3 Distributed Annotations of Fragmentary Texts

This chapter is devoted to distributed annotations of fragmentary texts in a digital environment. The first section (3.1) shows and explains the new idea of conceiving fragmentary texts as annotations of textual elements about reused authors and works. The second section (3.2) describes the *CITE Architecture*, which is a protocol for producing canonical citations, and its application to text reuse. The third section (3.3) describes an experiment developed by the *Perseus Project* for implementing a fragmentary texts editor.

3.1 Annotations of Reused Authors and Works

One of the main concerns when raising evidence of lost works is to reconstruct the complex relationship between the fragment and its source of transmission. This means weighing the level of interference played by the author who has reused and transformed the original context of the fragment — measuring the distance between the source text and the derived text — and trying to perceive the degree of text reuse and its effects on the resulting target text.[1] This interpretative process is usually explained in the commentary of an edition of fragmentary texts or in papers and monographs pertaining to various aspects of fragmentary authors and works, but is completely lost in the printed representation of the fragments, which are simply typographical reproductions of extracts of derived texts.[2]

As I described in section 2.2, textual fragments are a form of hypertext. In this respect, a digital environment offers the possibility to represent *fragments* as *text reuses* within their context of transmission, pointing directly to the elements that are traces of a text reuse and going beyond the problem of extracting and decontextualizing extended chunks of texts that preserve quotations and reuses of other texts.[3] A straightforward example is a quotation of the lost author Hecataeus of Miletus by Athenaeus of Naucratis in the *Deipnosphists*:

---

1   Lee (2007) 472.
2   See section 1.4.
3   On the difficulty of defining borders and "boundaries of the open text" in a hypertext, which is a "borderless electronic text," see Landow (2006) 112–118.

> *Deipn.* 10.67 = 447c: Ἑκαταῖος δ᾽ ἐν δευτέρῳ Περιηγήσεως εἰπὼν
> περὶ Αἰγυπτίων ὡς ἀρτοφάγοι εἰσὶν ἐπιφέρει· τὰς κριθὰς ἐς τὸ πῶμα
> καταλέουσιν. ἐν δὲ τῇ τῆς Εὐρώπης Περιόδῳ Παίονάς φησι πίνειν
> βρῦτον ἀπὸ τῶν κριθῶν καὶ παραβίην ἀπὸ κέγχρου καὶ κονύζης·
> ἀλείφονται δέ, φησίν, ἐλαίῳ ἀπὸ γάλακτος. καὶ ταῦτα μὲν ταύτῃ.[4]

In this passage the Naucratites quotes two works of Hecataeus that are now lost and adds a few pieces of information about the source texts: The name of the fragmentary author (Ἑκαταῖος), the titles of two fragmentary works (Περιήγησις and τῆς Εὐρώπης Περίοδος) with book number (δεύτερος) in the first case, and different text reuses, which are introduced by *verba dicendi* (εἰπών, ἐπιφέρει, φησι and φησίν) and seem to be partly paraphrases or summaries and partly *verbatim* repetitions of the original words of the lost works. This passage is usually split into two different chunks of text corresponding to two fragments of Hecataeus: see FHG I 8 fr. 123 and 20 fr. 290; FGrHist 1 FF 154 and 323a; BNJ 1 FF 154 and 323a.

If the concept of hypertext is nowadays taken for granted and if the annotations of the elements shown above seems to be an easy process, in fact a proper representation of all the components belonging to the domain of historical text reuse requires a complex infrastructure with layers of annotations and tagsets.[5] These annotations include not only the portion of text that can be considered a reuse, but also many pieces of information like names and geographic provenance of reused authors with variants, titles and/or descriptions of reused works, *verba dicendi*, expressions of literary criticism, and other linguistic and morphosyntactic features.

Building a digital library of text reuses of fragmentary authors means, first of all, to select the string of words that belong to the portion of text which is classifiable as a reuse and, secondly, to encode all those elements that signal the presence of the text reuse itself (named entities, grammar, syntax, etc.).

The next step is to align and encode all information pertaining to other witnesses that reuse the same original text with different words and a different syntax, parallel texts that deal with the same topic of the text reuse, and finally different editions and translations of both the source and the derived texts.[6]

---

4   "Hecataeus in the Book II of the *Tour* reports that the Egyptians eat bread, and then continues: They grind up barley to produce the substance they drink. And in his *Journey through Europe* he says that the Paeonians drink *brutos* made from barley, and *parabiē* made from millet and fleabane; and they smear themselves, he claims, with oil made from milk. So much for these topics." Trans. by Olson (2006–2012).

5   On the fast development of computing in the humanities in the last twenty years and on the consequent implicit ingestion of the theory of hypertext, see Landow (2006) xi–xiv. For an introduction to annotation and its literary, scholarly, civic, and everyday significance across historical and contemporary contexts, see Kalir/Garcia (2021).

6   Almas/Berti (2013) 1; Berti/Romanello et al. (2009).

3.2 The CITE Architecture

Before addressing methods and strategies for annotating historical fragmentary texts, it is necessary to introduce the *CITE Architecture*, which is a framework for producing machine actionable citations of texts including quotations and text reuses of extant and lost sources. CITE is a digital library architecture that stands for "Collections, Indices, Texts, and Extensions."[7] This architecture was originally developed for the *Homer Multitext* (HMT), which is a project of the Harvard's Center for Hellenic Studies to digitally represent language, structure, and manuscript tradition of Greek epic poetry.[8] The goal of the architecture is to provide a framework for identifying and retrieving machine actionable *citations* of texts and other data that are the basis of scholarly publications in the humanities.[9] The architecture makes use of the international standard of "Uniform Resource Names" (URNs). This standard allows to make unique, complete, precise, and machine actionable scholarly citations. Given that URNs can be hierarchical, the architecture gives the possibility to cite at different levels of granularity.[10]

The architecture is based on two main *data models*: 1) the standard CTS URN for identifying and retrieving *texts* and *passages of texts*, and 2) the standard CITE URN for identifying and retrieving *discrete objects* and *other data* (e.g., manuscript folios, images, syntactic structure, metrical feet, text reuse, etc.).[11] The CTS URN is based on the assumption that a text can be modelled as "an ordered hierarchy of citation objects" (OHCO^2) and defines "a citable text as a set of citable nodes" that belong to a bibliographic hierarchy and to a citation hierarchy, and that are ordered.[12] Implementations of the CTS URN have been developed by different projects, such as the *Perseus Digital Library*, *Open Greek and Latin*, the *First 1000 Years of Greek*, the *Scaife Viewer*, the *Canonical Text Service* of the University of Leipzig and *CapiTainS*.[13] An example of a scholarly citation that can be digitally represented is "Homer, *Iliad*, edition of Wolf (1804), Book 1, line 1." This citation refers to the first line of the *Iliad* of Homer in the 1804 edition by Friedrich August

---

7   http://cite-architecture.org
8   Dué/Ebbott (2009); Dué/Ebbott (2019).
9   Smith (2009).
10 Smith/Blackwell (2012); Blackwell/Smith (2016); Blackwell/Smith (2019).
11 CTS stands for "Canonical Text Services."
12 The acronym OHCO derives from DeRose/Durand et al. (1990), who argued that a text is an "ordered hierarchy of content objects." In Renear/Mylonas et al. (1996), the authors thought again their model and recognized that there are overlapping hierarchies when taking different perspectives on the content of a text. Smith/Weaver (2009) observed that a better functional definition of text has the "citation object" as its fundamental unit, and so proposed OHCO^2. CTS is built on this definition. See Smith/Blackwell (2012), Blackwell/Smith (2016) 3, and Berti/Blackwell et al. (2016) 124.
13 Crane/Almas et al. (2014); Tiepmar/Teichmann et al. (2014); Tiepmar (2018); Babeu (2019); Muellener (2019); Tiepmar/Heyer (2019).

