The Linked TEI: Text Encoding in the Web

TEI Conference and Members Meeting 2013

Book of Abstracts

edited by
Fabio Ciotti
Arianna Ciula

ASSOCIAZIONE PER
L'INFORMATICA UMANISTICA
E LA CULTURA DIGITALE

DIGILAB
CENTRO INTERDIPARTIMENTALE
DI RICERCA E SERVIZI

SAPIENZA
UNIVERSITÀ DI ROMA

The Linked TEI: Text Encoding in the Web

Book of Abstracts - electronic edition

Abstracts of the TEI Conference and Members Meeting 2013:
October 2-5, Rome

Edited by Fabio Ciotti and Arianna Ciula

DIGILAB Sapienza University & TEI Consortium
Rome 2013

The Linked Fragment: TEI and the encoding of text re-uses of lost authors

Berti, Monica; Almas, Bridget

The goal of this paper is to present characteristics and requirements for encoding quotations and text re-uses of lost works (i.e., those pieces of information about lost authors that humanists classify as 'fragments'). In particular the discussion will focus on the work currently done using components of Perseids (http://sites.tufts.edu/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 source documents in Classics.

Working with text re-uses of fragmentary authors means annotating information pertaining to lost works that is embedded in surviving texts. These fragments of information derive from a great variety of text re-uses that range from verbatim quotations to vague allusions and translations. One of the main challenges when looking for traces of lost works is the reconstruction of the complex relationship between the text re-use and its embedding context. Pursuing this goal means dealing with three main tasks: 1) weighing the level of interference played by the author who has reused and transformed the original context of the information; 2) measuring the distance between the source text and the derived text; 3) trying to perceive the degree of text re-use and its effects on the final text.

The first step for rethinking the significance of quotations and text re-uses of lost works is to represent them inside their preserving context. This 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 (i.e., named entities such as the onomastics of re-used authors, titles of re-used works and descriptions of their content, *verba dicendi*, syntax, etc.). The second step is to align and encode all 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 re-use (parallel texts), and finally different editions and translations of both the source and the derived texts.

This paper addresses the following requirements for producing a dynamic representation of quotations and text re-uses of fragmentary authors, which involve different technologies including both inline and stand-off markup:

- *Identifiers*: i.e. stable ways for identifying: fragmentary authors; different kinds of quotations and text re-uses; passages and works that preserve quotations and text re-uses; editions and translations of source texts; entities mentioned within the text re-uses; annotations on the text re-uses.
- *Links*: between the fragment identifier and the instances of text re-use, the fragment identifier and the attributed author, the fragment identifier and an edition which collects it; between the quoted passage and the entities referenced in it; between the quoted passage and translations.
- *Annotations*: the type of re-use; canonical citations of text re-uses; dates of the initial creation of the re-use, of the work which quotes it, author birth and death; editorial commentary on each text re-use; bibliography; morphosyntactic analysis of the quoted passage; text re-use analysis (across different re-uses of the same text); syntactic re-use analysis; translation alignments (between re-used passages and their translations); text reuse alignments (between different re-uses of the passage in the same language).
- *Collections* (the goal is to organize text re-uses into the following types of collections): all text re-uses represented in a given edition which includes re-uses from one or many authors; all text re-uses attributed to a specific author; all text re-uses quoted by a specific author; all text re-uses referencing a specific topic; all text re-uses attributed to a specific time period, etc.

In the paper we discuss in particular how we are combining TEI (http://www.tei-c.org), the Open Annotation Collaboration (OAC) core data model (http://www.openannotation.org/spec/core/), and the CITE Architecture (http://www.homermultitext.org/hmt-doc/cite/index.html) to represent quotations and text re-uses via RDF triples. The subject and object resources of these triples can be resolved by Canonical Text and CITE Collection Services to supply the TEI XML and other source data in real time in order to produce new dynamic, data-driven representations of the aggregated information.

The CITE Architecture defines CTS URNs for creating semantically meaningful unique identifiers for texts, and passages within a text. It also defines an alternate identifier syntax, in the form of a CITE URN, for data objects which don't meet the characteristics of citable text nodes, such as images, text re-uses of lost works, and annotations. As URNs, these identifiers are not web-resolvable on their own, but by combining them with a URI 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. In the paper we supply specific examples of URNs, and their corresponding URIs, for texts, citations, images and annotations.

