Journal of the Text Encoding Initiative

Issue 8 - PREVIEW (2014-2015)
Selected Papers from the 2013 TEI Conference

Monica Berti, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova and Gregory R. Crane

The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors

Electronic reference
Monica Berti, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova and Gregory R. Crane, « The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors », *Journal of the Text Encoding Initiative* [Online], Issue 8 - PREVIEW | 2014-2015, Online since 01 January 2015, connection on 09 June 2015. URL : http://jtei.revues.org/1218 ; DOI : 10.4000/jtei.1218

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on: http://jtei.revues.org/1218
This PDF document was generated by the journal.

The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors

Monica Berti, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova, and Gregory R. Crane

1. Introduction

1 This paper presents the research work of the Humboldt Chair of Digital Humanities at the University of Leipzig to establish a new Leipzig Open Fragmentary Texts Series (LOFTS).[1] The project is developed in conjunction with the Perseus Digital Library at Tufts University and the Harvard University Center for Hellenic Studies. LOFTS produces open editions of ancient classical works that survive through quotations and text reuses in later texts (i.e., textual fragments). These fragments include many formats that range from quotations to allusions and translations, which are only a shadowy image of the original depending on their distance from a literal citation (for a definition of fragmentary texts, see Berti 2012 and Berti 2013). Generations of scholars have looked for information about lost authors and have created huge print collections of fragmentary works that cover almost every literary genre in prose and verse, allowing for the rediscovery of authors otherwise lost and forgotten (Most 1997). As far as Greek literature from the eighth century BCE to

the third century CE is concerned, we know that the works of about sixty percent of known authors are preserved only in textual fragments, thus showing the importance of this form of evidence (Berti et al. 2009).

2 Print editions of fragmentary works are indexed collections of excerpts extracted from their contexts and from the textual data about those contexts. Editions of fragmentary works are fundamentally hypertexts and the goal of this project is to produce a dynamic infrastructure for a full representation of the complex relationships between sources, quotations, and annotations about them. In a digital edition, fragments are linked directly to the source text from which they are drawn and can be accurately aligned to multiple editions (Berti et al. 2009). Accordingly, digital fragments are contextualized annotations about reused authors and works. As new versions of, or scholarship on, the source texts emerge in a standard, machine-actionable form, these new findings are automatically linked to the digital fragments. Such a representation allows us to go beyond the limits of print collections of fragmentary authors, where quotations and text reuses are snippets of text completely isolated from their original context. Fragmentary texts also include quotations and reuses of surviving sources (e.g., Herodotus quoted by Athenaeus). In these cases it is possible to compare the original text with its reuses and, therefore, to produce different textual alignments that help scholars to infer the quotation habits of ancient authors.

3 The first step for rethinking the significance of text reuses of lost works is to represent them inside their preserving context. This entails the selection of the string of words that belong to the portion of text classifiable as reuse and to encode all those elements that signal the presence of the text reuse itself (i.e., named entities such as the onomastics of reused authors, titles of reused works and descriptions of their content, *verba dicendi*, syntax, etc.). The second step is to encode all the information pertaining to other sources that reuse the same original text with different words or a different syntax (witnesses), or that deal with the same topic of the text reuse (parallel texts), as well as alternative editions and translations of both the source and the derived texts (Almas and Berti 2013).

4 LOFTS has two main objectives: (1) to digitize paper editions of fragmentary works and link them to source texts, and (2) to produce born-digital editions of fragmentary works. In order to achieve these goals, LOFTS is implementing a Fragmentary Texts Editor within the Perseids Platform[2] (Almas and Berti 2013) and is producing a digital edition of the *Fragmenta Historicorum Graecorum*

edited by Karl Müller in the nineteenth century (Müller 1878–85; DFHG Project).[3] Alongside the DFHG project, LOFTS is working on two other subprojects: (1) the Digital Athenaeus, whose goal is to produce a digital edition of the *Deipnosophists*, with a complete survey of quotations of both lost and surviving sources preserved by Athenaeus of Naucratis;[4] and (2) the Digital Marmor Parium, which seeks to create a digital edition of the so-called Parian Marble, a Hellenistic chronicle on marble from the Greek island of Paros[5]. The latter is the work of a fragmentary author whose name is lost and the project serves as a test case to build a model that works with both fragmentary texts and physical fragments of ancient works such as inscriptions and papyri (Berti and Stoyanova 2014).

5 The present paper is divided into two parts. The first part (section 2) describes the implementation of a fragmentary texts editor within Perseids, the editorial platform developed by the Perseus Project for collaborative annotation of classical source documents. Perseids facilitates the annotation of text reuses, the production of syntactic annotations, the alignment of multiple texts, and the production of digital commentaries on fragmentary works. This section also elaborates on the complex nature of fragmentary texts and how different technologies need to be combined to reach our goals.

6 The second part (section 3) describes the Digital Fragmenta Historicorum Graecorum, a synergetic effort which is currently encoding print collections of fragmentary texts and creating a directory of fragmentary authors which feeds into the Perseus catalog.

2. Perseids Platform

7 The Perseids Platform supports collaborative editing, annotation, and publication of born-digital editions of source documents in the classics.[6] Perseids is not one single application but an integrated environment built from a loose coupling of heterogeneous tools and services from a variety of sources. The development of the Perseids platform was inspired and motivated by the work of several pre-existing projects: the Tufts Miscellany Collection at Tufts University,[7] the Homer Multitext Project at the Harvard's Center for Hellenic Studies,[8] and the Papyri.info project[9] (Almas and Beaulieu 2013). The Son of SUDA Online (SoSOL) application sits at the core of the Perseids platform. SoSOL is a Ruby on Rails[10] application, originally developed by the Papyri.info project, that serves as front end for a Git[11] repository of documents, metadata, and annotations.

It includes a workflow engine that enables documents and data of different types to pass through flexible review and approval processes. The SoSOL application includes user interfaces for editing XML documents, metadata, and annotations. While it does not include a full-featured XML editor, it supports alternative text-based input of XML markup, and can enforce XML schema validation rules on the documents being edited.

8 A key goal behind the initial development of the platform was to enable original undergraduate research in the field of Classics.[12] The workflows related to the encoding of text reuses and lost authors represent core use cases for the current phase of work on the platform (Almas and Berti 2013).[13] In developing features of the Perseids Platform to support these workflows, we are focusing first and foremost on the data. We expect that techniques for visually representing digital editions will change rapidly with technology. So, while our work includes prototype representations of digital editions suitable for publication on the web, our first priority is to enable scholars to create data about the authors, texts and related commentaries, annotations, links, and translations in a way that encourages and facilitates their preservation and reuse. We have identified the following core requirements to meet this goal:

- The ability to represent the texts themselves, links between them, and annotations and commentaries on them, in semantically and structurally meaningful ways that adhere to well-accepted and documented standard formats.
- Stable and resolvable identifiers for all relevant data points, including:
    - the lost authors and their works
    - the authors and extant texts that preserve quotations and text reuses of the lost works
    - different editions and translations of the lost and extant texts
    - named entities (e.g., persons, places, and events) mentioned within the texts
    - commentaries and annotations on the texts, from ancient times through the present
- The ability to group any of the data points into collections representing different contextual views of the data.
- The ability to accurately represent provenance information for data and workflows.

