# Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors

Bridget Almas
Perseus Project
Tufts University - Medford MA
bridget.almas@tufts.edu

Monica Berti
Universität Leipzig - Institut für Informatik
Digital Humanities
monberti@gmail.com

## ABSTRACT
The goal of this document is to present a fragmentary texts demo built under Perseids, a collaborative platform being developed by the Perseus Project that leverages and extends pre-existing open-source tools and services to support editing and annotating TEI XML documents in Classics: http://sites.tufts.edu/perseids/ [1]. The aim of this use case is to build a shared environment for multi-level annotations of text re-uses of ancient lost works: http://perseids.org/sites/berti_demo/index.html.

## Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *architectures, navigation.*

## General Terms
Documentation, Design, Standardization, Experimentation.

## Keywords
Fragmentary Texts, Text Re-use, Perseus Project, TEI, XML, OAC, JSON-LD, CTS/CITE Architecture, RDF, Annotations, Linked Data.

## 1. INTRODUCTION
Generations of scholars have collected a huge amount of information about lost works that is preserved in surviving sources. As a result, they have produced print editions of Greek and Latin fragmentary authors that are fundamental for reconstructing an otherwise lost past [4]. These pieces of information derive from a great variety of text re-uses that range from verbatim quotations to paraphrases, vague allusions and translations. In print culture these pieces of information are called "fragmenta" and are reproduced as *decontextualized extracts* from many different sources. Digital libraries offer the possibility to represent these re-uses inside their texts of transmission and therefore as *contextualized annotations* about lost works [5]. Such annotations include not only the portion of text that can be considered a re-use, but also much more information as names and geographic provenance of re-used authors with variants, titles and/or descriptions of re-used works, *verba dicendi*, expressions of literary criticism and many other linguistic and morphosyntactic features. Building a digital library of text re-uses of fragmentary authors means first of all to select the string of words that belong to the portion of text which is classifiable as re-use and secondly to encode all those elements that signal the presence of the text re-use (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/or syntax, parallel texts that deal with the same topic of the text re-use, and finally different editions and translations of both the source and the derived texts.


*DH-case '13*, September 10 2013, Florence, Italy

http://dx.doi.org/10.1145/2517978.2517986

## 2. PERSEIDS FRAGMENTARY TEXTS DEMO
The Perseids demo addresses many different requirements for producing for the first time a dynamic representation of quotations and text re-uses of fragmentary authors, using various methods of inline and stand-off markup to produce stable ways for identifying and annotating text re-uses, including canonical citations, morpho-syntactic analysis, translation and text re-use alignments. In this document we discuss in particular how we are combining TEI, the Open Annotation Core (OAC) data model, and the CITE Architecture to represent quotations and text re-uses via 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 [2]. The demo interface is based on the print edition of the fragments of Istros the Callimachean [3] and here we will focus on one example, which is a passage of the *Deipnosophists* of Athenaeus (3.6) that includes a text re-use from Istros (see Figure 1).

### 2.1 Canonical Citations of Text Re-Uses
The first function for a proper representation of text re-uses of lost works is to visualize them inside their embedding context. This means to select the string of words that belong to the portion of text which is classifiable as re-use. The Canonical Text Services (CTS) specification defines a URN-based identifier structure for identifying texts and related data objects, and network service application programming interfaces (APIs) for retrieving fragments of texts by canonical reference expressed as CTS URNs (http://www.homermultitext.org/hmt-doc/cite/index.html). 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, we 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 re-uses of lost works the