Wolf and is valid both if a scholar reads a printed text of the book or a digital version of it in *Google Books*, *Internet Archive*, or *HathiTrust*:

Μῆνιν ἄειδε, θεά, Πηληϊάδεω Ἀχιλῆος,

According to the *CITE Architecture*, it is possible to convert this citation into a machine actionable format:

`urn:cts:greekLit:tlg0012.tlg001.wolf1:1.1`

This CTS URN represents different levels of the citation: The CTS *domain* (`urn:cts:`) which is required in the URN syntax, the *namespace* `greekLit` that identifies works in ancient Greek, and a *hierarchy* that identifies the work and the edition of the work. In this case the hierarchy is analogous to that of the *Functional Requirements for Bibliographic Records* (FRBR).[14] Within this hierarchy, texts exist in a text-group (in this case the Homeric poetry) and a text-group contains one or more works (in this case the *Iliad*). Identifiers include numbers of the *Thesaurus Linguae Graecae* (TLG), where Homer is `tlg0012` and the *Iliad* is `tlg001`.[15] `wolf1` is a reference to the 1804 edition of the *Iliad* by the German scholar Friedrich August Wolf. After the colon, `1.1` refers to book 1, line 1. It is also possible to expand the citation to ranges of passages (e.g., `1.1-1.2`) and add a substring to cite a word of the text (e.g., `1.1@μῆνιν[1]`).

    `urn:cts:greekLit:tlg0012.tlg001` is the *work-level* identifier, where the *Iliad* is considered as an abstraction (notional work) that includes every edition and translation of the work.

    `urn:cts:greekLit:tlg0012.tlg001.wolf1` is the "version-level" identifier (in this case the 1804 printed edition of the *Iliad* by Wolf).

    `urn:cts:greekLit:tlg0012.tlg001.wolf1:1.1@μῆνιν[1]` is the passage component that also points to words (in this case the first occurrence of the string μῆνιν in book 1, line 1 of the edition of the *Iliad* by Wolf).

    The CITE component of the architecture allows to identify and retrieve citations of discrete objects and other data like physical manuscript folios of the *Homer Multitext* project.[16] For example, `urn:cite:hmt:msA.msA–12r` uniquely identifies a single object in the *Homer Multitext* collection, which is folio 12 *recto* of the *Venetus A* manuscript. The *CITE Architecture* offers also the possibility to cite other data, such as lexical tokens, metrical feet, syntax, and fragmentary texts.[17] As I described in chapter 1, there are two main kinds of fragmentary texts: *fragmentary texts of still extant works* and *fragmentary texts of lost works*. The first kind of fragmentary text is citable as an alignment of CTS URNs. Following the examples described in chapter 1 (p. 10), the reuse of a passage of the *Histories*

---

14   For the use of FRBR in the *Perseus Catalog*, see p. 26.
15   On TLG numbers of authors and works, see p. 19 n. 55.
16   Smith/Blackwell (2012).
17   Berti/Blackwell et al. (2016); Blackwell/Smith (2016); Blackwell/Smith (2019); Blackwell/Smith (2020).

of Herodotus (6.127.1) in the *Deipnosophists* of Athenaeus (12.58 = 541bc) can be documented as:

```
urn:cts:greekLit:tlg0008.tlg001:12.58@ἀπὸ[1]-12.58@ἀφίκετο[1]
```
quotes
```
urn:cts:greekLit:tlg0016.tlg002:6.127.1@ἀπὸ[1]-6.127.1@ἀπίκετο[1]
```

This alignment means that the string ἀπὸ μὲν Ἰταλίης Σμινδυρίδης ὁ Ἱπποκρά-τεος Συβαρίτης, ὃς ἐπὶ πλεῖστον δὴ χλιδῆς εἷς ἀνὴρ ἀφίκετο of Ath., *Deipn.* 12.58 is a *quotation* of the string ἀπὸ μὲν δὴ Ἰταλίης ἦλθε Σμινδυρίδης ὁ Ἱπποκράτεος Συβαρίτης, ὃς ἐπὶ πλεῖστον δὴ χλιδῆς εἷς ἀνὴρ ἀπίκετο of Hdt. 6.127.1.[18]

The second kind of fragmentary text (*lost text*) is citable as an alignment of CTS and CITE URNs. For example, Ath., *Deipn.* 3.6 (= 74e) is the only citable evidence of a passage of the work of Istros the Callimachean that is now lost: 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 συ-κοφάνται.[19] In this case the text of the *Deipnosophists* can't be aligned with any other texts because the original work of Istros is lost. Moreover, in order to avoid the duplication of the text of Athenaeus by extracting from it the chunk of text with the reuse of Istros (as it happens in printed collections of fragmentary texts), the *CITE Architecture* offers the possibility to represent the reuse within the text of Athenaeus:

```
urn:cite:lofts:berti.istros12
```
quotes
```
urn:cts:greekLit:tlg0008.tlg001:3.6@Ιστρος[1]-3.6@συκοφάνται[1]
```

This alignment means that a passage of Ath., *Deipn.* 3.6 has been *quoted* and classified as fragment 12 (`istros12`) of Istros the Callimachean in the critical edition of the scholar Berti.[20] The CITE URN represents the level of the *edition*, while the CTS URN represents a string of *text* that is cited for specific purposes. The *CITE Architecture* allows different editors to cite and classify strings of text in different ways. For example, the same or a shorter or longer string of text referring to Istros in Ath., *Deipn.* 3.6 could be cited and classified by different editors and with different numbers in different editions of the fragmentary author Istros.

CITE URNs can be used for many other kinds of *citable analyses* within the domain of fragmentary literature. In this case the *data model* developed by the

---

18 On the use of the Kaibel reference system in the CTS URNs of the *Deipnosphists*, see section 5.4.

19 "Istrus in his *Attic History* says that the dried figs produced by these trees were not exported from Attica, in order that only the inhabitants of the country could enjoy them. When many people were found to be evading the law, those who informed the jurors about them were then for the first time referred to as sycophants." Trans. by Olson (2006–2012).

20 Berti (2009b) 99–102.

CITE Architecture has been experimented with the *Deipnosophists* of Athenaeus of Naucratis and specifies five *subjects of analyses* with *properties*:[21]

1. **Authors** (properties)
   - *EntityUrn*: a CITE URN that uniquely identifies a reused author. An example is `urn:cite:digAth:authors.auth3` that identifies "Archilochus" and is unique for all occurrences of his name in the *Deipnosophists*.[22]
   - *CtsUrn*: a CTS URN that identifies a passage where the author is mentioned. `urn:cts:greekLit:tlg0008.tlg001:1.2@Ἀρχίλοχον[1]` is one of the passages of the *Deipnosophists* where Archilochus is mentioned. This passage serves to justify the author's inclusion in the list.[23] When an author is reused often, the passage here should be a clear, unambiguous reference (e.g., "Homer says […]").
   - *Label*: a human-readable name for the author. E.g., "Archilochus of Paros."
   - *OptionalCtsGroupUrn*: a group-level CTS URN that identifies still extant authors. E.g., `urn:cts:greekLit:tlg0012:` (Homer).