2.1 Data Formats

2.1.1 Texts

9 We use TEI encoding to represent the source texts and textual fragments preserved within them.[14] TEI provides the markup syntax and vocabulary needed to produce XML that enables citable passages of text to be unambiguously identified and linked to *within their preserving context*, a key requirement of the representation of text reuses as discussed above. For example, the following excerpt from Athenaeus' *Deipnosophists* 3.6 uses the TEI `<div>` element to demarcate the book and chapter:

```
<div type="textpart" subtype="book" n="3">
 ...
 <div type="textpart" subtype="chapter" n="6">
  <p>
  ...Ἴστρος δ' ἐν τοῖς Ἀττικοῖς οὐδ' ἐξάγεσθαί φησι τῆς Ἀττικῆς τὰς ἀπ'
αὐτῶν γινομένας ἰσχάδας, ἵνα μόνοι ἀπολαύοιεν οἱ κατοικοῦντες: καὶ ἐπεὶ πολλοὶ
ἐνεφανίζοντο διακλέπτοντες, οἱ τούτους μηνύοντες τοῖς δικασταῖς ἐκλήθησαν τότε
πρῶτον συκοφάνται. ...
  </p>
 </div>
</div>
```

2.1.2 Identifiers

10 The CITE Architecture developed by the Homer Multitext Project[15] provides us with a standard identifier syntax for texts, passages, and related objects and with APIs for services which can retrieve objects identified via these protocols. CITE defines Canonical Text Services (CTS) URNs for creating semantically meaningful unique identifiers for texts and passages within texts. CITE also defines an alternate URN syntax for data objects that are not citable text notes, such as images, annotations, and the lost texts themselves. As URNs, these identifiers are not web-resolvable on their own. By combining them with a URL prefix and deploying CTS and CITE services to serve the identified resources at those addresses, we have resolvable, stable identifiers for our texts, data objects, and annotations (Smith and Blackwell 2012). One of our key motivations for using CTS URNs is that they give us a robust means of targeting annotations at specific substrings of text within a canonical work. The pointers to the text are specific to the location of these strings within

their canonical citation structure, and not to the XML markup of any particular digital edition of the text. Further, the URN syntax degrades gracefully to allow us to reference either the notional work or a very specific edition of that work. Our goal in using this syntax, together with RDF and stand-off markup techniques as discussed below, is to enable the assertions we are making to stand on their own as data, independent of the encoding techniques used to digitize the text.

11 For example, the following set of identifiers might be used to represent a reuse of a lost work of Istros at book 3, chapter 6 in Athenaeus' *Deipnosophists* (Almas and Berti 2013):

`urn:cts:greekLit:tlg0008.tlg001.perseus-grc1:3.6@Ἴστρος%5B1%5D-συκοφάνται%5B1%5D`

12 This is a CTS URN for a subset of passage `3.6` in the `perseus-grc1` edition of the work identified by `tlg001` in the group of texts associated with Athenaeus, identified by `tlg0008`. The URN further specifies a string of text in that passage, starting at the first instance of the word *Ἴστρος* and ending at the first instance of the word *συκοφάνται*.

13 `urn:cite:perseus:lci.2`

is a CITE URN identifier for the instance of lost text being reused. This URN identifies an object whose name `2` is unique within the Perseus Collection of Lost Content Items (`lci`). Every item in this collection points to a specific text reuse of a lost author as it is represented in a modern edition (see below for further discussion of CITE Collections).

14 These URNs represent distinct technology-independent identifiers for the two cited objects, and by prefixing them with the `http://data.perseus.org` URL prefix (representing the web address at which they can be resolved) we create stable web identifiers for them, making them compatible with linked data best practices:[16]

`http://data.perseus.org/citations/urn:cts:greekLit:tlg0008.tlg001.perseus-grc1:3.6%40Ἴστρος%5B1%5D-συκοφάνται%5B1%5D`[17]

`http://data.perseus.org/collections/urn:cite:perseus:lci.2`

15 The RDF data model also gives us more precision than XML with respect to targets of annotations and subjects of assertions. A human reader can usually tell from context whether an assertion concerns the wording of a sentence, the writing of a sentence as an event, the author of a sentence, or another scholar's assertion about the sentence. XML has no standardized semantics, and interpretations of relationships among elements and attributes are often underspecified in

XML vocabularies (Renear et al. 2002). In RDF, subject, predicate, and object roles in a statement are explicit at the data structure level. Distinctions among domain entities (such as author vs. authorship event, or morpheme vs. sentence) are encoded as a class identity for the resource.

2.1.3 Annotations

16 The term "Annotations" covers a potentially wide variety of data types. We can have simple annotations in the form of typed links between data points, such as: a textual fragment and a proposed author of that fragment; detailed textual commentaries making an assertion about a text; a complex morphosyntactic analysis of a section of text; an alignment between editions or translations of a text. The Open Annotation (OA) data model "specifies an interoperable framework for creating associations between related resources, annotations, using a methodology that conforms to the Architecture of the World Wide Web."[18] The OA model enables us to serialize every annotation in its most simple form, as a link between one or more *target* items being annotated, and one or more *bodies* representing the contents of the annotation. OA also gives us a standard vocabulary for categorizing the motivation for the annotations. URIs are used to specify both the target and the body of the annotation. We use the OA data model both as the primary representation of an annotation, in cases where the annotations are created by linking two identifiers (such as a link between a passage in a text and an identifier for a named entity or event), and also as a serialization method for more complex annotations, where the annotation process involves the creation of complex documents as the annotation bodies which we can then reference by their URI identifiers. In the latter case, we use a variety of standard formats for the actual annotation bodies, including:

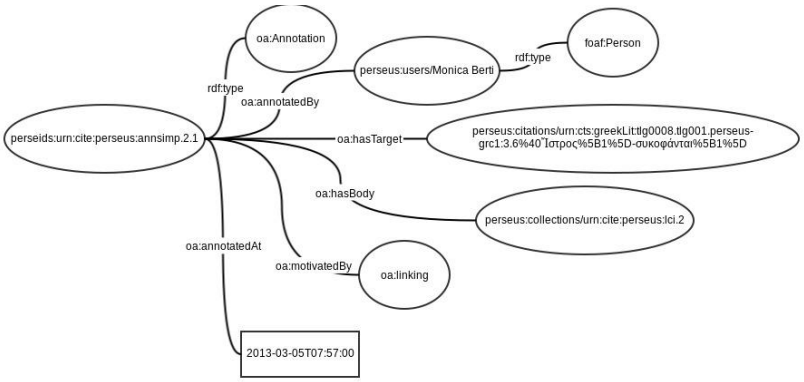
- The Perseus Ancient Greek and Latin Treebank Schema[19] for morphosyntactic analyses.
- The Alpheios Translation Alignment Schema[20] for text alignments.
- Markdown Syntax[21] for short textual commentaries.
- TEI XML for primary and secondary source texts.