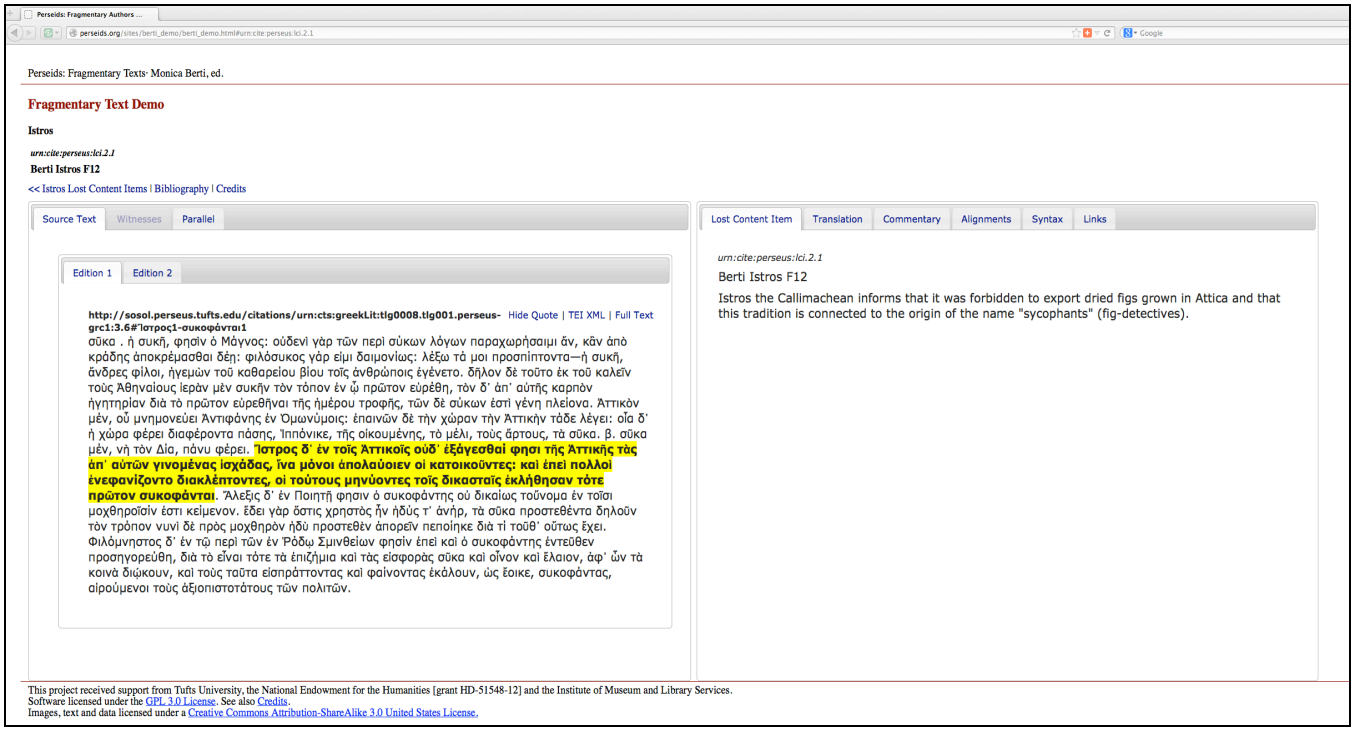


**Figure 1. Perseids Fragmentary Texts Demo**

situation is different, because the original text of the re-used author is lost and we have just the text of the re-using author, which is the only citable evidence. Accordingly, we have created a Perseus Collection of Lost Content Items (urn:cite:perseus:lci). These LCIs are assigned CITE URNs as unique identifiers, and assigned descriptive properties, for example naming a specific text re-use 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 (the CITE URN identifier for the Perseus Collection Object representing the text re-use of Istros with a reference to the edition of [3], where this portion of Athenaeus' text is reproduced and classified as Istros F12) *quotes* urn:cts:greekLit:tlg0008.tlg001:3.74e#Ἴστρος[1]-συκοφάνται[1] (the CTS URN identifier for Athen., *Deipn.* 3.6 with the addition of substring reference for greater precision[1]). 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).

## 2.2 Source Text, Witnesses, and Parallel Texts

Figure 1 shows the interface of the Perseids demo display with different functions for annotating text re-uses of fragmentary authors. On the left side the source text preserving the re-use (Athen., *Deipn*. 3.6) is visible through different editions and with links to the TEI XML file and the full text stored in Perseus. By "showing quote" the system highlights annotations of the portion of text classifiable as re-use according to different editors. The substring of the CTS URN specifies the range of words to be highlighted in the source text. It is also possible to visualize and annotate other sources that re-use the same text with different words or syntax (witnesses) and/or that deal with the same topic of the re-used text (parallel texts). The right side of the screen shows information about the text re-use (lost content item) that we are annotating (Istros re-used by Athenaeus with a reference to the edition of [3]) accompanied by its CITE URN, a title and a description of the content. Note that, as will be described further below, all source text, translations, commentaries and lost content item descriptions are retrieved at display time via asynchronous requests to remote services.

## 2.3 Annotating Text Re-Uses

On the right side of the interface, different editors can work on other information in order to build and implement a shared environment of multi-level annotations of text re-uses: (1) *Translations alignments* in different languages of the text re-use through the Alpheios Translation Alignment Editor (http://alpheios.net/). The translations in the demo were made using the Perseids Platform [1]. (2) *Commentaries* on the same text re-use for which we have created a Perseus Collection of Commentaries on Lost Content Items (urn:cite:perseus:lcicomm). (3) *Alignments* of witnesses and parallel texts (see above). (4) *Syntactic annotations of text re-uses*. Text re-use 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 re-uses with the Alpheios Treebank Editor in order to detect different examples of syntactic re-uses (e.g., different words with the same syntax and/or same words with different syntax). (5) *Links* to various resources such as scanned editions of sources and commentaries via Google Books and the Internet Archive.

[1] In between the publication of the demo and the writing of this paper, the CTS syntax for symbol separating the subreference from the passage changed from "#" to "@". We will be updating our demo code accordingly.

