Dialogue and linking between TEI and other semantic models

TEI started a deep dialogue with other semantic models - i.e. CIDOC-CRM and FRBR/FRBR(OO) -, with the purpose of data interchange and in order to increase the digital editors possibilities to formally declare hermeneutical positions. If the TEI schema proposes most of the elements/attributes (and classes) useful to describe interpretation instances, other schemas as well as other value vocabularies and metadata element sets, could improve some potentiality of the model. On one hand other schemas could contribute to refine the function of some TEI elements; while on the other existing ontologies could enhance the interpretation effectiveness. The aim of this panel is to introduce 3 different approaches to documents representation, where TEI could draw some hints from other models.

First the role of EAC (Encoded Archival Context) is studied in order to enrich the description of people, starting from the archival approach to context intended as the reading key for defining individuals' roles and functions. A second aspect is the dialogue between TEI and the existing ontologies, with a special attention to the geographic data. Finally, the application of "semantic lenses" as an exploratory tool for annotated documents could open up the relationships between TEI and specific ontologies devoted to semantic publishing.

All these approaches adopt a linked data perspective, enriching TEI element with @ref, URI and the RDF model for assertions. Exposing TEI annotations as data sets could improve the interchange of both the schema, and the documents based on it, with other exiting data sets, enhancing the information retrieval possibility. Digital editions based on TEI could start a dialogue with the WWW resources in a global vision of heritage as a connection of cultural data where digital editions play a crucial role in the preservation of cultural memory as an interlink between literary texts, archival documents, museum objects and books.

1. TEI <person> towards EAC: the identity between functions and context

Francesca Tomasi (University of Bologna)

Among the main revisions of the P5 version TEI Schema, the section Biographical and Prosopographical Data [1] constitutes a challenging innovation. TEI decided to invest on the 'person' concept, defining a taxonomy of elements useful to describe individuals. In 2006 a special workgroup called 'Personography' was chartered "to investigate how other existing XML schemes and TEI customization handle data about people" [2] and a "Report on XML mark-up of biographical and prosopographical data" was published [3].

A fundamental approach to describing people is the unique identification of individuals and the enrichment of descriptions through features classification. However, we must never forget that people are strongly connected to the context in which they appear, performing activities; as a result, roles and functions, intended as individuals' features, naturally change depending on the context, i.e. on the source that attests the individual. It's therefore possible to state that: 1) some features are static over the time and they are theoretically constant with respect to context (i.e. birth, death, nationality, persName); 2) other features vary depending on date and place (i.e age, affiliation, education, event, state); 3) roles and functions (i.e. author, actor, editor, speaker) are elements that identify people based on the context.

We can then assert that a person is a *complex entity*, because individuals are connected to different phenomena typologies: some of them are unchangeable, some others depend on a time period, on a place or on the context, and are able to transform a string in a concept. With concept we mean an assertion originated from the relation between the elements necessary to provide meaning.

The element `<person>` in TEI could be associated to different roles or functions. Let's consider a digital edition of a literary text. A person is, respectively: the one who created the digital edition - at the different levels -, the author of the analogic source, the editor of the printed version, the whole of individuals cited in the text. The concept of person extends its boundaries: although individuals are strictly related to the source, that provides the appropriate semantic background, they are also entities with a function able to connect the same person to different documents, or to other resources in general, and a person with other people that share the same role. Multiple relationships then arise:

between individuals, between a person and a document in which the person is mentioned and between a person and other resources.
This reflection links TEI to one particular XML schema that is EAC (Encoded Archival Context) [4] developed in order to formalize the ISAAR (CPF) standard (International Standard Archival Authority Record for Corporate Bodies, Persons and Families)[5] and now represented also as ontology [6]. EAC contributes to the reasoning on individuals, pointing out the importance of both the context and the relationships. The approach here described aims to extend the domain of digital editions to the archival studies one. The archival science declares the principle of the separation between the description of records (documents) and the description of people (corporate bodies, persons and families) [7], focusing on the context as a fundamental keyword. The same approach could be followed in TEI mostly, if the final aim is to expose data sets to be used by the Web community.
It becomes important to think of EAC as a schema able to suggest how to extend the concept of `<relation>` in TEI. EAC (CPF) is based on the principle of entity intended as corporate body, person, or family that manage relations – between entities and between one entity and a resource linked at some level - each of which is described, dated and categorized. Besides the elements connected to the "relation" principle (`<cpfRelation>` and `<resourceRelation>`), EAC describes the `<function>` element that "provides information about a function, activity, role, or purpose performed or manifested by the entity being described" on a specific date. The element `<functionRelation>` describes a "function related to the described entity. [...] Includes an attribute `@functionRelationType`" that could support a taxonomy of values [4].
A new model of *authority record,* intended as complex structure, able to document the context in which the identity is attested, could be therefore proposed: the authority is generated not only by the controlled form of the name, and the related parallel forms, but it is also the result of relationships that are born from the context to determine a concept.
Following the RDF model it's possible to say that an identified entity (URI) manages relationships (predicate) with different objects: another entity (URI), a place (URI), a date (URI), an event (URI), a contextual resource (URI), i.e. the document, an external resource (URI), that is another object related to the entity, because it shares the same role or function (i.e. the same actor, identified with a specific and controlled `<persName>`, involved in two different drama, covering a different function in two different time).
We could try to exemplify this procedure on an individual's responsibility: a contributor of a digital edition who, on a specific date, performed a specific activity. TEI metadata propose two places for the responsibility description (`<fileDesc>` e `<revisionDesc>`):

```
<fileDesc><titleStmt><respStmt> <resp>,<name>
<revisionDesc><respStmt> <resp>,<persName>
```

Each person is associated to a responsibility able to identify the function that the entity covered in that specific document, linking people to resource. The same person could cover the same, but also a different, responsibility in other editions; in this way relationships could be extended to other documents. Other individuals could be then connected to the first person because of the sharing of the same responsibility.
This process could be declared and exposed as data set with RDF and URI for the syntax and TEI/EAC for classes and predicates in order to build a collection of authorities of people that covered a role or a function in a certain time and in a specific context. Declaring connections as relationships, through the EAC model, a knowledge base of people, with a context-originated function, could be developed.
The digital editions open in this way the vision to the cultural heritage domain, defining connections between heterogeneous objects and "creating efficiencies in the re-use of metadata across repositories, and through open linked data resources" [8]. Linked Data describing persons acting in specific roles would be considerably improved by using specifications for these persons' function using the context as interpretative key: "the description of personal roles and of the statuses of documents needs to vary in time and according to changing contexts [...] such roles and statuses need to be handled formally by ontological models." [9]

Bibliography

[1] TEI Consortium (eds.). "13.3 Biographical and Prosopographical Data". In *Guidelines for Electronic Text Encoding and Interchange*. Last updated on 21 December 2011. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html#NDPERS
[2] TEI: Personography Task Force. http://www.tei-c.org/Activities/Workgroups/PERS/index.xml
[3] Wedervang-Jensen, Eva, and Matthew Driscoll, *Report on XML mark-up of biographical and prosopographical data*. 16 Feb 2006. http://www.tei-c.org/Activities/Workgroups/PERS/persw02.xml
[4] EAC-CPF, Encoded Archival Context for Corporate Bodies, Persons, and Families. http://eac.staatsbibliothek-berlin.de/
[5] CBPS - Sub-Committee on Descriptive Standards. "ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families". 2nd Edition, 2003. http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html
[6] Mazzini, Silvia, and Francesca Ricci. 2011. "EAC-CPF Ontology and Linked Archival Data". In *Semantic Digital Archives (SDA) Proceedings of the 1st International Workshop on Semantic Digital Archives*. http://ceur-ws.org/Vol-801/.
[7] Pitti, Daniel. 2004. "Creator Description: Encoded Archival Context". *Authority control in organizing and accessing information: definition and international experience*. Ed. Arlene G. Taylor, 1941-, Barbara B. Tillett, Murtha Baca and Mauro Guerrini, 201-226. Binghamton N.Y.: Haworth Information Press.
[8] Larson, Ray R., and Krishna Janakiraman. 2011. "Connecting Archival Collections: The Social Networks and Archival Context Project". In *Research and Advanced Technology for Digital Libraries. Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL 2011)*. Ed. Stefan Gradmann, Francesca Borri, Carlo Meghini and Heiko Schuldt, 3-14. Heidelberg, Germany: Springer. DOI: 10.1007/978-3-642-24469-8_3.
[9] Peroni, Silvio, David Shotton, and Fabio Vitali. 2012. "Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents". In *Proceedings of the 8th International Conference on Semantic Systems*, 9-16. ACM, New York. DOI: 10.1145/2362499.2362502