2. **Works** (properties)
   - *EntityUrn*: a CITE URN that uniquely identifies a reused work. For example, `urn:cite:digAth:works.work1` identifies the gastronomic work by Archestratus of Syracuse or Gela.[24]
   - *AuthorUrn*: a CITE URN that uniquely identifies a reused author and corresponds to an author cataloged in **Authors**. `urn:cite:digAth:authors.auth20` identifies Archestratus of Syracuse or Gela, who was the author of a gastronomic work. Athenaeus writes that Archestratus was from Syracuse

---

21 The *Deipnosophists* is a work full of many different kinds of quotations and reuses of other texts and this is the reason why it fits well with experiments for producing annotations of fragmentary texts in their context. On the *Digital Athenaeus* project and on the content of the *Deipnosophists*, see chapter 5. The *data model* of the *CITE Architecture* has been developed in collaboration with D. Neel Smith and Christopher W. Blackwell and was presented at the international conference *Digital Humanities 2016*: see Berti/Daniels et al. (2016) and Berti/Blackwell et al. (2016) (texts and examples mentioned in the following pages are taken from these papers).

22 In the syntax of the CITE URN, `auth3` represents a sequence number in an *ordered collection*. Each item has a sequence number that reflects the item's sequence in the text of the *Deipnosophists*. This value is programmatically generated by a CTS-aware script before publication of the collection: Berti/Blackwell et al. (2016). Given that text reuses in the same text can be identified and cited in different ways by different scholars, sequence numbers may differ in different *collections*. Examples provided here do not correspond to a complete analysis of the *Deipnosophists*, but to a first theoretical experiment with the text of Athenaeus.

23 This example is interesting because the citation of Archilochus is part of a passage of the *Deipnosophists* that mentions "Archilochus' successors" (τῶν μετ᾿ Ἀρχίλοχον ποιητῶν) and therefore includes two references: one to Archilochus and the other to the poets who came after him. The *CITE Architecture* allows to represent both references with different *citable analyses* of the same text.

24 For the syntax of the CITE URN that includes a sequence number for the items in the collection, see n. 22.

or Gela (Ἀρχέστρατος ὁ Συρακούσιος ἢ Γελῷος). Like for ancient titles (see n. 26), also places of origin of ancient authors are often uncertain and sources reflect these uncertainties. The *CITE Architecture* allows to cite the two traditions about the ethnic origin of Archestratus by generating two *citable analyses* that can be aligned to the same CITE URN that identifies the author (see below the *subject* **Mentions**). In this case the *EntityURN* `urn:cite:digAth:authors.auth20` will correspond to the *CtsUrn* `urn:cts:greekLit:tlg0008.tlg001:1.7@Ἀρχέστρατος[1]-1.7@Γελῷος[1]`, whose textual content corresponds to both Ἀρχέστρατος ὁ Συρακούσιος and Ἀρχέστρατος ὁ Γελῷος.[25]

- *CtsUrn*: a CTS URN that identifies a passage in Athenaeus where the work is mentioned. E.g., `urn:cts:greekLit:tlg0008.tlg001:1.7@ὅτι[1]-1.7@φησί[1]` identifies the passage of the *Deipnosophists* where the gastronomic work of Archestratus is mentioned.[26] This passage serves to justify the work's inclusion in the list. When a work is reused often, the passage here should be a clear, unambiguous reference (e.g., "Eupolis says, in the *Demoi* [...]").
- *Label*: a human-readable name for the work. E.g., "*Demoi.*"
- *OptionalCtsWorkUrn*: a work-level CTS URN that identifies still extant works. E.g., `urn:cts:greekLit:tlg0012.tlg001`: (*Iliad*).[27]

3. **Characters** (properties)
    - *EntityUrn*: a CITE URN that uniquely identifies sophists (characters) who take part in the banquet described by Athenaeus in the *Deipnosphists*.[28] For example, `urn:cite:digAth:characters:character1` identifies Aemilianus Maurus.
    - *CtsUrn*: a CTS URN that identifies a passage in Athenaeus where the character is mentioned. `urn:cts:greekLit:tlg0008.tlg001:3.25@Αἰμιλιανὸς[1]` is an example of one of the passages of the *Deipnosophists* where Aemilianus is mentioned and talks. This passage serves to justify the character's inclusion in the list; when a character is reused often, the passage here should be a clear, unambiguous reference (e.g., "Ulpianus says [...]").
    - *Label*: a human-readable name for the character. E.g., "Aemilianus Maurus."
    - *OptionalCtsGroupUrn*: for characters who were authors and whose texts are

---

25 For the annotation of this example with INCEpTION, see section 5.6.3.

26 The work of Archestratus is one of the many examples of ancient Greek literature with different titles in the tradition. In the passage of the *Deipnosophists* cited here (1.7 = 4e) Athenaeus testifies that this work was entitled *Gastronomy* (Γαστρονομία) according to Chrysippus, *Life of pleasure* (Ἡδυπάθεια) according to Lynceus and Callimachus, *Science of dining* (Δειπνολογία) according to Clearchus, and *Art of cooking* (Ὀψοποιία) according to others. The *CITE Architecture* allows to cite all these four titles and align them to a unique identifier that represents them, because they are different expressions of the same work.

27 An interesting example are names for specific books of the *Iliad*, which can have more precise CTS URNs in the *OptionalCtsWorkUrn* field: e.g., `urn:cts:greekLit:tlg0012.tlg001:18` = *Making of arms* (Ὁπλοποιία), which is the name of the eighteenth book of the *Iliad*.

28 This is a specific *subject of analysis* for the *Deipnosophists*, where the author describes a group of twenty-two learned men (sophists) who take part in the banquet described in the book and who cite many texts of ancient literature: see section 5.5.3.

still extant, a group-level CTS URN that identifies the character. For example, `urn:cts:greekLit:tlg0057:` (Galenus of Pergamum).

4. **Mentions** (properties)
   – *EntityUrn*: a CITE URN that uniquely identifies a reference in the text of the *Deipnosophists* of Athenaeus. E.g., `urn:cite:digAth:mentions.1` identifies Archestratus of Syracuse.[29]
   – *CiteUrn*: a CITE URN (from **Authors** or **Works** above) that identifies the author or work mentioned in Athenaeus. `urn:cite:digAth:authors.auth20` identifies Archestratus in the list of **Authors**.[30]
   – *CtsUrn*: a CTS URN that specifies a passage in Athenaeus that mentions the author like `urn:cts:greekLit:tlg0008.tlg001:1.7@Ἀρχέστρατος[1]–1.7@Γελῷος[1]`.[31]
   – *Text*: the relevant textual content of the passage specified by the CTS URN (above). E.g., Ἀρχέστρατος ὁ Συρακούσιος.
   – *Notes*: human-readable notes. May be empty.