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. Particularly applicable are the TEI elements for representing the hierarchy of citable nodes in a text. The Open Annotation Core data model provides with us a controlled vocabulary to identify the motivation for the annotations and enables us to express our annotation triples according to a defined and documented standard.

In the paper we present practical examples of annotations of text re-uses of lost works that have been realized using components of the Perseids platform. In Perseids we are combining and extending a variety of open source tools and frameworks that have been developed by members of the Digital Classics communitity in order to provide a collaborative environment for editing, annotating and publishing digital editions and annotations. The two most prominent components of this platform are the Son of SUDA Online tool developed by the Papyri.info (http://papyri.info) project and the CITE architecture, as previously mentioned.

The outcome of this work is presented in a demonstration interface of Perseids, The Fragmentary Texts Demo (http://services.perseus.tufts.edu/berti_demo/). We also present the data driving the demo, which contains sets of OAC annotations (http://services.perseus.tufts.edu/berti_demo/berti_annotate.js) serialized according to the JSON-LD specification.

The final goal is to publish the annotations and include all the information pertaining to fragmentary texts in the collection of Greek and Roman materials in the Perseus Digital Library. The purpose is to collect different kinds of annotations of text re-uses of fragmentary authors with a twofold perspective: 1) going beyond the limits of print culture collections where text re-uses are reproduced as decontextualized extracts from many different sources, and representing them inside their texts of transmission and therefore as contextualized annotations about lost works; 2) allowing the user to retrieve multiple search results using different criteria: collections of fragmentary authors and works, morphosyntactic data concerning text re-uses, information about the lexicon of re-used words, cross-genre re-uses, text re-use topics, etc.

Bibliography

- Almas, Bridget and Beaulieu, Marie-Claire (2013): *Developing a New Integrated Editing Platform for Source Documents in Classics*. In: Literary and Linguistic Computing (Digital Humanities 2012 Proceedings) (forthcoming).
- Berti, Monica (2013): *Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres*. In: Ancient Society 43.
- Berti, Monica, Romanello, Matteo, Babeu, Alison and Crane, Gregory R. (2009): *Collecting Fragmentary Authors in a Digital Library*. In: Proceedings of the 2009 Joint International Conference on Digital Libraries (JCDL '09). Austin, TX. New York, NY: ACM Digital Library, 259-262. http://dl.acm.org/citation.cfm?id=1555442
- Büchler, Marco, Geßner, Annette, Berti, Monica, and Eckart, Thomas (2012): *Measuring the Influence of a Work by Text Reuse*. In: Dunn, Stuart and Mahony, Simon (Ed.): Digital Classicist

Supplement. Bulletin of the Institute of Classical Studies. Wiley-Blackwell.

- Crane, Gregory R. (2011): *From Subjects to Citizens in a Global Republic of Letters*. In: Grandin, Karl (Ed.): Going Digital. Evolutionary and Revolutionary Aspects of Digitization. Nobel Symposium 147. The Nobel Foundation, 251-254.
- Romanello, Matteo, Berti, Monica, Boschetti, Federico, Babeu, Alison and Crane, Gregory R. (2009):*Rethinking Critical Editions of Fragmentary Texts by Ontologies*. In: ELPUB 2009: 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies. Milan, 155-174. http://hdl.handle.net/10427/70403
- Smith, D. Neel and Blackwell, Chris (2012): *Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture*. In: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends. The Center of Hellenic Studies of Harvard University. http://folio.furman.edu/projects/cite/four_urls.html

“Reports of My Death Are Greatly Exaggerated”: Findings from the TEI in Libraries Survey

Dalmau, Michelle; Hawkins, Kevin S.

Historically libraries, especially academic libraries, have contributed to the development of the TEI Guidelines, largely in response to mandates to provide access to and preserve electronic texts (Engle 1998; Friedland 1997; Giesecke, McNeil, and Minks 2000; Nellhaus 2011). At the turn of the 21st century, momentum for text encoding grew in libraries as a result of the maturation of pioneering digital library programs and XML-