2. Geolat: a digital geography for Latin literature

Maurizio Lana (University of Piemonte Orientale); Fabio Ciotti (University of Roma Tor Vergata) and Diego Magro (University of Torino)

This paper presents the "Geolat" project, which aims to make accessible the Latin literature through a query interface of geographic / cartographic type. The project, under the name DAGOCLaT (Digital Atlas with Geographical Ontology for Classical Latin Texts) in 2012 was presented in response to the call of "Compagnia di San Paolo Foundation" and at the end of a blind peer evaluation managed by European Science Foundation was funded for exploratory and initial activities. In January 2013, under the name ALTUSS (Advanced Latin Texts Uses for School and Society) the project, revised and enriched among other things by an advisory board composed by Gregory Crane (Perseus, Pelagios), Tom Elliott (Pleiades) and Leif Isaksen (Google Ancient Places), was presented in response to the European call ERC Synergy.

The first objective of the project is to set up a digital library that contains the works of Latin literature from its origins to the end of the Roman Empire (conventional date, the 476 d. C.). This stage involves the integration of various already existing repository of Latin texts of high philological quality, which will be integrated starting from their already existing TEI/XML encoding. Building a (someone could say "the") global digital library of ancient Latin literature is a very important field where APA is working [1], where Gregory Crane recently called [2] to start working, and where the "Geolat" project too will build its global library, because the library is a pre- condition for all the subsequent activities. All the library texts will be encoded with a very light TEI subset of tags.

In a second phase the works so collected are analyzed at morphological level by means of a parser (that of Lasla of Liège [3]) so as to associate with each word its analysis / morphological description, which includes the identification of proper names. After that, by means of manual intervention, geographic references will be progressively encoded in a formal manner by adopting the TEI elements <placeName> and <geogName> (described in the TEI Guidelines in chapter 13 "Names, Dates, People, and Places"). Each occurrence of place names and geographical references will be identified by a URI

(using the @ref attribute) that will point to a formal description of the place in a formal ontology of the ancient Latin world geography (the traditional printed reference was and still is the Barrington Atlas [4]).
This ontology will be built ad hoc, reusing the data offered by the Pleiades gazetteer [5], and establishing relationships with other relevant geographic ontologies, where possible, such as Geonames. In general the ontology will be structured in a two tier fashion (following the tradition in DL ontology modelling): a T-box modelling geospatial classes of locations their properties and their relationships and an A-box with geospatial information about individual places and location. At this level the sites of antiquity will be associated with a variety of information:

- URI (and eventual links to URIS in other data sets)
- GPS coordinates
- different names, time frames of validity and etymology
- belonging to an itinerary (pilgrimage, military expedition, etc.)
- typology
- historical, geographical, cultural annotations
- links to other relevant Linked Data sets

A third level of modeling will be tied to the logical relationship between textual references (and their annotations by an encoder) and their referent in the ontology. In fact, you can easily detect that the textual context in which each geographical word (or phrase) is placed determines different modes of reference. From this point of view it seems necessary to introduce into the system an ontology of (geographic) annotations that can account for this variety of reference. In our work we will also discuss the various operational opportunities to formalize this information at the level of inline markup or through links to RDF statements in stand-off markup.
All the resources produced in our project, as he primary sources as the geographic thesaurus and the list of textual annotations that link geographic locations and places text (identified by URI) will be made available on the Web according to the principles of Linked Data, and will help to enrich the "Web of Data" with new content.

Bibliography

[1] APA Digital Latin Library Project
http://www.apaclassics.org/index.php/research/digital_latin_library_project [2] Gregory Crane call
http://sites.tufts.edu/perseusupdates/2013/02/14/possible-jobs-in-digital- humanities-at-leipzig/
[2] LASLA http://www.cipl.ulg.ac.be/Lasla/
[3] Talbert R. (ed.), The Barrington Atlas of the Greek and Roman World, Princeton University Press 2000
[4] Pleiades Project, http://pleiades.stoa.org/
[5] GAP – Google Ancient Places, http://googleancientplaces.wordpress.com/
[5] GAPvis http://googleancientplaces.wordpress.com/gapvis/
[6] Tom Heath and Christian Bizer, Linked Data: Evolving the Web into a Global Data Space. S7nthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool 2011
[8] Tom Elliiot, S. Gillies, "Digital Geography and Classics, in Digital Humanities Quarterly 3.1 (Winter 2009), http://www.digitalhumanities.org/dhq/vol/3/1/000031.html
[9] Open Annotation Data Model, Open Annotation Community Group 2013,
http://www.openannotation.org/spec/core/

