of interpreting texts ("purpose, assumption, interpretation, supposition") and with the different degrees of confidence about the fact that an interpretation is true ("authenticity, uncertainty, possibility"). The word "fragment" sits within this cluster and almost in the middle of the whole chart meaning that it is slightly attracted towards the "text criticism" cluster as it was labelled before.

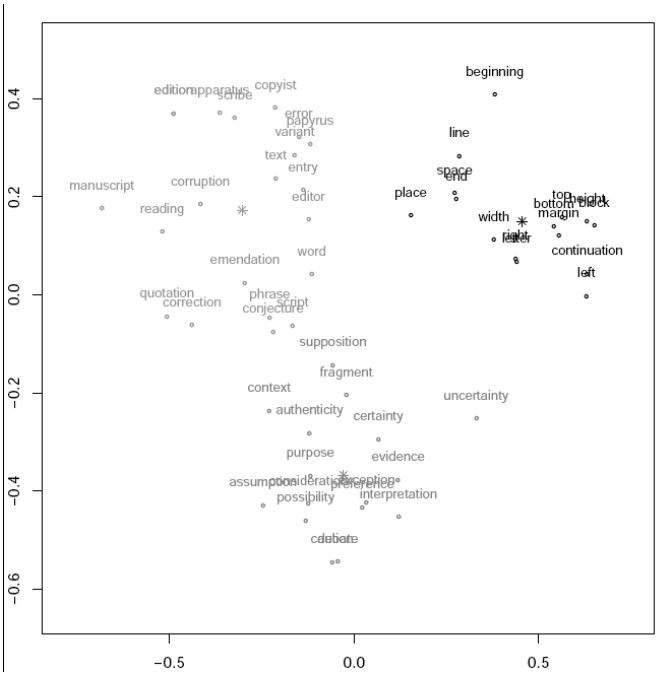

Figure 1: Cluster chart of relevant terms related to "fragment".

The experiment shows how the words used by scholars in a sample of journal papers related to fragmentary texts can be divided into three distinct semantic spaces. The word 'fragment', which is of crucial importance for our analysis, co-occurs together with words related to interpretation and uncertainty. Arguably every text is the result of an interpretation consisting in the choices made by the editor of that text. However in the case of fragmentary authors, the fact of not having extant manuscripts of the lost works makes such interpretation even more crucial because hard to verify.

A digital edition of fragments should adequately represent the multiplicity of interpretations implied by the identification of a literary fragment. The author and the work a fragment is attributed to are an example of the interpretations that are in play when we talk about fragmentary texts. It is often difficult to establish with a great degree of certainty where precisely a fragment starts and where it ends. This is the case for example of those fragments that are a description of the content of the lost work as opposed to those that are direct evidence for it.

As interpretations their existence is quite difficult to capture. A text passage might be considered a fragment by a scholar but may not be so according to another. The abstract notion of a fragment is so tightly bundled with the edition it contains that it is not possible to think of a fragment without specifying which editions we are referring to. In order to spot a first difference about the nature of fragmentary texts as opposed to texts that were entirely preserved, let us pause briefly to consider how those texts are referred to. Indeed, Classics scholars may refer to the *incipit* of Homer's *Iliad* by using a reference such as "Hom. *Il.* I 1", that is without specifying to which edition of the text they are referring to. In contrast it is impossible to refer to a fragment without specifying the exact edition of the text that is being referred. In other words, a fragment is so intrinsically an interpretation that it is always necessary to declare whose interpretation we are basing our analysis upon.

Finally, by looking at the gigantic work of Jacoby on the library's bookshelf it is impossible not to see the flexibility that the electronic medium enables as a great benefit of having a digital edition of fragments. For instance cases where fragments are repeated over several print volumes because they need to be classified under multiple thematic categories can be easily avoided in a digital environment. Commentary and translation of a fragment, or even of its witness (i.e. the text bearing the fragment), will be linked to the fragment itself and displayed to the user side by side with the text, variants and conjectures of the editions of both fragment and witness.

4. Technical Solutions

The technical solutions proposed here aim mainly to address the characteristics of the nature of fragments described above.