5. **Reuses** (properties)
   – *EntityUrn*: a CITE URN that uniquely identifies an instance of text reuse in the *Deipnosophists* of Athenaeus. For example, `urn:cite:digAth:reuse.1` identifies a reuse of a lost verse of Eratosthenes of Cyrene.
   – *CiteUrn*: a CITE URN (from **Authors** or **Works** above) that identifies the author or work mentioned in Athenaeus. `urn:cite:digAth:authors.auth13` identifies Eratosthenes of Cyrene.[32]
   – *CtsUrn*: a CTS URN that specifies a passage in Athenaeus containing the text reuse. This should be a range that includes language marking the passage as text reuse (*verbum dicendi*, etc.). E.g., `urn:cts:greekLit:tlg0008.tlg001:1.3@τρὶς[1]–1.3@ἄμεινον[1]` is the passage with the verse of Eratosthenes.[33]
   – *TextContent*: a string that contains the precise textual content (from Athenaeus) that is reused; this will exclude *verba dicendi*, etc. For example,

---

29 As we have seen before, Archestratus has two possible places of origin and the *CITE Architecture* cites both. In this case another CITE URN refers to the possible origin of Archestratus from Gela (`urn:cite:digAth:mentions.2`). Both CITE URNs (**Mentions**) correspond to `urn:cite:digAth:authors.auth20` (**Authors**): see below.

30 Archestratus' work is transmitted with four different titles and therefore there are four different CITE URNs for **Mentions**: `urn:cite:digAth:mentions.3` (Γαστρονομία), `urn:cite:digAth:mentions.4` (Ἡδυπάθεία), `urn:cite:digAth:mentions.5` (Δειπνολογία) and `urn:cite:digAth:mentions.6` (Ὀψοποιία). All these CITE URNs correspond to `urn:cite:digAth:works.work1` (**Works**).

31 This passage refers to Archestratus with his two possible places of origin and the CTS URN is valid for both `urn:cite:digAth:mentions.1` (Ἀρχέστρατος ὁ Συρακούσιος) and `urn:cite:digAth:mentions.2` (Ἀρχέστρατος ὁ Γελῷος).

32 In this case the text of the *Deipnosophists* doesn't mention the name of Eratosthenes, but the expression "the Cyrenean poet" (ὁ Κυρηναῖος ποιητής) that has to be referred to Eratosthenes (cf. *Suda* [E 2898] s.v. Ἐρατοσθένης).

33 Fr. 30, p. 65 Powell.

"τρὶς δ᾽ ἀπομαξαμένοισι θεοὶ διδόασιν ἄμεινον" is the verse of Eratosthenes.

– *Analytical Edition URN* [may be empty; only for extant works]: a CTS URN that attaches the reused text (from Athenaeus) to the ordered, hierarchical citation scheme of the reused work. Where Athenaeus reuses text from extant works, which exist in other editions with citation schemes, we can produce an Analytical Edition of that work, the "Athenaeus Edition"; this edition can be cited by CTS URNs. For lost works, there is no citation scheme, nor any inherent order to the text. For these, we will produce a collection of text reuses. This Collection can be cited by CITE URNs.

– *Alignment URN* [may be empty; only for extant works]: a CTS URN that specifies text in another edition of the reused work, used to assert an explicit alignment between Athenaeus' language and the language of another text. For example, the use of βουλεύεσθαι at Ath., *Deipn.* 1.18 can be interpreted as an allusion to βουληφόρε at *Iliad* 20.83; this is not a literal string match, nor are the two instances of the same lexical entity; we are asserting an alignment that is not discoverable by any automated process; the alignment urn allows us to make this alignment explicit.

– *Commentary* [may be empty]: a commenary that explains the interpretation of the text reuse.

– *Resp.*: The editor responsible for asserting the existence of, and documenting, this instance of text reuse.

Further experiments with this *data model* have been performed for annotating reuses of Homeric poetry in the *Deipnosophists* of Athenaeus with examples from the *Iliad*, which means working with reuses of a still extant work.[34] In this case the data model developed by the *CITE Architecture* specifies six pieces of information (*records*) to document text reuse:[35]

1. **Analysis Record URN**. Every documented instance of text reuse has a CITE URN, that uniquely identifies this instance in a CITE collection. E.g., `urn:cite:opdata:ahri:100` (item 100 in the `ahri` collection [Athenaeus' Homeric Reuse: *Iliad*], in the `opdata` namespace [open philology data]).

2. **Sequence Number**. The collection of instances of Homeric text reuse is an ordered collection; each item has a sequence number, reflecting the item's sequence in the text of the *Deipnosophists*. This value is programmatically generated by a CTS-aware script before publication of the collection.

3. **Analysed Text URN**. A CTS URN that indentifies (as precisely or imprecisely as necessary) the span of text in the *Deipnosophists* that is the subject

---

34 As part of these experiments, citable analyses of text reuse of Homer's *Iliad* in the *Deipnosophists* were collected by Ellie Daniel, Kimbell Dobbins, and Samantha Strickland from Furman University during their internship at the University of Leipzig in the Summer 2015 under the supervision of Christopher W. Blackwell and myself.

35 A detailed description of the data model with examples is available in Berti/Blackwell et al. (2016) 126–127, from which are taken the texts and the example mentioned in the following pages.

of this analysis of text reuse. The scope of the *Analysed Text* is determined by the nature of the text reuse.

4. **Reused Text**. While the *Analysed Text URN* (above) identifies a coherent and contiguous span of text, as it appears in the Edition being analysed, the *Reused Text* is a string that identifies only the text being reused. The *Analysed Text URN* provides context and a basis for alignment, while the *Reused Text* gives us the flexibility to call out non-contiguous text, to normalize text, or even to promote morphological forms determined by indirect statement to those appropriate for direct speech, without doing violence to our source-Edition.

5. **Alignment URN**. This collection documents reuse of Homeric poetry, for which there are extant editions with canonical citation. The *Alignment URN* is a CTS URN that points to one specific edition of the *Iliad* that (a) justifies the claim of text reuse, and (b) is the basis for attaching an Iliadic citation to this analysis. The *Perseus* edition of the *Iliad* of Homer (`urn:cts:greekLit:tlg0012.tlg001.perseus-grc:`) is used for the Alignment URNs (edition by Thomas W. Allen).

6. **Analytical Edition URN**. The collected instances of Iliadic text reuse in the *Deipnosophists* represent a new edition of the *Iliad*, whose text-content is based on the analysis of the project's edition of Athenaeus. The *Analytical Edition URN* is a CTS URN to an "Athenaeus Edition" of the *Iliad*; the citation-value is based on that of the alignment URN; the text-content of this edition is the reused text in Athenaeus. The *Analytical Edition* gives us an orthogonal view of the Homeric text reuse in Athenaeus; it allows us to navigate Athenaeus according to the OHCO^2 structure of the *Iliad*. We cite two notional editions in this field:
   – `urn:cts:greekLit:tlg0012.tlg001.ogl1:` is "Athenaeus' edition of the *Iliad*." We also cite, in at least one analysis,
   – `urn:cts:greekLit:tlg0012.tlg001.ogl2:` that is "Zenodotus' edition of the *Iliad*, according to Athenaeus." `ogl1` and `ogl2` are 1) an *Open Greek and Latin* edition of the "*Iliad* of Athenaeus," consisting of collected Iliadic language of the *Deipnosophists*, and 2) another *Open Greek and Latin* edition of the "*Iliad* of Zenodotus, according to Athenaeus," consisting of Iliadic language attributed to Zenodotus' edition in the *Deipnosophists*.[36]

---

36 For the example with a reference to the ancient scholar Zenodotus, see Ath., *Deipn.* 1.21 (= 12e–f) and Berti/Blackwell et al. (2016) 134–136 (for a detailed description) and 127.