3. Bringing semantic publishing in TEI: ideas and pointers

Silvio Peroni and Fabio Vitali (University of Bologna)

TEI has a full set of elements that can be used to describe facts about the publication details of a text, such as editionStmt, publicationStmt, and sourceDesc. A numerous list of sub-elements allows a zealous editor to provide a rich overview of publication aspects of the paper editions of the text, of this specific XML document, and of the steps through which an original source has made this XML possible.
Several collections of allowable values for these elements exist, as thesauri, authority lists or simple value lists, that simplify the task to describe frequent or common situations, and that homogenize

similar occurrence in different documents of the same collection. In a way, we could characterize value thesauri as *external aids to improve internal quality* of digital collections of texts.

In the last few years, a new discipline has arisen, *semantic publishing*, that tries to improve the scientific communication by using of web and semantic web technologies to enhance a published document so as to enrich its meaning, to facilitate its automatic discovery, to enable its linking to semantically related articles, to provide access to data within the article in actionable form, and to allow integration of data between papers [1,2]. Its main interest lies in the organization and description of scientific literatures, trying to tame the incredible complexity of the modern scientific publishing environment, both in terms of size and credibility of publishing venues, authors, research groups and sponsors. For instance, SPAR [3,4,5] is a suite of orthogonal and complementary ontology modules for creating comprehensive machine-readable RDF metadata for all aspects of semantic publishing and referencing, each of them precisely and coherently covering one aspect of the publishing domain using terms with which publishers are familiar. Together, they provide the ability to describe bibliographic entities such as books and journal articles, reference citations, the organization of bibliographic records and references into bibliographies, ordered reference lists and library catalogues, the component parts of documents, and publishing roles, publishing statuses and publishing workflows. SPAR ontologies have been already used in different projects such as *JISC Open Citations Project* [6] – a database of biomedical literature citations, harvested from the reference lists of all open access articles in PubMed Central that reference ~20% of all PubMed Central papers (approx. 3.4 million papers), including all the highly cited papers in every biomedical field – and *Semantic Web Applications in Neuromedicine (SWAN) Project* [7].

One of the main aims of semantic publishing therefore is to create a rich network of interconnected facts about publications from which interesting patterns can emerge to discover, for instance, clusters of similar publications, intrinsic values of publication venues, emerging trends in publication topics, etc. In a way, we could characterize annotations coming from actual documents as *internal aids to improve the external qualities* of digital collections of texts, especially regarding emerging characteristics of the collections themselves rather than belonging to individual documents.

We believe that the combination of these aspects could be mutually beneficial both in the increased quality of the individual documents, as well as in the increased quality and explorability of the emerging properties of document collections.

Being able to associate a full set of related facts to individual values in individual elements of the publication and edition details of the electronic version of a text provides the end user with a large and interesting network of considerations that go well beyond the individual text, and using standard tools from the Semantic Web may well allow reader to connect and exploit, for instance, the vast and growing collections of facts that embody the Linked Data initiative.

The actual syntax for this mesh is not particularly relevant. What is relevant is that through some syntactical mechanisms, it ends up being possible for an individual TEI document to feed Linked Data new and interesting facts about the corresponding publications and the involved actors, and conversely for Linked Data collections to enrich the amount of information about the publication and the involved actors that is made available to the interested reader, directly or after explicit queries, automatically or through the filtering and selecting action of an electronic editor.

The actual link between TEI documents and Linked Data resources is already feasible by adopting particular techniques and tools. Mainly, there are two ways to enable annotations linking existing TEI documents to Linked Data resources: either one embeds the annotation in the document itself (*embedding* techniques) or the annotations are stored in a separate document with references to the parts of the document each annotation refers to (*standoff* techniques). Neither the use embedding nor the use of standoff annotations is wrong or correct on its own; each technique has its own pros and cons that must be evaluated case by case before using them.