The first aspect concerns how adequately to represent the fact that fragments are essentially citations or in other words their inherent hypertextual nature. Although it is technically possible to produce a digital edition of a fragment by embedding the

text of its witness, it seems not to be correct from a theoretical point of view. Indeed this operation leads to the duplication of words within our corpus, a situation that is well exemplified by the aforementioned TLG.

The TLG, which is currently the most comprehensive electronic corpus of Greek literature, contains both fragmentary texts and texts preserved to us by direct sources (e.g. manuscripts)[11]. From a technical point of view both kind of texts in the TLG are marked up in a structural markup language. If we consider the TLG as a corpus that we can query for instance to know the frequency of a given word, a first problem will suddenly appear. When several fragments attributed to different authors are witnessed by the same text passage, what happens in the TLG is that the text of the witness – and therefore its words – are duplicated several times. In turn this will result in inconsistent results from any quantitative analysis on this corpus.

A possible solution to this problem is to apply what in Computer Science is called 'transclusion'. The idea of transclusion was proposed by T. Nelson in his formulation of the Xanadu project[12]. Nelson (2007) defines transclusion as "the same content knowably in more than one place" of which transquotation, namely an "explicit quotation which remains connected to its origins" is a special case.

Going back to our fragments, when collected in a digital corpus they should be treated as cases of transclusions – if not transquotations – of their witnesses. The words of the fragment witness do not need to be duplicated by being embedded in the edition of the fragment since they can be transcluded from another resource. However, some differences apply and should be carefully represented. In particular the fact that the text of a fragment established by a given editor is a transquotation of the text of its witness in a given edition plus the conjectures proposed and/or the variant readings accepted by the editor of the fragment. A consequence of this approach is that variants and conjectures in the transcluded text should also be possibly displayed in a digital edition of a fragment.

Before moving on to the question of how to model interpretations in a digital environment, let us consider briefly how the references to our transcluded (or transquoted) texts will be expressed. Indeed, whereas in Xanadu's model what matters about texts is their version, when considering a digital edition the focus is on editions. A suitable reference should allow us to specify precisely which edition of a text is to be transcluded. In the framework we propose, the Canonical Text Services (CTS) protocol as defined by Smith (2009) will be used to serve this purpose.

[11] The TLG uses a best edition approach meaning that for each work it provides access to the text as it was established in one critical edition among those available. Instead the digital edition described in this paper aims to account for multiple editions and translations of the same text or fragment.

[12] T. Nelson (1983; 2007) defined several terms related to the Web that are now of common use, such as "docuverse", "hyperlink" and "transclusion". Di Iorio and Lumley (2009) have recently proposed an enhanced model for XML inclusions inspired by the Xanadu project and the concept of transclusion.

Let us examine briefly the main characteristics of this protocol and in which respects it fits our purposes.

The CTS is a protocol that allows us to make machine actionable references to texts. Such protocol is used to serve a collection of texts encoded in TEI/XML in a client-server architecture. For each text contained in a CTS instance two distinct hierarchies are specified: the hierarchy of versions and the hierarchy of citation levels that allow us to cite that text. One aspect particularly emphasised by this protocol is the concept of logical citation scheme as opposed to a physical one. Put it simply a logical citation scheme allows us to create reference to a text passage that can be resolved into several editions of that passage. The reference "Ath. *Deipn.* XV 694 e-f" to Athenaeus' *Deipnosophistae* can be resolved by a human reader into the text of that precise passage as established in Dindorf's, Kaibel's and Meineke's edition.

The CTS allows us to express the same mechanism within a persistent identifier, namely a CTS Uniform Resource Name (URN), that can be resolved into the text of different editions. The simple and at the same time essential feature of references expressed as CTS URNs is that their semantics are understandable also by non-human agents. As an example urn:cts:greekLit:tlg0008.tlg001:1.1 and urn:cts:greekLit:tlg 0008.tlg001.fhg01:1.1 are equivalent references to a text passage, whereas the former can be resolved into multiple editions and the latter is implying a specific edition.