An example of this data model is a passage in the first book of the *Deipnosophists* (1.18 = 11a) where Athenaeus discusses how Homer equates drunkeness with madness:

καὶ Ἀγαμέμνων δὲ λέγει που περὶ αὐτοῦ "ᴬἀλλ᾽ ἐπεὶ ἀασάμην φρεσὶ λευγαλέῃσι πιθήσας |ᴮἢ οἴνῳ μεθύων, ἤ μ᾽ ἔβλαψαν θεοὶ αὐτοί," εἰς τὴν αὐτὴν τιθεὶς πλάστιγγα τὴν μέθην τῇ μανίᾳ.[37]

The Homeric text under analysis is ἀλλ᾽ [...] αὐτοί, but the "analysed text" begins from καὶ Ἀγαμέμνων [...] because the introductory clause is the signal that Athenaeus is quoting from Homer.[38] The two lines of Homer generate two different records in the text reuse data: (A) and (B).

A (ἀλλ᾽ ἐπεὶ ἀασάμην φρεσὶ λευγαλέῃσι πιθήσας) is most straightforward because it is a direct quotation of Homer that matches established editions of the *Iliad*. It is possible to assign an *Analysis Record URN* and fill in the other data fields (fig. 3.1). The first line of poetry in this passage of Athenaeus is found *verbatim* in *Iliad* 9.119, when Agamemnon is expressing remorse for the quarrel with Achilles. This will be the 100th instance of text reuse in the collection of Iliadic text reuse in Athenaeus. This instance of reuse emerges when we analyse the passage that begins "And Agamemnon says, somewhere [...]." The specific reused text in the passage under analysis is ἀλλ᾽ [...] πιθήσας. This instance aligns with 9.119 in the *Perseus* edition of the *Iliad*. In the "Iliad according to Athenaeus' *Deipnosophists*," we can identify this text as 9.119, following the canonical citation of the poem.

Instance	Field	Value
A	Analysis Record URN	`urn:cite:opdata:ahri:100`
A	Sequence	`100`
A	Analyzed Text URN	`urn:cts:greekLit:tlg0008.tlg001.berti:1.18@καὶ[17]-1.18@αὐτοί[1]`
A	Reused Text	ἀλλ᾽ ἐπεὶ ἀασάμην φρεσὶ λευγαλέῃσι πιθήσας
A	Alignment URN	`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:9.119`
A	Analytical Edition URN	`urn:cts:greekLit:tlg0012.tlg001.ogl01:9.119`

Figure 3.1. Athen., *Deipn.* 1.18 (11a) – record (A)

B (ἢ οἴνῳ μεθύων, ἤ μ᾽ ἔβλαψαν θεοὶ αὐτοί) is more complicated because Athenaeus is quoting a line that does not appear in any (other) edition of the *Iliad*. It will share some data values with (A), but differ in oth-

---

37 "And Agamemnon says, somewhere, about himself, 'But since I acted foolishly, obeying my addled thoughts | either I was drunk with wine, or the gods themselves harmed me,' placing drunkenness on the same balance as insanity." Trans. Berti/Blackwell et al. (2016) 128.

38 `urn:cts:greekLit:tlg0008.tlg001.berti:1.18@καὶ[17]-1.18@αὐτοί[1]`. This urn refers to an ongoing digital version of the *Deipnosophists* that I have been working on (`berti`).

Instance	Field	Value
B	Analysis Record URN	`urn:cite:opdata:ahri:101`
B	Sequence	`101`
B	Analyzed Text URN	`urn:cts:greekLit:tlg0008.tlg001.berti:1.18@καὶ[17]-1.18@αὐτοί[1]`
B	Reused Text	ἢ οἴνῳ μεθύων ἤ μ᾽ ἔβλαψαν θεοὶ αὐτοί
B	Alignment URN	`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:9.119`
B	Analytical Edition URN	`urn:cts:greekLit:tlg0012.tlg001.ogl01:9.119a`

Figure 3.2. Athen., *Deipn.* 1.18 (11a) – record (B)

ers (fig. 3.2). In this editorial judgement, the dactylic hexameter text ἢ οἴνῳ μεθύων, ἤ μ᾽ ἔβλαψαν θεοὶ αὐτοί is an instance of Homeric text reuse. It is assigned its own Analysis Record URN. This finding is the result of the analysis of the text at `urn:cts:greekLit:tlg0008.tlg001.berti:1.18@καὶ[17]-1.18@αὐτοί[1]`. (B) shares the same *Analysed Text URN* as (A). The text of (B) is after that of (A) in the text of Athenaeus, so (B) has a *Sequence number* one higher than that of (A). The *Analysed Text* passage presents (A) and (B) as a natural sequence, rather than two quotations juxtaposed by Athenaeus. Because the Alignment URN locates (A) in the *Iliad*, and because the Analysed Text unites (A) and (B), it is possible to use `urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:9.119` as the Alignment URN for (B) as well. However, in this "Athenaeus Edition" of the *Iliad*, (B) is an additional citeable passage, which is identified as 9.119a.

3.3 Perseids Fragmentary Texts Editor

To annotate quotations and text reuses of fragmentary authors in digital source texts experiments have been performed within *Perseids*, which "offers a free and open online environment to produce collaborative data-driven editions of ancient documents."[39]

The work was developed in 2012 and 2013 as part of a collaboration between the Alexander von Humboldt Chair of Digital Humanities at the University of Leipzig and the *Perseus Digital Library*. The result is a demo version of a *Fragmentary Texts Editor* (FTE) within *Perseids*, whose aim was to produce a pro-

---

39   See http://www.perseids.org. The development of *Perseids* was inspired by the work of several pre-existing projects, such as the *Tufts Miscellany Collection* at Tisch Library at Tufts University, the *Homer Multitext* project, and *Papyri.info*. The *Son of SUDA OnLine* (SoSOL) application is at the core of *Perseids*. For more information, see Almas/Berti (2013), Berti/Almas et al. (2014–2015), Almas/Beaulieu (2016), Berti/Almas et al. (2016), and Almas (2017). From these publications derive texts and examples reproduced in this section.

totype of a dynamic representation of quotations and reuses of fragmentary texts to help scholars annotate information about fragmentary authors by providing a shared environment for multi-level annotations of text reuses of ancient works. The demo is available at http://pubs.perseids.org/berti_demo/src/index.html and the code at https://github.com/PerseusDL/lci-demo.[40]

Figure 3.3. *Perseids Fragmentary Texts Editor*: Istros F 12 Berti

The FTE demo uses methods of inline and stand-off markup to produce stable ways for identifying and annotating text reuse, including canonical citations, morpho-syntactic analyses, translation and text alignments. The FTE demo interface collects texts from the printed edition of the fragments of Istros the Callimachean that I published in Berti (2009b). In this section I focus on a passage of the *Deipnosophists* of Athenaeus (3.6 = 74e) that preserves a text reuse of Istros (F 12 Berti).[41] Figure 3.3 shows the interface of the FTE demo with different functions for visualizing fragmentary texts.[42] The left side of the demo is devoted to the source of the text reuse:

---

40   Almas/Berti (2013); Berti/Almas (2013); Berti/Almas et al. (2014–2015); Berti/Almas et al. (2016).

41   Berti (2009b) 99–102.

42   http://pubs.perseids.org/berti_demo/src/berti_demo.html#urn:cite:perseus:lci.2.1

- **Source Text**. This tab shows the passage of the source preserving the reuse according to different editions. In this case the text of *Deipn.* 3.6 is presented in two tabs according to the two editions by Charles B. Gulick and Georg Kaibel. The interface provides CTS URNs of Edition 1[43] and of Edition 2,[44] and a function for visualizing the entire TEI XML file and the full text of the two editions in separate windows.[45] The "Show/Hide Quote" hyperlink allows to highlight and hide in both editions the passage with the reference to Istros, providing also the CTS URNs of the passage in the two editions of the source text with a substring corresponding to the highlighted range of text.[46] The portion of text highlighted in yellow corresponds to the chunk of text of *Deipn.* 3.6 classified as fragment 12 in Berti (2009b).[47]
- **Witnesses**. This tab is for representing other sources that preserve the same or a similar text reuse of a fragmentary text. In the example described here the tab is not active because there are no other witnesses for this fragment of Istros.[48]
- **Parallel**. This tab allows to visualize parallel sources of the fragment of Istros, which means other sources about the same topic.[49]

On the right side of the interface it is possible to visualize information about the fragment annotated in the source text on the left side:

- **Lost Content Item**. This tab provides a short summary of the content of the fragment with information about its editor. Given that this is a reuse of a lost text, the fragment is cited with a CITE URN: `urn:cite:perseus:lci.2.1`. In the syntax of the CITE URN, `lci` stands for *Lost Content Items*, which is the collection of text reuses of lost texts in the FTE demo. Number `2` identifies the fragment in the collection. On the use

---

43 `urn:cts:greekLit:tlg0008.tlg001.perseus-grc1:3.6`

44 `urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:3.6`

45 This function is not anymore available in the demo. The goal of the interface was to include more than one edition of the same source text.

46 The CTS URN has a URL prefix in order to be part of a resolvable web address. In the CTS syntax the FTE demo still uses the symbol "#" instead of "@" for separating the subreference from the passage (e.g., `#Ιστρος1-συκοφάνται1`): see Almas/Berti (2013), n. 1.

47 The FTE was implemented to include more editions of the same fragment, which means representing different annotations of the same text reuse by different editors. For this function, cf. the *demo.fragmentarytexts.org* described in section 2.2.

48 Berti (2009b) 99. Berti Istros F4 in the FTE demo provides the text of *Suda* [Π 152] s.v. Παναθήναια as one of the "witnesses" of the fragment of Istros: https://pubs.perseids.org/berti_demo/src/index.html. In this case the fragment has another witness (Photius), but, being a demo, the FTE provides only the text of sources available in a digital format. Cf. Berti (2009b) 59.

49 In this case the two parallel sources in the FTE are Plut., *Sol.* 24.1 (whose text is visualized with its CTS URN) and *Suda* [Σ 1330] s.v. συκοφαντεῖν (with a link to the *Suda On Line* project). For other parallel sources of this fragment, see Berti (2009b) 99–102.

of CITE URNs for text reuses of lost sources, see section 3.2.

– **Translation**. This tab provides translations of the text of the fragment.[50]
– **Commentary**. This tab provides a commentary to the fragment. As for other annotations, the FTE demo was originally planned to include more commentaries about the same reuse and each of them was collected and identified in a *Perseus Collection of Commentaries on Lost Content Items*.[51]
– **Aligment**. This tab provides alignments of the text of the fragment with its witnesses and/or parallel texts.[52] Translation alignments in the demo were produced in the *Perseids* platform using the *Alpheios* Translation Alignment Editor.
– **Syntax**. This tab shows morpho-syntactic annotations of the text of the fragment. Text reuse works not only at a word level, but also at a syntactic one, because reusing a text means not only quoting and readapting words in a new context, but also reproducing syntactic features. In this case the goal is to produce annotations of text reuses with the *Alpheios* Treebank Editor in order to collect and detect different examples of syntactic reuses (e.g., reuse of different words with the same syntax and/or reuse of the same words with a different syntax).[53]
– **Links**. This tab provides links to printed editions of the source text and of the fragmentary author through available resources, as for example *Google Books* and *Internet Archive*, or by uploading PDF files in the FTE demo.[54]

The work behind the *Fragmentary Texts Editor* combines TEI XML files, the *Open Annotation Core* (OAC) data model, and the *CITE Architecture* to represent quotations and text reuses via *Resource Description Framework* (RDF) triples. All of the textual and data elements presented in the display are defined as OAC annotations made available to the display code in a JSON-LD data structure. The subject and object resources of these triples are resolved by Canonical Text and CITE Collection Services to the TEI XML and other source data in real time in order to produce new dynamic, data-driven representations of the aggregated information.[55]

---

50 In this case there are the English translation of the passage of the *Deipnosophists* from the edition by Charles D. Yonge and the Italian translation of the fragment of Istros from Berti (2009b).
51 `urn:cite:perseus:lcicomm`
52 In this case the alignments are with the two parallel texts of Plut., *Sol.* 24.1 and *Suda* [Σ 1330] s.v. συκοφαντεῖν.
53 On syntactic text reuse detection and for a visualization of syntactic annotation of Istros F 12 Berti, see http://demo.fragmentarytexts.org/en/istros.html.
54 All source texts, translations, commentaries and lost content item descriptions are retrieved at display time via asynchronous requests to remote services: Almas/Berti (2013).
55 Almas/Berti (2013); Berti/Almas (2013); Berti/Almas et al. (2014–2015); Berti/Almas et al. (2016).

As I showed in section 3.2, a quotation of a still surviving text can be represented with a RDF triple: **[subject cts-urn-1]** *quotes* **[object cts-urn-2]**. For example, I represent the annotation of a quotation of Homer in Athenaeus as:

urn:cts:greekLit:tlg0008.tlg001:3.X.x

(Athen., *Deipn.* passage X.x)

quotes

urn:cts:tlg0012.tlg001:X.xx

(Hom., *Il.* passage X.xx)

When working with text reuses of lost works the situation is different, because the original text of the reused author is lost and we have just the text of the reusing author, which is the only citable evidence. For this reason, a *Perseus Collection of Lost Content Items* (urn:cite:perseus:lci) was created as part of the FTE. These LCIs are assigned CITE URNs as unique identifiers, and assigned descriptive properties, for example naming a specific text reuse of a lost author as it is represented in a modern edition because we don't have the original text of the lost author and we have to express the citation at an edition-level. In our example (Athen., *Deipn.* 3.6), the annotation triple is represented in the following way:

urn:cite:perseus:lci.2.1[56]

quotes

urn:cts:greekLit:tlg0008.tlg001:3.6#Ιστρος[1]-συκοφάνται[1][57]

This triplet expresses the relation between an object in a CITE Collection (an edition of a fragment of Istros) and a passage of a text (the *Deipnosophists* of Athenaeus who quotes Istros).