17 The annotation representing the assertion (discussed above) that text at Athen., *Deipn.* 3.6 describes a reuse of a lost work of Istros identified by `urn:cite:perseus:lci.2`, serialized in OA using the JSON-LD[22] format, might be formalized as follows (Almas and Berti 2013):

```
{
  "@context": "http://www.w3.org/ns/oa-context-20130208.json",
  "@id": "http://perseids.org/annotations/urn:cite:perseus:annsimp.2.1",
  "@type": "oa:Annotation",
  "annotatedAt": "2013-03-05T07:57:00",
  "annotatedBy": {
      "@id": "http://data.perseus.org/sosol/users/Monica Berti",
      "@type": "foaf:Person",
      "name": "Monica Berti"
  },
  "hasBody": "http://data.perseus.org/collections/urn:cite:perseus:lci.2",
  "hasTarget": "http://data.perseus.org/citations/
urn:cts:greekLit:tlg0008.tlg001.perseus-grc1:3.6%40Ἴστρος%5B1%5D-συκοφάνται
%5B1%5D",
  "oa:motivatedBy":"oa:linking"
}
```

18 Figure 1 shows a possible visual representation of the same annotation:

Figure 1: Visual representation of an OA annotation of a text reuse in Athen., *Deipn*. 3.6



2.2 Collections

19 We need to be able to organize text reuses into various types of collections of data, including those represented in a given traditional print edition which comprises: reuses from one or many authors; all text reuses attributed to a specific author; all text reuses quoted by a specific author; all text reuses referencing a specific topic; all text reuses attributed to a specific time period. CITE Collection Services define a protocol for identifying and retrieving digital representations of

objects identified by CITE URNs. CITE URNs uniquely identify objects in a set of objects of a similar kind: for example, images, manuscripts, and entries in commentaries. (CITE stands for Collections, Indices, Texts, and Extensions.) Examples of CITE collections used to support the encoding of text reuses for our project include the abstract lost text entities themselves, digital images of manuscripts of the extant source texts that quote those lost texts, commentaries on instances of text reuse, and linguistic annotations of the quoted text.

2.3 Provenance

20 Scholars produce data through a variety of activities, observations, and other events. Some data is created, other data discovered. The *provenance* of a digital object is an account of its origin and change over time. Documenting and preserving data provenance in a structured, machine-readable format enables us to more precisely track and document shared resources, ultimately improving data quality and encouraging further sharing. Specifically, we intend our retrieval, editorial annotation, and communication tools to create records of the research transactions in which they are used. These records are expressed in RDF vocabularies that are based on abstract provenance models.

21 According to Groth et al. (2006), two key principles for provenance data are that (1) actors must only record propositions that they know to be true, through statements of what they observe; and (2) each statement of provenance must be attributable to a particular actor.

22 In our use case, we need to be able to (1) reference ancient data that can be identified but that did not literally come into existence as the result of any modern computational interaction (and which may in fact no longer be extant in any preserved source); and (2) identify the role a data item, such as an ancient scholarly assertion, plays as the vehicle for the modern scholarly claims. A third requirement, which results from the second, is that we need to be able to represent the assertions of the ancient scholars, on which our modern assertions depend, in a format that can be included computationally in a common data set with the modern claims.

23 Although our usual workflow discourse describes domain entities coming into existence and undergoing genuine change, that understanding poses problems for tracking identity and other semantic relationships. For example, if a text is enriched by adding TEI markup, what exactly is preserved in the derivation of the newer version from the older? On what basis do we identify the

same original work in each reuse despite the superficial differences between them? And how do we represent assertions concerning works that have no physical exemplars? Our solution to dilemmas like these is to employ complementary data provenance models: one with a transformational view and another with a semiotic view of the same research events.

24 The W3 Consortium's PROV is a specification for expressing provenance records with descriptions of the entities and activities involved in producing and delivering or otherwise influencing a given digital object.[23] A PROV *activity* is an event through which entities come into existence and/or change to become new entities. Activities are dynamic aspects of the world, such as actions and processes. For example, a PROV account of the OA annotation shown in the previous section would document its coming into existence at a particular time through a researcher's use of annotation software.

25 In contrast to PROV, the Systematic Assertion Model (SAM) treats events in scholarship as linguistic acts (Wickett et al. 2012). SAM is concerned not with mutable data structures that change from one form to another, but with unchanging abstract objects that come to stand in new contingent relationships during a project. A SAM account of the OA annotation, for example, would be concerned with its role as a vehicle for the scholar's claim of a proposition's truth (that is, the proposition that Athen., *Deipn*. 3.6, quotes LCI 2). A SAM description links an annotation expression to the act of annotation, the claim advanced at that time and place, the annotator's justification for the claim, and to interpretive contexts necessary for understanding the annotation. In the present example these contexts would include the OA vocabulary and the JSON-LD serialization syntax.

26 SAM's formal account of research data and its content supplies contextual information that PROV lacks, but PROV can record functional relationships that SAM hides: because SAM's abstract symbol structures and propositions are immutable, their roles in a project are incompletely specified without the PROV view of their use as inputs and outputs to activities.

27 In the scenario discussed in this paper, a scholar wants to annotate what she thinks is a reuse of a lost text attributed to Istros in an extant source text by Plutarch (*Sol.* 24.1), where Istros is not named.[24] The scholar, using the Perseids annotation tools to produce a concrete representation of this assertion, *indicates* the text that represents the reuse. An example of an RDF triple documenting this action—the selection of the string starting with the first instance of *ἐξαγωγὴ* to the first instance of *συκοφαντεῖν* according to the SAM data model—is:

```
  <system:actions/berti/textsel1> sam:indicates [ xsd:string
 "ἐξαγωγὴ1,συκοφαντεῖν1"];
  event:agent <perseids:users/berti>;
  event:time [xsd:dateTime "2013-05-20T09:01:00"];
  a sam:Indication.
```

28 Although we could represent this action as an *activity* involving the scholar's physical interaction with the system, we would lose the significance of the linguistic force behind the scholar's identification of the text.

29 In reaction to her selection of text, the system computes a URI identifier for that text. This event can be represented as a SAM *computation*, but in order to represent the fact that the URI *came into being* as a result of a user-system interaction, it is also appropriate to describe this as a PROV *activity* that was informed by the scholar's previous action of selecting a string of text (Almas et al. 2013):

```
  <system:urncompute1> sam:indicates <data:citations/
urn:cts:greekLit:tlg0007.tlg007.perseus-grc1:24.1%40ἐξαγωγὴ%5B1%5D-συκοφαντεῖν
%5B1%5D>;
  event:agent <system:installation1>;
  event:time [xsd:dateTime "2013-05-20T09:01:01"];
  a sam:Computation, prov:Activity;
  prov:wasInformedBy <system:actions/berti/textsel1>.
```

2.4 Dissemination and Presentation

30 As discussed above, our primary focus thus far has been on capturing the data about the authors, texts and related commentaries, annotations, links, and translations in a way that is accurate and also encourages and facilitates its preservation and reuse. Visual representation of the data is one type of reuse, and the data format selections have been made with the need to support disseminations for online presentation in mind. The JSON-LD syntax recommended by OA allows us to easily build a dynamic display interface in Javascript which navigates the JSON-LD data object and retrieves the datasets identified as the targets and bodies of the annotations at their addressable URIs, as served by supporting CTS and CITE services. Our prototype interface[25] provides a demonstration of one possible approach to a digital representation of text reuse data. Similarly, although we hope eventually to be able to represent the rich provenance metadata

discussed above in the visual representation of our digital editions, we are above all concerned with capturing the data in a way that ensures they can be preserved and serve as the basis for further research.