# 3. DATA-DRIVEN APPROACH

Annotations, and the texts and entities that they annotate, are the primary data type behind our demo. The demo combines the TEI XML 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. This application of standards and data enables us to present a new dynamic data-driven display leveraging linked open data and also to publish our own annotation data in a standard format to facilitate its reuse.

## 3.1 Text and Annotation Identifiers

We use CTS URNs to create semantically meaningful unique identifiers for texts, and passages within a text. We can reference either an abstract notional work or a precise expression of that work. The CITE protocol defines the following properties for a citable text node and the CTS URN syntax to identify text nodes that adhere to them: (1) belongs to a specific version of a work in a FRBR-like hierarchy; (2) belongs to a citation hierarchy of one or more levels; (3) is ordered; and (4) may have mixed content (text and nodes). A CTS URN is made up of the following distinct parts:

***urn:cts:NAMESPACE:TEXTGROUP.WORK.VERSION.EXEMPLAR:PASSAGE@SUBREF.***

In the example for Athenaeus' *Deipnosophists* provided above, the identifier *urn:cts:greekLit:tlg0008.tlg001:3.6* references Athen., *Deipn*. 3.6. By adding a version component to the identifier, *urn:cts:greekLit:tlg0008.tlg001.**perseus-grc1**:3.6*, we reference the same passage 3.6 but this time in the specific "perseus-grc1" edition of this work in the Perseus Digital Library.

The CITE architecture defines an alternate identifier syntax, in the form of a CITE urn, for data objects which don't meet the above-mentioned four characteristics of citable nodes. CITE urns can be used for images, fragments of lost texts, and a variety of different annotation types, and the syntax includes an image extension which supports identifying coordinates on an image.

In our demo we use the CTS and CITE data models to mint identifiers for the texts themselves, the specific passages of those texts which are the targets of the annotations, translations of those texts, the lost content items, and the annotations themselves. As URNs, these CTS and CITE identifiers are not web-resolvable on their own, but by combining them with a URI prefix, such as "http://data.perseus.org/citations" and deploying CTS and CITE services to serve the identified resources at those addresses, we have resolvable, stable and semantically meaningful URI identifiers for our texts, data objects and annotations (for details see http://sites.tufts.edu/perseusupdates/beta-features/perseus-stable-uris/). The CTS API for passage retrieval depends upon the availability of well-formed XML from which citable passages of texts can be retrieved by XPath. The TEI standard provides the markup syntax and vocabulary needed to produce XML which meets these requirements, and is a well-accepted standard for digitization of texts.

In our demo, the source text is served by the Perseus CTS API (http://sites.tufts.edu/perseusupdates/beta-features/perseus-cts-api/), translated text is served by an instance of the Alpheios CTS API (http://alpheios.net/content/alpheios-cts-api) and the Commentary annotations and Lost Content Item objects are served by an instance of the Google Fusion table implementation of the CITE Collections Service (https://bitbucket.org/neelsmith/citefusioncoll).


**Figure 2. OAC Annotations**

## 3.2 Using OAC for Data Publication and Display

The Open Annotation Core 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" (http://www.openannotation.org/spec/core/). This model enables us to express our annotations according to a defined and documented standard, increasing the feasibility of their reuse. Using the OAC data model we express annotations as simple URI based triples, 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 re-use annotations, where the text being re-used (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 re-use. 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 link 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 re-use 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. We 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, our 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". We link additional supporting resources, including other witnesses, translation alignments and morphosyntactic annotations in a similar manner.

Using the JSON-LD syntax recommended by OAC (http://www.w3.org/TR/json-ld-syntax) allows us to build a