Another consequence of using a CTS-based system is that several editions are aligned with one another on the basis of a common citation scheme which is edition-independent. Let us suppose we have a CTS instance containing the text of Athenaeus' *Deipnosophistae* according to Kaibel, Meineke and Dindorf, and that this CTS instance allows us to access that text by Casaubon's section. The text corresponding to the reference "Ath. *Deipn.* XV 694 e-f" in Kaibel's edition as a result will be automatically aligned to the text of the same passage respectively in Meineke's and Dindorf's edition as depicted in Fig. 2. Moreover, different CTS URNs pointing to words in the text can be used when editors (e.g. Müller and Jacoby) consider to be a fragment different spans of the same witness.

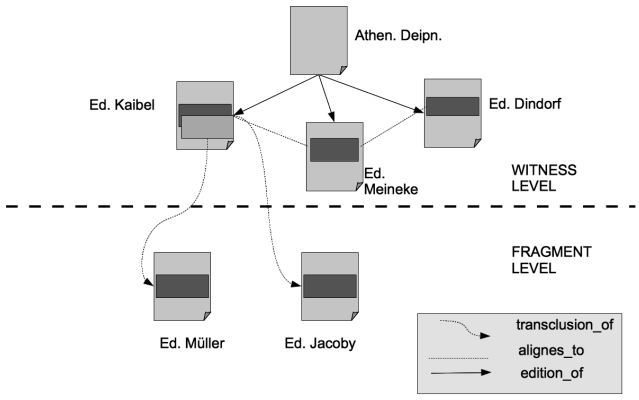

Figure 2: Diagram of the fragment-witness relationship in a multi edition scenario.

As far as concerns the interpretation aspect of fragments, the use of an ontology has been proposed by Romanello et al. (2009). Although this topic requires a level of detail which is beyond the scope of this paper, let us see briefly what is being proposed. Ontologies – in the Computer Science meaning of the term – are a formalism for knowledge representation. The main reason for this choice is adequately to represent in a digital environment what is ultimately carried by a critical edition: the formulation of an interpretation. The ontology we are planning to use is based on work that has been done to model philosophical thinking as described by Pasin (2009). The use of such an ontology should allow us specifically to address the interpretations about texts such as conjectures, attributions and identification of fragments. When considering fragments, there are several interpretations that scholar makes, some of which have been already mentioned: the identification of the boundaries of a fragment; the attribution of a fragment to an author and determining to which work it might have belonged; dating and classifying the fragment according to the genre, content, type, and so forth. The aim of such an ontology is to make explicit such statements that are often disguised within a print edition and record them according to the conceptual model they imply.

However, when creating a digital edition for a network environment a first essential task to be accomplished is the creation (if necessary) of persistent unique identifiers for the objects that will be represented[13]. In this case what is needed is a unique identifier for each fragment. Having identifiers for fragmentary authors will not suit the purpose since, as we have seen in the previous section, one of the attributions in play is that of a fragment to an author. Those identifiers will then be used to group together all the interpretations that the act of identifying a fragment implies.

As far as concerns the fragments of Greek historians, we are just now completing the creation of persistent unique identifiers for all the fragments identified by F. Jacoby and with a correspondence in Müller's edition[14]. For this purpose we have digitised and run the OCR on Jacoby's concordances as a way to acquire some basic knowledge about all the fragments that are mentioned here. After the scanning and the OCR, the text of the concordances was parsed[15] in order to extract information such as: a) the number of the fragment both in Müller's and Jacoby's editions; b) the author the two editors attributed the fragment to; c) correspondences between one or more fragment in Müller's to one or more fragments in Jacoby's edition. We hoped that the unique identifiers so obtained will be a starting point for future digital editions of these texts in a collaborative and networked environment.

[13] Current examples of initiatives aimed to define persistent identifiers are Pleiades <http://pleiades.stoa.org/> for ancient geographical places and the Canon of Greek Authors and Works by Harvard's Center for Hellenic Studies <http://chs75.chs.harvard.edu/registries/cts/chsCanon>.

[14] The results will be openly licensed and made available on a code repository accessible at <https://github.com/mromanello/Digital-Editions>.

[15] For further technical details about parsing print indexes see Romanello et al. (2009). On the automatic parsing of critical apparatuses see Boschetti (2009).

5. Conclusion and Further Work

The creation of digital critical editions of fragmentary texts has still a long way to go. A first attempt to tackle this problem has taken us just beyond the definition of the model that was described in this paper. However, the contribution of such editions to Classical philology is extremely valuable. Not only a digital edition will make easier to find and visualise information about fragments but it is also a more adequate representation of their hypertextual nature.