3. The Digital Fragmenta Historicorum Graecorum (DFHG) Project

31 As mentioned before, one of the goals of LOFTS is to digitize paper editions of fragmentary works and link them to the source texts from which the fragments have been excerpted. Such a work aims both at preserving a philological heritage and at providing a methodological foundation for implementing a new generation of born-digital editions of fragmentary texts. Accordingly, LOFTS is encoding the *Fragmenta Historicorum Graecorum* (*FHG*), which is one of the most important collections of textual fragments of Classical authors.

3.1 The *FHG* Texts

32 The five volumes of the *FHG* edited by Karl Müller (1878–85) comprise the first big collection of fragments of Greek historians' work ever realized. The work collects excerpts from textual sources pertaining to more than six hundred fragmentary authors. Excluding the first volume, these authors are chronologically distributed and range from the sixth century BCE to the seventh century CE. Under each author the fragments are numbered sequentially and arranged according to works and book numbers (when such information is available), and every fragment is translated into Latin. The first volume also includes the text of the inscription of the *Marmor Parium* with a Latin translation, a chronological table, and a commentary, and the Greek text of the Rosetta Stone with a French literal translation, as well as a critical, historical, and archaeological commentary. The fifth volume includes fragments of the writings of Greek and Syriac historians preserved in Armenian sources.

33 The basic structure of the *FHG* has introductions to the volumes with discussions of the content, citations of sources, and references to bibliographical records. Sometimes short introductions precede the fragments of the fragmentary authors with biographical information and related

textual evidence. The Greek fragments are arranged into two columns and the Latin translations are printed at the bottom of the page below a horizontal line. The *FHG* does not include a critical apparatus for the Greek text.

3.2 The DFHG Project

34 LOFTS is producing the first digital edition of the *FHG* as a means of providing an open, linked, machine-actionable text for the study and advancement of Greek fragmentary literature.[26] The text is encoded in accordance with the latest TEI EpiDoc standards.[27] The decision to adopt the EpiDoc subset was dictated by two factors: (1) EpiDoc markup lends itself well to ancient texts, be these papyri, inscriptions, manuscripts, or editions, allowing for various levels of detail; (2) encoding in EpiDoc ensures compatibility with existing papyrological and epigraphic resources, with the EAGLE-Europeana network,[28] and with the forthcoming EpiDoc-compliant Perseus 5.0 databank.

35 The Leipzig Humboldt Chair and Perseus Digital Library partnership entails a synergetic effort not only to enrich the Perseus Catalog[29] but also to provide new testing material for the continuous development of the Perseids Platform, where the first EpiDoc version of the *FHG* will be placed for further annotation. This undertaking is proving to be invaluable not only for the growth of partner resources but also for the development of the EpiDoc schema itself.[30]

36 The project team is also working on the text of the so-called *Marmor Parium* (*IG* 12.5.444), which is a fragmentary inscription from the island of Paros that preserves a Hellenistic chronicle from the reign of Cecrops (1581/1580 BCE) to the archonship of Euctemon (299/298 BCE).[31] The epigraphical text was edited in the first volume of the *FHG* because of its historiographical value (Müller 1878–85, 1:533–90). The author of the text is unknown, but the content reflects his choices and it consists of a list of historical events mainly based on the Athenian history. In this respect, this evidence is a perfect example of a fragmentary author, whose work is preserved not through quotations in later texts, but in a fragmented original form. The text of the inscription is also being encoded in EpiDoc. An important part of the project is the identification of named entities mentioned in the inscription (such as names of kings and magistrates, personal names, and place names). The Pleiades gazetteer has been referenced for the place names.[32] The identification of individuals will make use of and feed into the Standards for Networking Ancient Prosopographies project (SNAP).[33] The team is also producing a visualization of the chronology preserved by the *Marmor Parium* with

the open source tool TimelineJS, which allows the comparison of the text not only with other ancient chronologies but also with different chronological interpretations of the content of the inscription made by modern scholars (Berti and Stoyanova 2014).[34]

3.3 The EpiDoc Guidelines and Encoding Process

37 The digital text of the *FHG* was obtained by feeding the volume scans available at the Internet Archive[35] to an Optical Character Recognition (OCR) engine, which transformed the printed text into digital form. The error-laden Greek output was corrected semi-automatically and stored in text files. Any remaining errors are manually rectified during the encoding process.