Annotations, and the texts and entities that they annotate, are the primary data type behind the FTE demo. The demo combines the TEI XML (EpiDoc subset) in which the source texts are encoded, with the CTS and CITE data models for URN based text and data object identifiers, the CTS and CITE service APIs, and the OAC standard for serialization of annotations.[58] This application of standards and data enables to present a new dynamic data-driven display leveraging *Linked Open Data* and also to publish annotation data in a standard format to facilitate its reuse.

The use of the OAC model enables to express FTE annotations according to a defined and documented standard, increasing the feasibility of their reuse. Using the OAC data model, annotations are expressed as simple URI based triples,

---

56   This is the CITE URN identifier for the *Perseus Collection Object* that represents the text reuse of Istros with a reference to the edition of Berti (2009b), where this portion of Athenaeus' text is reproduced and classified as Istros F12.

57   This is the CTS URN identifier for Athen., *Deipn.* 3.6 with the addition of substring reference for greater precision.

58   On the *Perseus* CTS API, see Almas/Berti (2013).

with a controlled vocabulary to identify the motivation for the annotation. According to OAC, an annotation "target" is the resource being annotated and the annotation "body" is the resource containing the contents of the annotation. The URIs used for annotation bodies and targets can resolve to anything from simple text strings and vocabulary terms, to complex morpho-syntactic annotations. OAC also supports many-to-many relationships between annotation targets and annotation bodies. This is particularly useful for text reuse annotations, where the text being reused (and/or the instance of its reuse) cannot be expressed by a single contiguous range of text and instead is surrounded by words which are not explicitly part of the reuse. In this case, we can use multiple CTS URN identifiers for the substrings within the passage, the set of which become the target and/or body of the annotation.

The primary set of annotations driving the demo links the passages from the extant source text to the lost content item. These annotations identify the URI of the extant source text in which a reuse occurs as the *target* of the annotation and the URI of the CITE object representing the lost content item as the *body* of the annotation. I use the OAC vocabulary term *classifying* to define the motivation for these annotations, as we are classifying the passage in the extant source text as an occurrence of text reuse. By contrast, my commentary annotations reference the URI for the lost content item itself as the annotation target, and the URI for the commentary as the annotation body. Translations of source texts reference the URIs for the source text passages as their targets, and the URIs of the translated passages as their bodies.

The OAC vocabulary term chosen for the motivation in this case is *linking*. I link additional supporting resources, including other witnesses, translation alignments and morpho-syntactic annotations in a similar manner. The OA model enables 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 a standard vocabulary for categorizing the motivation for the annotations. URIs are used to specify both the target and the body of the annotation.

The OA data model was used 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 can be then referenced by their URI identifiers. In the latter case, a variety of standard formats was used for the actual annotation bodies, including the *Perseus Ancient Greek and Latin Treebank* schema for morpho-syntactic analyses, the *Alpheios* translation alignment schema for text alignments, Mark-

down Syntax for short textual commentaries, TEI XML for primary and secondary source texts.[59]

Using the JSON-LD syntax recommended by OAC allows to build a dynamic display interface in Javascript that navigates the JSON-LD data object and retrieves the datasets identified as the targets and bodies of the annotations at their addressable URIs. The demo code retrieves the resources that are identified by CTS and CITE URN enabled URIs (as served by the CTS and CITE services discussed above) asynchronously as the page loads and in response to user interaction with interface widgets, and uses XSLT stylesheets to transform the XML content of the resources returned to HTML for display. The non CTS and CITE enabled resources are served by various other web applications, presenting various formats of data, and, due to time constraints, the demo currently presents these resources as links which open the original resource in a new tab or window. The annotation that represents the assertion according to which a 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 format, might be formalized as follows:[60]

```
{
    "@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"
                    } ,
    "has Body": "http://data.perseus.org/collections/urn:cite:
perseus:lci.2",
    "has Target": "http://data.perseus.org/citations/urn:cts:
greekLit:tlg0008.tlg001.perseus-grc1:3.6@Ἰστρος[1]-3.6@
συκοφάνται[1]",
    "oa:motivatedBy": "oa:linking"
}
```

The *Perseids* platform has at its core the *Son of SUDA OnLine* (SoSOL) application, which is a Ruby on Rails application originally developed by *Papyri.info* that serves as front end for a Git repository of documents, metadata, and annotations. It includes a workflow engine that enables documents and data of different types

---

59   Berti/Almas et al. (2014–2015) 7.
60   Almas/Berti (2013).

to pass through flexible review and approval process.[61] The SoSOL application includes user interfaces for editing XML documents, metadata, and annotations (fig. 3.4).
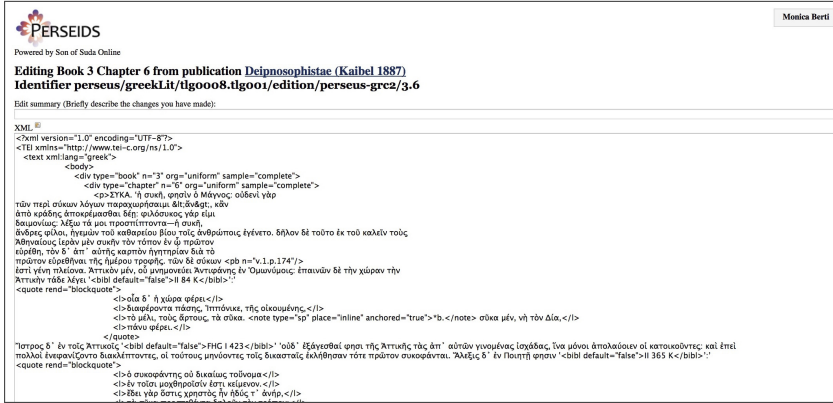

Figure 3.4. *Perseids* XML editing environment

When developing features of *Perseids* to support these workflows, the focus was first and foremost on the data. Considering that technologies change in a rapid way and while including prototype representations of digital editions suitable for publication on the web, the first priority was 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.

The following core requirements were identified 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.

As part of the FTE demo project and in order to represent the workflow of a scholar who identifies and edits a text reuse of a lost work, an experiment was done by applying and extending the *Systematic Assertion Model* (SAM), which is