The technical solutions that have been proposed are based on open and distributed protocols and thus they allow for a collaborative and networked environment. Digital editions of fragmentary texts would remain impossible enterprises if digital editions of the fragment witnesses had to be replicated every time from scratch. Instead the enterprise will more likely succeed if we exploit the principle of transclusion and apply it to those editions that are now appearing, which are openly licensed and made available through a distributed protocol, such as the CTS[16].

Acknowledgements

This work was supported by grants from the National Endowment for the Humanities (http://www.neh.gov), the Mellon Foundation (http://www.mellon.org/) and the Joint Information Systems Committee (http://www.jisc.ac.uk/). I also want to gratefully acknowledge Gregory Crane, Federico Boschetti, Monica Berti, Gabriel Bodard and Pietro Liuzzo for their support and useful suggestions.

References

Berti, Monica et al., 2009, "Collecting fragmentary authors in a digital library". In: *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. Austin, TX, USA: ACM: 259-262. Available at: http://portal.acm.org/citation.cfm?id=1555400.1555442 [Accessed January 12, 2010].

Bodard, Gabriel / Garcés, Juan, 2009, "Open Source Critical Editions: A Rationale". In: M. Deegan & K. Sutherland, (eds.) *Text editing, print and the digital world*. Farnham England; Burlington VT: Ashgate: 3-98.

Boschetti, Federico, 2007, "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)*. Available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf [Accessed January 4, 2009].

[16] The *First thousand years of Greek* <http://chs75.chs.harvard.edu/projects/diginc/first1kyears> is currently the main example of such projects.

BOSCHETTI, Federico, 2010, *A Corpus-based Approach to Philological Issues*. Available at: http://eprints-phd.biblio.unitn.it/185/ [Accessed December 10, 2010].

CRANE, Gregory ET AL., 2006, "Beyond digital incunabula: Modeling the next generation of digital libraries". In: *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*: 353-366.

JACOBY, Felix, 1958, *Die Fragmente der Griechischen Historiker, von Felix Jacoby.*, Leiden, Brill.

SMITH, Neel, 2009, "Citation in Classical Studies". In *Digital Humanities Quarterly*, 3(1). Available at: http://www.digitalhumanities.org/dhq/vol/003/1/000028.html [Accessed March 15, 2009].

NELSON, Ted, 1983, *Literary machines : the report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics*, Sausalito (Ca.), Mindful Press.

NELSON, Theodor Holm / SMITH, Robert Adamson / MALLICOAT, Marlene, 2007, "Back to the future." In: *Proceedings of the 18th conference on Hypertext and hypermedia - HT '07*. the 18th conference. Manchester, UK: 227. Available at: http://portal.acm.org/citation.cfm?id=1286303 [Accessed December 7, 2010].

PASIN, Michele / MOTTA, Enrico / ZDRAHAL, Zdenek, 2007, "Capturing knowledge about philosophy". In: *Proceedings of the 4th international conference on Knowledge capture*. Whistler, BC, ACM: 47-54. Available at: http://portal.acm.org/citation.cfm?id=1298406.1298416 [Accessed January 29, 2009].

ROMANELLO, Matteo ET AL., 2009, "When printed hypertexts go digital: information extraction from the parsing of indices". In: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, Torino, ACM: 357-358.
Available at: http://portal.acm.org/citation.cfm?id=1557914.1557987 [Accessed June 30, 2009].

ROMANELLO, Matteo ET AL., 2009, "Rethinking Critical Editions of Fragmentary Texts By Ontologies". In: MORNATI, S. / T. HEDLUND, (eds.), *Proceedings of 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies*, Milano: 155-174.
Available at: http://conferences.elpub.net/index.php/elpub/elpub2009/paper/view/158/66 [Accessed September 11, 2009].

RUBENSTEIN, Herbert / GOODENOUGH, John B., 1965, "Contextual correlates of synonymy". In *Commun. ACM*, 8(10): 627-633.
Available at: http://portal.acm.org/citation.cfm?id=365657 [Accessed November 14, 2009].

Matteo Romanello
Centre for Computing in the Humanities
King's College
London
matteo.romanello@kcl.ac.uk