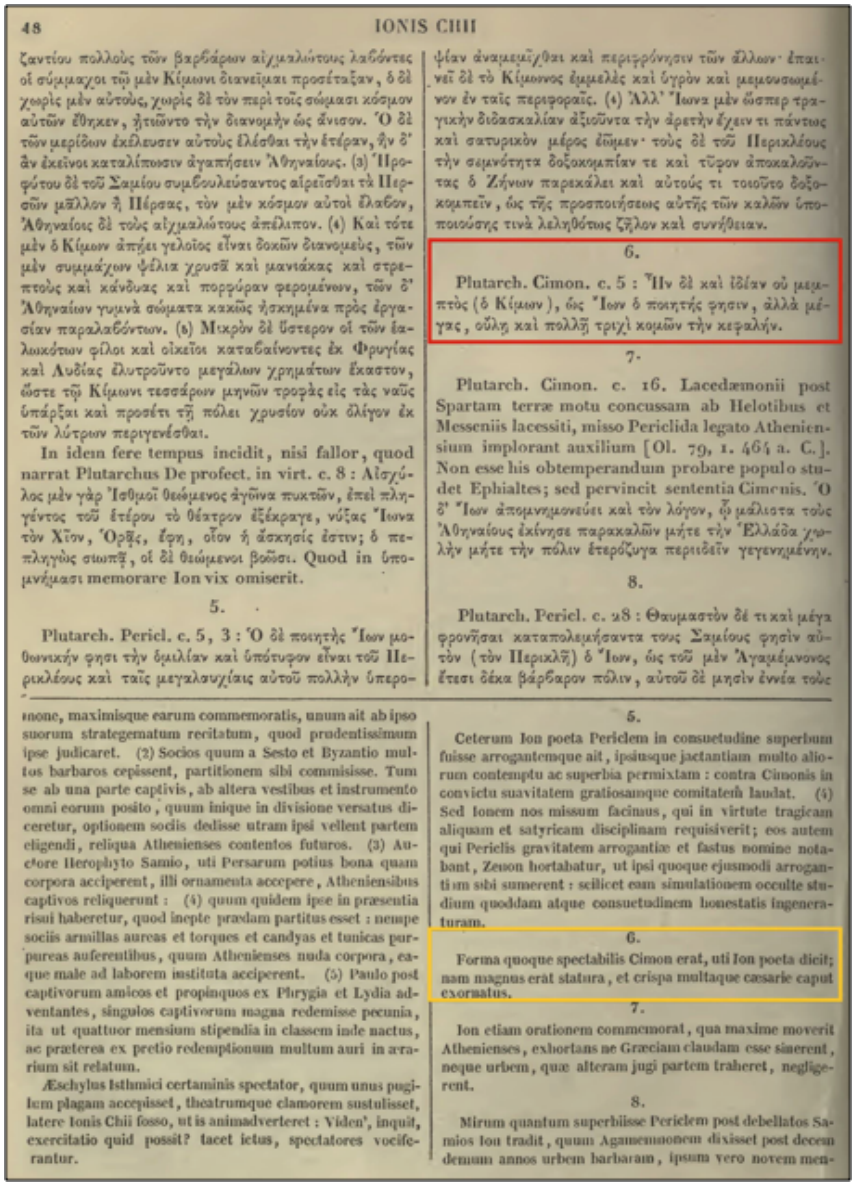
38 LOFTS has drawn up an EpiDoc template that is compatible with the Perseus Digital Library corpus and applicable to all types of texts produced by the University of Leipzig Open Greek and Latin (OGL) Project, which is producing a comprehensive open collection of Classical sources.[36] The template covers basic layout and citation structures with room for alteration and improvement and it is integrated with encoding guidelines, produced not only to document the LOFTS editorial choices but also to help external individuals contribute to the encoding process. The project production line consists of two stages: (1) the encoding process is implemented by LOFTS and tackles layout, citation elements, and philological issues; (2) the initial editorial endeavor and the guidelines are made available for crowdsourcing, where contributors are encouraged to further tag the EpiDoc files with information regarding places, personal names, and any other relevant entities.

3.3.1 Stage One: First Level of Encoding by LOFTS

39 The editorial team of LOFTS creates one XML file per fragmentary author. Every file contains a `<teiHeader>` with specific information about the author, volume, or book in question. As for the `<text>`, Müller's layout is not kept since the Greek text is separated from its Latin translation. The structure within `<text>` reflects the structure of each volume, using the `<div type="textpart">` with different `@subtype` values, whenever needed. Müller's Latin translations of the fragments are encoded in a separate XML file to facilitate text alignment.

40 For example, fragment 6 of Ion of Chios will be encoded as follows (figures 2, 3, and 4):

Figure 2: Page image from FHG 2:48: Greek text of fragment 6 (outlined in red) and Latin translation (outlined in yellow)



48 IONIS CHII

ζαντίου πολλοὺς τῶν βαρβάρων αἰχμαλώτους λαβόντες οἱ σύμμαχοι τῷ μὲν Κίμωνι διανεῖμαι προσέταξαν, ὁ δὲ χωρὶς μὲν αὐτοὺς, χωρὶς δὲ τὸν περὶ τοῖς σώμασι κόσμον αὐτῶν ἔθηκεν, ἠτιῶντο τὴν διανομὴν ὡς ἄνισον. Ὁ δὲ τῶν μερίδων ἐκέλευσεν αὐτοὺς ἑλέσθαι τὴν ἑτέραν, ἣν δ' ἂν ἐκεῖνοι καταλίπωσιν ἀγαπήσειν Ἀθηναίους. (3) Προφύτου δὲ τοῦ Σαμίου συμβουλεύσαντος αἱρεῖσθαι τὰ Περσῶν μᾶλλον ἢ Πέρσας, τὸν μὲν κόσμον αὐτοὶ ἔλαβον, Ἀθηναίοις δὲ τοὺς αἰχμαλώτους ἀπέλιπον. (4) Καὶ τότε μὲν ὁ Κίμων ἀπῆει γελοῖος εἶναι δοκῶν διανομεὺς, τῶν μὲν συμμάχων ψέλια χρυσᾶ καὶ μανιάκας καὶ στρεπτοὺς καὶ κάνδυας καὶ πορφύραν φερομένων, τῶν δ' Ἀθηναίων γυμνὰ σώματα κακῶς ἠσκημένα πρὸς ἐργασίαν παραλαβόντων. (5) Μικρὸν δὲ ὕστερον οἱ τῶν ἁλωκότων φίλοι καὶ οἰκεῖοι καταβαίνοντες ἐκ Φρυγίας καὶ Λυδίας ἐλυτροῦντο μεγάλων χρημάτων ἕκαστον, ὥστε τῷ Κίμωνι τεσσάρων μηνῶν τροφὰς εἰς τὰς ναῦς ὑπάρξαι καὶ προσέτι τῇ πόλει χρυσίον οὐκ ὀλίγον ἐκ τῶν λύτρων περιγενέσθαι.

In idem fere tempus incidit, nisi fallor, quod narrat Plutarchus De profect. in virt. c. 8 : Αἰσχύλος μὲν γὰρ Ἰσθμοῖ θεώμενος ἀγῶνα πυκτῶν, ἐπεὶ πληγέντος τοῦ ἑτέρου τὸ θέατρον ἐξέκραγε, νύξας Ἴωνα τὸν Χῖον, Ὄρᾳς, ἔφη, οἷον ἡ ἄσκησίς ἐστιν; ὁ περιπληγὼς σιωπᾷ, οἱ δὲ θεώμενοι βοῶσι. Quod in ὑπομνήμασι memorare Ion vix omiserit.

5.

Plutarch. Pericl. c. 5, 3 : Ὁ δὲ ποιητὴς Ἴων μοθωνικήν φησι τὴν ὁμιλίαν καὶ ὑπότυφον εἶναι τοῦ Περικλέους καὶ ταῖς μεγαλαυχίαις αὐτοῦ πολλὴν ὑπεροψίαν ἀναμεμῖχθαι καὶ περιῤῥόνησιν τῶν ἄλλων· ἀπαιτεῖ δὲ τὸ Κίμωνος ἐμμελὲς καὶ ὑγρὸν καὶ μεμουσωμένον ἐν ταῖς περιφοραῖς. (4) Ἀλλ' Ἴωνα μὲν ὥσπερ τραγικὴν διδασκαλίαν ἀξιοῦντα τὴν ἀρετὴν ἔχειν τι πάντως καὶ σατυρικὸν μέρος ἔῶμεν· τοὺς δὲ τοῦ Περικλέους τὴν σεμνότητα δοξοκομπίαν τε καὶ τῦφον ἀποκαλοῦντας ὁ Ζήνων παρεκάλει καὶ αὐτούς τι τοιοῦτο δοξοκομπεῖν, ὡς τῆς προσποιήσεως αὐτῆς τῶν καλῶν ὑποποιούσης τινὰ λεληθότως ζῆλον καὶ συνήθειαν.

6.