---

61   Berti/Almas et al. (2014–2015) 3–4.

Figure 3.5. Complementary provenance models: text reuse workflow: Almas/Berti et al. (2013)
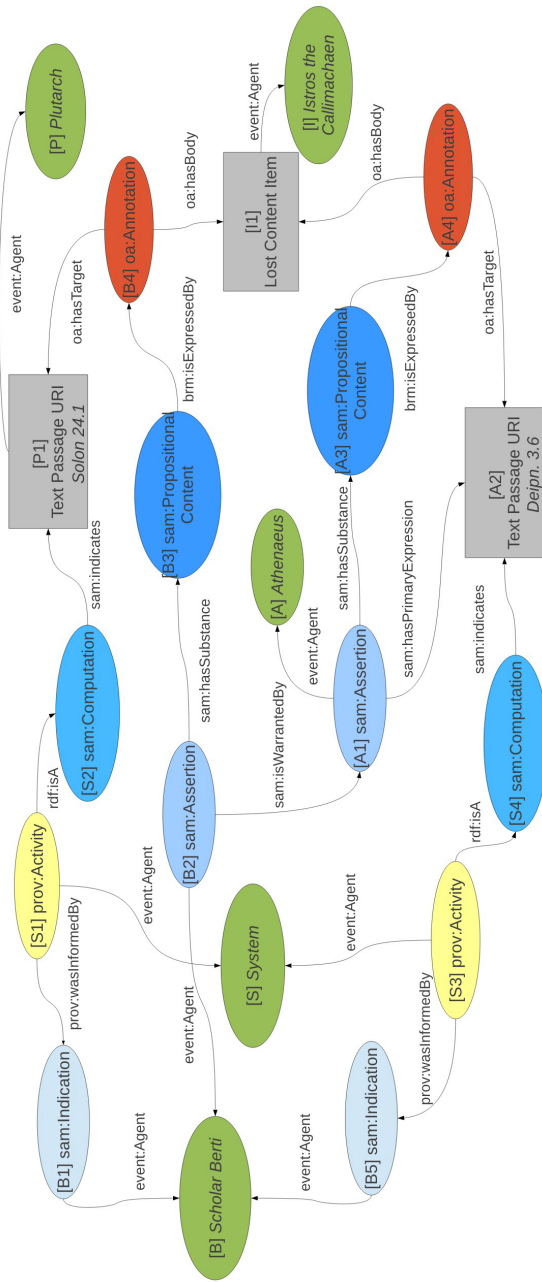
Summary: A scholar wants to annotate what she thinks is a quotation or text re-use of a lost text attributed to Istros in an extant source text by Plutarch (a reference to the origin of the word 'sycophants' where Istros is not named). To substantiate her argument, the scholar must also identify corroborating 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).

a framework developed at the University of Illinois for the description of provenance roles and agents essential to the identity of scientific data that accounts for the events and roles essential to the creation of text-like resources.[62]

The focus of many provenance models, such as the W3C PROV model, is on functional elements and processes of computational activities, while SAM supplies a detailed account of the particular ways symbol structures are used in scholarly data and discourse. The experiment was done by presentig an example text reuse scenario drawn from *Perseids*, demonstrating the use of a SAM-based RDF vocabulary extended to support textual research in the humanities, along with elements from PROV and the *Open Annotation* data model. This integrated account provides a rich, contextualized view of the encoding and use of data in humanities research (fig. 3.5).[63]

To test the extensions to the SAM framework, the project team modeled an example of a scholarly assertion of text reuse. In this use case, the scholar (Berti) identifies a section of text from Plutarch, *Solon* 24.1 that she believes was drawn from the lost work of Istros the Callimachean:[64]

> τῶν δὲ γινομένων διάθεσιν πρὸς ξένους ἐλαίου μόνον ἔδωκεν, ἄλλα
> δ' ἐξάγειν ἐκώλυσε· καὶ κατὰ τῶν ἐξαγόντων ἀρὰς τὸν ἄρχοντα ποι-
> εῖσθαι προσέταξεν, ἢ ἐκτίνειν αὐτὸν ἑκατὸν δραχμὰς εἰς τὸ δημό-
> σιον. καὶ πρῶτος ἄξων ἐστὶν ὁ τοῦτον περιέχων τὸν νόμον. οὐκ ἂν
> οὖν τις ἡγήσαιτο παντελῶς ἀπιθάνους τοὺς λέγοντας ὅτι καὶ σύκων
> ἐξαγωγὴ τὸ παλαιὸν ἀπείρητο, καὶ τὸ φαίνειν ἐνδεικνύμενον τοὺς
> ἐξάγοντας κληθῆναι συκοφαντεῖν.

The scholar wants to create an annotation that connects Plutarch's text back to Istros. To substantiate this argument, she identifies corroborating material from another primary source, which is Athenaeus' reference to Istros in the *Deipnosophists* 3.6:[65]

---

62 Wickett/Sacchi et al. (2013); Almas/Berti et al. (2013); Berti/Almas et al. (2014–2015) 7–9.

63 Almas/Berti et al. (2013); Berti/Almas et al. (2014–2015) 7–9.

64 "Of the products of the soil, he allowed oil only to be sold abroad, but forbade the exportation of others; and if any did so export, the archon was to pronounce curses upon them, or else himself pay a hundred drachmas into the public treasury. His first table is the one which contains this law. One cannot, therefore, wholly disbelieve those who say that the exportation of figs also was anciently forbidden, and that the one who showed up, or pointed out such exporters, was called a 'sycophant,' or fig-shower." Trans. Perrin (1914). See Berti (2009b) 99–102.

65 "Istrus in his *Attic History* says that the dried figs produced by these trees were not exported from Attica, in order that only the inhabitants of the country could enjoy them. When many people were found to be evading the law, those who informed the jurors about them were then for the first time referred to as sycophants." Trans. by Olson (2006–2012).

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.

This scenario could include references to other primary sources that directly quote, allude to, or paraphrase Istros or the consultation of additional resources, such as comparisons of syntactic analyses and translations, or scholarly commentaries.[66]

Figure 3.5 shows the workflow of this text reuse scenario: Scholar Berti [B] is the *agent* of an Indication [B1] — the selection of a string of text. [B1] *informs* an Activity [S1] by the system — a Computation [S2] of the URI for a text passage [P1] by Plutarch [P]. [B] makes an Assertion [B2], the *substance of* which is Propositional Content [B3] as *expressed by* an Annotation [B4] *targeting* the Plutarch text [P1] as a reuse of a Lost Content Item [l1] attributed to Istros the Callimachaen [l]. [B] *indicates* another string of text [B5] that *informs* an Activity [S3] by the system — a Computation [S4] of the URI for a text passage [A2] by Athenaeus [A]. The text at [A2] is the *primary expression* of an Assertion [A1], the *substance of* which is Propositional Content [A3] as *expressed by* an Annotation [A4] *targeting* [A2] as a reuse of Lost Content Item [l1] attributed to Istros. B's Assertion [B2] is *warranted by* A's ancient Assertion [A1].

Combining SAM entities and properties with those from other models, such as W3C's PROV and *Open Annotation* (OA), provides a richer, more contextualized view of data encoding and use in humanities research:

– **SAM Indication**. Event in which some abstract structures (e.g., a series of characters) are indicated in the pursuit of some intellectual or creative goal.
– **SAM Assertion**. Event in which an agent advances a claim. The primary expressive may be a natural language sentence or any abstract arrangement of symbols. Assertions that are warranted by observations or computations are *systematic assertions*.
– **SAM Prospositional Content**. Language-independent bearer of truth values.
– **SAM Computation**. An event that may reflect contingencies such as scholarly interpretation. Does not necessarily imply *creation*. E.g., an analytic process may indicate a string of text which already existed.
– **PROV Activity**. Event through which entities come into existence and/or change to become new entities. Activities are dynamic aspects of the world, such as actions, processes, etc.

---

66   Berti (2009b) 99–102.

– **OA Annotation**. Expresses the relationship between two or more re-
sources, including metadata about the relationship concerning creation and
intent.

Combining complementary data provenance models enables to more precisely
track and document shared resources, ultimately improving data quality and en-
couraging further sharing. Using PROV Activities, it was possible to share precise
details about system actions and processes leading to the extraction of text and
creation of URI endpoints. SAM was used to identify the contingent aspect of
the underlying resources as things which are subject to interpretation and which
were in existence prior to their use as data in our analysis. OA was used to share
concrete serializations of the analyses in the form of annotations.[67] In this use
case the model enabled 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 (3) requirement, which
results from the second, is that it was necessary 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.

---

67 Almas/Berti et al. (2013).