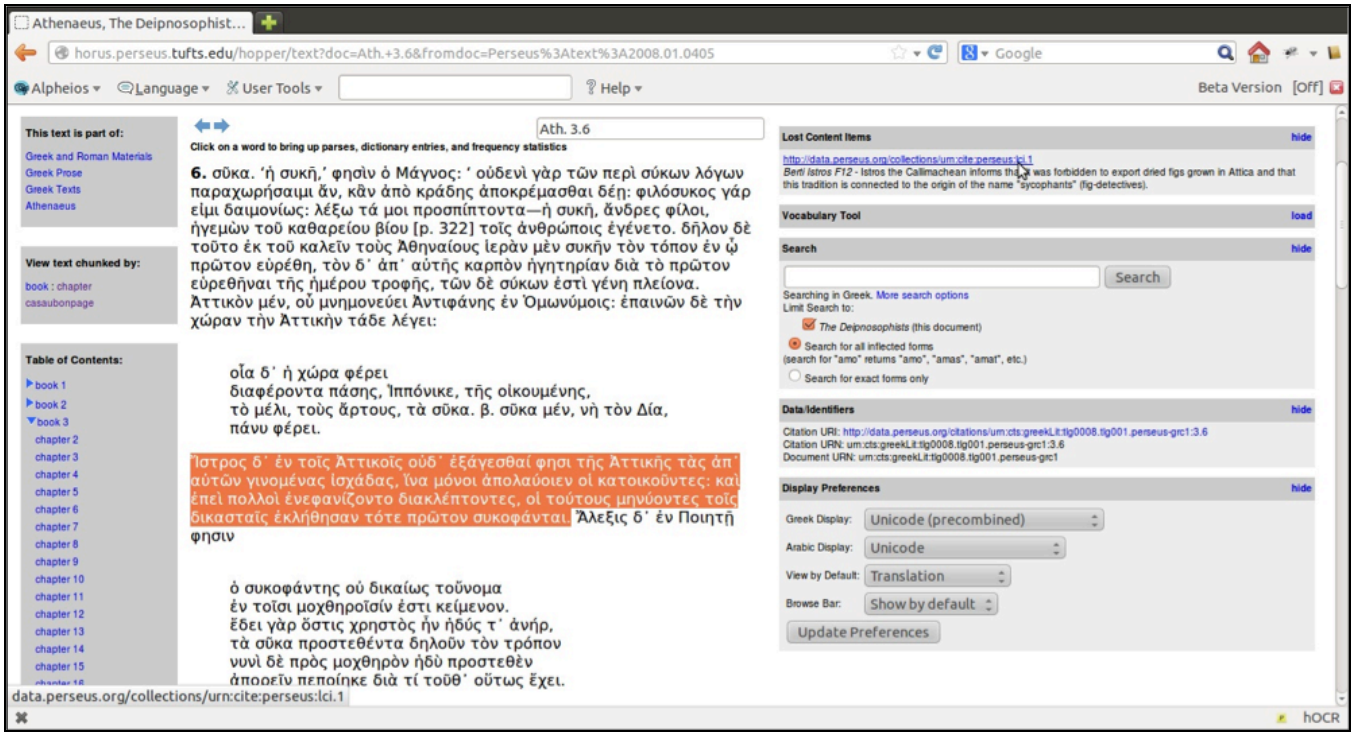


**Figure 3. Fragmentary Texts in Perseus**

dynamic display interface in Javascript that navigates the JSON-LD data object (see Figure 2) 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. In the future we may decide to process and present some or all of these resources inline in the display as well.

The demo interface code (https://github.com/PerseusDL/lci-demo) extends the CTS Kit from the Homer Multitext project (http://homermultitext.blogspot.com/2012/07/html-cts-kit-abstract-announcing-for.html) with customized stylesheets and display code, and to add processing of this JSON-LD structure containing the annotations.

## 4. CONCLUSIONS

The goal is to publish the annotations and include all this information in the collection of Greek and Roman materials in the Perseus Digital Library (see Figure 3). It's important to note that, while the work presented here is a demonstration of one specific use case and its implementation, it is part of a larger effort of the Perseids project to define and support a new model of scholarly publication in a born-digital environment. This model requires a platform which supports a wide variety of interoperable tools to collect, analyze, preserve and display textual data and annotations in various contexts, as those being developed by GERTRUDE (http://prezi.com/yfrrshdaiacd/the-tool-gertrude/), Hypothes.is (http://hypothes.is/) and the Shared Canvas project (http://www.shared-canvas.org/). Leveraging standard data models to facilitate integration of tools and data from various sources is a core premise behind the development of the Perseids platform.

## 5. ACKNOWLEDGMENTS

This work was supported by grants from Tufts University, the National Endowment for the Humanities [grant HD-51548-12] and the Institute of Museum and Library Services. We also thank the Homer Multitext Project for the development of the CITE Architecture and supporting services. New support from the Mellon Foundation will enable us to bring the demonstration code presented here from prototype to production quality (http://sites.tufts.edu/perseusupdates/2013/07/16/mellon-funds-perseids-project/).

## 6. REFERENCES

[1] Almas, B. and Beaulieu, M. C. 2013. Developing a New Integrated Editing Platform for Source Documents in Classics. *Literary & Linguistic Computing* DOI=10.1093/llc/fqt046.

[2] Almas, B. and Berti, M. 2013. The Linked Fragment: TEI and the Encoding of Text Re-uses of Lost Authors. *TEI Conference 2013*.

[3] Berti, M. 2009. *Istro il Callimacheo. Testimonianze e frammenti su Atene e sull'Attica*. Tored Edizioni, Tivoli.

[4] Berti, M. 2013. Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres. *Ancient Society* 43, 269-288.

[5] Berti, M., Romanello, M., Babeu, A. and Crane, G. 2009. Collecting Fragmentary Authors in a Digital Library. *JCDL '09* (Austin, TX, June 15-19, 2009). ACM, New York, NY, 259-262.DOI= http://doi.acm.org/10.1145/1555400.1555442.