Plutarch. Cimon. c. 5 : Ἦν δὲ καὶ ἰδέαν οὐ μεμπτὸς (ὁ Κίμων), ὡς Ἴων ὁ ποιητής φησιν, ἀλλὰ μέγας, οὔλῃ καὶ πολλῇ τριχὶ κομῶν τὴν κεφαλήν.

7.

Plutarch. Cimon. c. 16. Lacedæmonii post Spartam terræ motu concussam ab Helotibus et Messeniis lacessiti, misso Periclida legato Atheniensium implorant auxilium [Ol. 79, 1. 464 a. C.]. Non esse his obtemperandum probare populo studet Ephialtes; sed pervincit sententia Cimonis. Ὁ δ' Ἴων ἀπομνημονεύει καὶ τὸν λόγον, ᾧ μάλιστα τοὺς Ἀθηναίους ἐκίνησε παρακαλῶν μήτε τὴν Ἑλλάδα χωλὴν μήτε τὴν πόλιν ἑτερόζυγα περιιδεῖν γεγενημένην.

8.

Plutarch. Pericl. c. 28 : Θαυμαστὸν δέ τι καὶ μέγα φρονῆσαι καταπολεμήσαντα τοὺς Σαμίους φησὶν αὐτὸν (τὸν Περικλῆ) ὁ Ἴων, ὡς τοῦ μὲν Ἀγαμέμνονος ἔτεσι δέκα βάρβαρον πόλιν, αὐτοῦ δὲ μησὶν ἐννέα τοὺς

mone, maximisque earum commemoratis, unum ait ab ipso suorum strategematum recitatum, quod prudentissimum ipse judicaret. (2) Socios quum a Sesto et Byzantio multos barbaros cepissent, partitionem sibi commisisse. Tum se ab una parte captivis, ab altera vestibus et instrumento omni eorum posito, quum inique in divisione versatus diceretur, optionem sociis dedisse utram ipsi vellent partem eligendi, reliqua Atheniensis contentos futuros. (3) Auctore Herophyto Samio, uti Persarum potius bona quam corpora acciperent, illi ornamenta accepere, Atheniensibus captivos reliquerunt : (4) quum quidem ipse in præsentia risui haberetur, quod inepte prædam partitus esset : nempe sociis armillas aureas et torques et candyas et tunicas purpureas auferentibus, quum Atheniensis nuda corpora, eaque male ad laborem instituta acciperent. (5) Paulo post captivorum amicos et propinquos ex Phrygia et Lydia adventantes, singulos captivorum magna redemisse pecunia, ita ut quattuor mensium stipendia in classem inde nactus, ac præterea ex pretio redemptionum multum auri in ærarium sit relatum.

Æschylus Isthmici certaminis spectator, quum unus pugilem plagam accepisset, theatrumque clamorem sustulisset, latere Ionis Chii fosso, ut is animadverteret : Viden', inquit, exercitatio quid possit? tacet ictus, spectatores vociferantur.

5.

Ceterum Ion poeta Periclem in consuetudine superbum fuisse arrogantemque ait, ipsiusque jactantiam multo aliorum contemptu ac superbia permixtam : contra Cimonis in convictu suavitatem gratiosamque comitatem laudat. (4) Sed Ionem nos missum facimus, qui in virtute tragicam aliquam et satyricam disciplinam requisiverit; eos autem qui Periclis gravitatem arrogantiæ et fastus nomine notabant, Zenon hortabatur, ut ipsi quoque ejusmodi arrogantiam sibi sumerent : scilicet eam simulationem occulte studium quoddam atque consuetudinem honestatis ingeneraturam.

6.

Forma quoque spectabilis Cimon erat, uti Ion poeta dicit; nam magnus erat statura, et crispa multaque cæsarie caput exornatus.

7.

Ion etiam orationem commemorat, qua maxime moverit Athenienses, exhortans ne Græciam claudam esse sinerent, neque urbem, quæ alteram jugi partem traheret, negligerent.

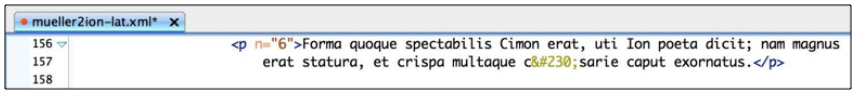
8.

Mirum quantum superbisse Periclem post debellatos Samios Ion tradit, quum Agamemnonem dixisset post decem demum annos urbem barbaram, ipsum vero novem men-

Figure 3: Image from the XML file for the Greek translation

```
mueller2ion-grc.xml
500
501          <cit n="6" xml:lang="grc">
502              <bibl><author role="witness">Plutarch</author>. Cimon. c. 5</bibl>
503              <quote>Ἦν δὲ καὶ ἰδέαν οὐ μεμπτὸς (ὁ Κίμων), ὡς Ἴων ὁ ποιητής φησιν, ἀλλὰ
504                  μέγας, οὔλῃ καὶ πολλῇ τριχὶ κομῶν τὴν κεφαλήν.</quote>
505          </cit>
```

Figure 4: Image from the XML file for the Latin translation



41 While retaining the same number (6), the Greek is encoded in a `<cit>` and the Latin translation in a `<p>`. The `<cit>` is broken down into `<bibl>`, which contains the source `<author>` (Plutarch) and work, and a `<quote>`, containing the Greek text. Whenever there is a note, the `<note>` element is also placed within `<cit>`. Any information that does not strictly pertain to the fragment is encoded in a `<p>`. This initial editorial stage also involves replacing special characters, such as the *æ* diphthong, with their Unicode entities (see the Latin translation, where *æ* is displayed as `&#230;`) in order to avoid font and potential display issues and confusion between graphically similar but semantically different characters (e.g., capital Latin c and Greek capital lunate sigma). Numbers, including ordinal, are also tagged. Footnotes are given arabic numbers (Müller uses asterisks) for clarity.

42 An XSL transformation[37] is used to help the encoder better visualize the marked-up text and links the main body to the footnotes. All references to other works (often abbreviated by Müller) are tagged as `<bibl>` and will eventually point to a master bibliography file. Work titles appearing in the Greek and Latin texts are wrapped in a `<title>` element. Where we have both the Greek and the Latin titles of a work, the two titles are enclosed in another `<title>` tag:

```
  <title><title xml:lang="grc">Περὶ τελετῶν</title> sive <title type="alt"
xml:lang="la">De mysteriis</title></title>
```

43 The *FHG* does not include a critical apparatus, but sometimes Müller signals variant readings in the text, and these are tagged as `<rdg>` elements inside an `<app>` *within* the main text. Personal names are simply tagged as `<persName>`. Names often carry additional information, such as patronymics and epithets or clues about an individual's profession. Encoding such specificity is not a stage one priority but a stage two expectation.

44 Another problematic feature when working with editions of fragmentary works is how to encode specific editorial decisions concerning the attribution of a quotation or a text reuse to an author whose original works are lost. Müller sometimes adds different marks (parentheses, square brackets, or question marks) to the fragment number in order to signal uncertainty about the

attribution of that fragment.[38] In these cases the uncertainty is encoded using the `@ana` attribute to the `<cit>` element, with one of the following values: `"#dubia"`, `"#incerta"`, and `"#anonyma"`. These values are defined in the `<classDecl>` element in `<encodingDesc>` in the header. The reason for this choice depends on the fact that when the project was started not all relevant elements had the `@cert` attribute. Even if they did, Müller reasons for marking something as uncertain vary a lot, and he is also inconsistent in his use of these sigla. The use of the `<certainty>` element also proved problematic, because of this inconsistency, since it was at times difficult to determine whether the uncertainty was the content of the citation, the attribution to an author, or the extent of the quotation. So it was decided to roughly map Müller's three sigla to what seemed the most frequent and probable uncertainty categories across the collection, and rather than have `@cert` or `<certainty>` sometimes in `<cit>` sometimes in `<bibl>` sometimes in `<quote>`, to have three different values of the `@ana` attribute in `<cit>`.

3.3.2 Stage Two: Crowdsourced Annotation

45 Once complete, these basic XML files are deposited in the Perseids Platform for further annotation by third parties. While still in testing mode, this second stage encourages annotators to tag any additional information, including—but not limited to—the following:

- Names (person).[39]
- Names (place).[40]
- Bibliographic references: expansion of bibliographic references and linking to bibliography file.
- Numbers.
- Titles within Greek and Latin text.
- Stage 2 encoding of the Latin translations (after initial encoding at stage one).

46 The editorial board provides the final review for each file. EpiDoc-encoded DFHG files are being progressively added to the DFHG GitHub repository[41] for everyone to download, improve, and share in accordance with our CC BY-SA 4.0 International License.[42]

4. Conclusions

47 The Leipzig Open Fragmentary Texts Series (LOFTS) aims not only at producing an open series of Greek and Latin fragmentary authors, but also at building a model for representing quotations and text reuses of lost works in a digital environment. In order to achieve these goals, different technologies are being used and implemented beyond the TEI XML standard. The final outcome is the production of dynamic excerpts from source texts. In the case of digitized print editions of fragmentary works, this means linking the fragments directly to the source text and annotating their metadata within it. In the case of new-born digital editions of fragmentary texts, fragments become multilayered annotations of information concerning fragmentary authors and reuses of their lost works. These dynamic excerpts contribute to the production of a real "multitext," where each version of the same text embodies a different step in its transmission and a reconstruction of philological conjectures (on the concept of multitext see Blackwell and Crane 2009). The ultimate goals are the creation of open, linked, machine-actionable texts for the study of textual reuses of Classical works and the development of a collaborative environment for crowdsourced annotations that involve both scholars and students.

BIBLIOGRAPHY

Almas, Bridget, and Marie-Claire Beaulieu. 2013. "Developing a New Integrated Editing Platform for Source Documents in Classics." *Literary & Linguistic Computing* 28(4): 493–503. doi:10.1093/llc/fqt046.

Almas, Bridget, and Monica Berti. 2013. "Perseids Collaborative Platform for Annotating Text Re-uses of Fragmentary Authors." In *DH-Case 2013. Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environments: Metadata, Vocabularies and Techniques in the Digital Humanities*, edited by Francesca Tomasi and Fabio Vitali, article no. 7. New York, NY: ACM Digital Library. doi:10.1145/2517978.2517986.

Almas, Bridget, Monica Berti, Sayeed Choudhury, David Dubin, Megan Senseney, and Karen M. Wickett. 2013. "Representing Humanities Research Data Using Complementary Provenance Models." Poster presented at Building Global Partnerships—RDA Second Plenary Meeting, Washington, DC, September 16–18, 2013.

Berti, Monica. 2012. "Citazioni e dinamiche testuali. L'intertestualità e la storiografia greca frammentaria." In *Tradizione e Trasmissione degli Storici Greci Frammentari II. Atti del Terzo Workshop Internazionale. Roma, 24–26 febbraio 2011*, edited by Virgilio Costa, 439–58. Tivoli: Edizioni Tored. http://www.monicaberti.com/wp-content/uploads/2014/08/CitazioniDinamicheTestuali.pdf.

———. 2013. "Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres." *Ancient Society* 43: 269–88. http://poj.peeters-leuven.be/content.php?url=article&id=2992614&journal_code=AS. doi: 10.2143/AS.43.0.2992614.

Berti, Monica, and Simona Stoyanova. 2014. "Digital Marmor Parium. For a Digital Edition of a Greek Chronicle." In *Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the First EAGLE International Conference*, edited by Silvia Orlandi, Raffaella Santucci, Vittore Casarosa, Pietro Maria Liuzzo, 319–24. Roma: Sapienza Università Editrice. http://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf.

Berti, Monica, Matteo Romanello, Alison Babeu, and Gregory R. Crane. 2009. "Collecting Fragmentary Authors in a Digital Library." In *JCDL '09: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, 259–62. New York, NY: ACM Digital Library. doi:10.1145/1555400.1555442.

Blackwell, Christopher, and Gregory R. Crane. 2009. "Conclusion: Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital Age." *Digital Humanities Quarterly* 3(1). http://www.digitalhumanities.org/dhq/vol/3/1/000035/000035.html.

Groth, Paul, Simon Miles, and Steve Munroe. 2006. "Principles of High Quality Documentation for Provenance: A Philosophical Discussion." *Lectures Notes in Computer Science* 4145: 278–86. doi:10.1007/11890850_28.

Most, Glenn W., ed. 1997. *Collecting Fragments. Fragmente sammeln*. Göttingen: Vandenhoeck & Ruprecht.

Müller, Karl, ed. 1878–85. *Fragmenta Historicorum Graecorum.* 5 vols. Paris: A. F. Didot.

Renear, Allen, David Dubin, and C. M. Sperberg-McQueen. 2002. "Towards a Semantics for XML Markup." *DocEng '02: Proceedings of the 2002 ACM Symposium on Document Engineering*, 119–26. New York, NY: ACM Digital Library. doi:10.1145/585058.585081.

Smith, D. Neel, and Christopher W. Blackwell. 2012. "Four URLs, Limitless Apps: Separation of Concerns in the *Homer Multitext* Architecture." In *Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends*. Cambridge, MA: Center for Hellenic Studies, Harvard University. http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=4846.

Wickett, Karen M., Andrea Thomer, Simone Sacchi, Karen S. Baker, and David Dubin. 2012. "What Dataset Descriptions Actually Describe: Using the Systematic Assertion Model to Connect Theory and Practice." Poster presented at the 2012 ASIS&T Research Data Access and Preservation Summit, New Orleans, March 22–23. http://hdl.handle.net/2142/30470.

NOTES

1 http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/the-leipzig-open-fragmentary-texts-series-lofts/.

2 For a prototype interface, see http://perseids.org/sites/berti_demo/.

3 On this project, see below, section 3.

4 Digital Athenaeus, Universität Leipzig, http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/digital-athenaeus/.

5 Digital Marmor Parium, Universität Leipzig, http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/digital-marmor-parium/.

6 http://perseids.org/.

7 http://www.library.tufts.edu/tisch/ematlocalstorage/miscellany_collection/home.html.

8 http://www.homermultitext.org/.

9 http://www.papyri.info/.

10 http://rubyonrails.org/.

11 http://git-scm.com/.

12 This work was supported by grants from Tufts University, the National Endowment for the Humanities (grant HD-51548-12), and the Institute from Museum and Library Services. Funding from the Mellon Foundation is now allowing us to expand the platform to support classroom-based collaboration on digital editions, beginning-to-end scholarly workflows, and the development of dynamic syllabi using the resources managed by the platform.

13 For the prototype functionality used by students of the Digital Philology course at the University of Leipzig in the fall of 2013, see http://sites.tufts.edu/perseids/workflows/fragmentary-author-workflows/fragmentary-text-prototype-fall-2013/.

14 The texts will adhere to the EpiDoc subset of the TEI standard, as discussed further below.

15 Documentation: http://www.homermultitext.org/hmt-doc/cite/index.html.

16 "Perseus Stable URIs," Perseus Digital Library Updates, http://sites.tufts.edu/perseusupdates/beta-features/perseus-stable-uris/.

17 At the time of this writing, complete implementation of the CTS standard for resolution of passage subreferences at the `data.perseus.org` address is still pending.

18 Open Annotation Data Model, February 8, 2013, http://www.openannotation.org/spec/core/20130208/index.html.

19 http://nlp.perseus.tufts.edu/syntax/treebank/agdt/1.7/data/treebank-1.5.xsd.

20 https://svn.code.sf.net/p/alpheios/code/xml_ctl_files/schemas/trunk/aligned-text.xsd.

21 John Gruber, "Markdown: Syntax," http://daringfireball.net/projects/markdown/syntax.

22 JSON for Linking Data, http://json-ld.org/.

23 PROV-DM: The PROV Data Model, W3C Recommendation, April 30, 2013, http://www.w3.org/TR/prov-dm/.

24 To substantiate her argument, the scholar must also identify corroborating material, including instances of Istros' text in other primary sources—in this example that of Athenaeus, *Deipn.* 3.6, who does name Istros as the source.

25 http://perseids.org/sites/berti_demo/index.html (source code at https://github.com/PerseusDL/lci-demo).

26 Digital Fragmenta Historicorum Graecorum (DFHG) Project, Universität Leipzig, http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/digital-fragmenta-historicorum-graecorum-dfhg-project/.

27 EpiDoc: Epigraphic Documents in TEI XML, http://sourceforge.net/p/epidoc/wiki/Home/.

28 http://www.eagle-network.eu/.

29 http://catalog.perseus.org/.

30 There is little room or intention in this article to elaborate further on the topic. Let it suffice to say this endeavor has contributed to the new release of the schema (published in February 2014) which now allows, among other additions, the following: (1) the `@rend` attribute of `<hi>` can now have `"stacked"` as a value to encode stacked symbols or characters (such features are often found in manuscripts); (2) the `<anchor>` element is no longer omitted. All this has been explained in version 8.20 of the EpiDoc Guidelines (Tom Elliott, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt, et al., *EpiDoc Guidelines: Ancient Documents in TEI XML*, December 4, 2014, http://www.stoa.org/epidoc/gl/latest/).

31 http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/digital-marmor-parium/

32 http://pleiades.stoa.org/.

33 http://snapdrgn.net/.

34 http://timeline.knightlab.com/.

35 https://archive.org/details/fragmentahistori01mueluoft.

36 http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/.

37 http://www.w3.org/TR/xslt.

38 See, for example, Müller 1878–85, 1:1, frr. 5 and 7; 1:56, fr. 83; 2:29, fr. 2.

39 For example, Κάδμος Ἀρχελάου Μιλήσιος, where we have the name of the author (Κάδμος), the name of the father (Ἀρχελάου), and the ethnicon (Μιλήσιος). Further details on, and matching of, name components will be part of a later, spin-off prosopography project.

40 GeoNames geographical database, http://www.geonames.org/.

41 https://github.com/OpenGreekAndLatin/dfhg-dev/wiki.

42 https://creativecommons.org/licenses/by-sa/4.0/.

ABSTRACT

This paper presents a joint project of the Humboldt Chair of Digital Humanities at the University of Leipzig, the Perseus Digital Library at Tufts University, and the Harvard Center for Hellenic Studies to produce a new open series of Greek and Latin fragmentary authors. Such authors are lost and their works are preserved only thanks to quotations and text reuses in later texts. The project is undertaking two tasks: (1) the digitization of paper editions of fragmentary works with links to the source texts from which the fragments have been extracted; (2) the production of born-digital editions of fragmentary works. The ultimate goals are the creation of open, linked, machine-actionable texts for the study and advancement of the field of Classical textual fragmentary heritage and the development of a collaborative environment for crowdsourced annotations. These goals are being achieved by implementing the Perseids Platform and by encoding the *Fragmenta Historicorum Graecorum*, one of the most important and comprehensive collections of fragmentary authors.

INDEX

Keywords: fragmentary text, digital edition, Classics, CTS, SAM, PROV, EpiDoc

AUTHORS

MONICA BERTI

Monica Berti is Assistant Professor at the Humboldt Chair of Digital Humanities at the University of Leipzig, where she teaches courses on digital philology. Since 2008 she has been working with the Perseus Digital Library at Tufts University, where she has been visiting professor in the Department of Classics. Her current research work is focused on representing quotations and text reuses of ancient lost works.

BRIDGET ALMAS

Bridget Almas is the lead software developer for the Perseus Digital Library at Tufts University. She has been working in software development since 1994, and for Perseus since 2010. She also currently serves as a member of the Technical Advisory Board of the Research Data Alliance.

DAVID DUBIN

David Dubin is a research associate professor at the University of Illinois Graduate School of Library and Information Science. His research areas are the foundations of information representation and description, and issues of expression and encoding in documents and digital information resources.

GRETA FRANZINI

Greta Franzini completed her Classics BA and Digital Humanities MA degrees at King's College London. Greta is currently doing a PhD at the UCL Centre for Digital Humanities, where she conducts research in the fields of Classical Philology, Manuscript Studies and Literary Criticism. She is also currently working as a Researcher at the Göttingen Centre for Digital Humanities. Previously, she worked as a Research Associate at the Humboldt Chair of Digital Humanities at the University of Leipzig.

SIMONA STOYANOVA

Simona Stoyanova is a classicist who specializes in epigraphy, and a digital humanist. She works as a research associate for the Open Philology Project at the University of Leipzig. She is also a PhD candidate in Digital Classics at King's College London. Her research is focused on the Greek and Latin epigraphic traditions in the mixed-language population of the province of Thrace, with particular interest on palaeographical issues and their possible investigation through the DigiPal framework.

GREGORY R. CRANE

Gregory R. Crane is the editor-in-chief of the Perseus Project at Tufts University, where he is professor of Classics and holds the Winnick Family Chair of Technology and Entrepreneurship. He is also Alexander von Humboldt Professor of Digital Humanities at the University of Leipzig.