Even though many techniques have been devised in the past, usually the more technical solutions address only the problem of how to store the annotations, without dealing with the meaning of the annotations themself. In the case of embedded annotations, these solutions offer a generic way to augment existing markup with annotations (e.g. RDFa [9]). In the case of standoff annotations, the existing technical solutions provide a way to address content (e.g. EARMARK [10-11] and NIF [12]). In addition to other approaches, EARMARK offers an extension [13] to actually express the meaning of the annotation and allows one to easily link bunch of text in TEI documents to external resources. It

also provides a Java API [14] to support users in creating (even overlapping) annotations upon the same text, keeping track of provenance information such as the author who made the annotation and the time in which the annotation has been created.
The technical solutions are only one half of what is needed to annotate documents. The other half is the use of an annotation model and vocabulary. There are many such vocabularies available, ranging from very generic annotation frameworks (e.g. the Open Annotation Data Model (OADM) [15] or the Annotation Ontology [16]), to more specific frameworks (e.g. the Linguistic Annotation Framework (LAF) [17], used to annotate the various linguistic features of a speech through its transcript, or Domeo [18], that describes annotations used to connect scholarly documents).

Bibliography

[1] Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing, Learned Publishing 22 (2): 85–94. DOI: 10. 1087/2009202
[2] Shotton, D., Portwin, K., Klyne, G., Miles, A. (2009). Adventures in semantic publishing: exemplar semantic enhancements of a research article, PLoS Computational Biology 5 (4): e1000361. DOI: 10.1371/journal.pcbi. 1000361
[3] Semantic Publishing and Referencing Ontologies: http://purl.org/spar.
[4] Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 17 (December 2012): 33-43. DOI: 10.1016/j.websem.2012.08.001
[5] Peroni, S., Shotton, D., Vitali, F. (2012). Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents. In Presutti, V., Pinto, H. S. (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9-16. DOI: 10.1145/2362499.2362502
[6] JISC Open Citations homepage: http://opencitations.net.
[7] Ciccarese, P., Wu, E., Kinoshita, J., Wong, G., Ocana, M., Ruttenberg, A., Clark, T. (2008). The SWAN biomedical discourse ontology, Journal of Biomedical Informatics 41 (5: 739–751. DOI: 10.1016/j.jbi.2008.04.010
[8] Huitfeldt, C., Sperberg-McQueen, C. M. (2001). Texmecs: An experimental markup meta-language for complex documents. Working paper of the project MLCD, University of Bergen.
[9] Adida, B., Birbeck, M., McCarron, S., Herman, I. (2012). RDFa Core 1.1. W3C Recommendation, 7 June 2012. World Wide Web Consortium. http: //www.w3.org/TR/2012/REC-rdfa-core-20120607/
[10] Di Iorio, A., Peroni, S., Vitali, F. (2011). Using Semantic Web technologies for analysis and validation of structural markup. In International Journal of Web Engineering and Technologies, 6 (4): 375-398. Olney, Buckinghamshire, UK: Inderscience Publisher. DOI: 10.1504/IJWET.2011.043439
[11] Di Iorio, A., Peroni, S., Vitali, F. (2011). A Semantic Web Approach To Everyday Overlapping Markup. In Journal of the American Society for Information Science and Technology, 62 (9): 1696-1716. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. DOI: 10.1002/asi.21591
[12] Hellmann, S., Lehmann, J., Auer, S. (2012). Linked-data aware uri schemes for referencing text fragments. In ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Aquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (Eds.), Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Lecture Notes in Computer Science 7603: 398-412. Berlin, Germany: Springer. DOI: 10.1007/978-3-642-33876-2_17
[13] Peroni, S., Gangemi, A., Vitali, F. (2011). Dealing with Markup Semantics. In Ghidini, C., Ngonga Ngomo, A., Lindstaedt, S. N., Pellegrini, T. (Eds.), Proceedings the 7th International Conference on Semantic Systems (I-SEMANTICS 2011): 111-118. New York, New York, USA: ACM. DOI: 10.1145/2063518.2063533
[14] Barabucci, G., Di Iorio, A., Peroni, S., Poggi, F., Vitali, F (2013). Annotations with EARMARK in practice: a fairy tale. Submitted for publication in the 1st Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities (DH-CASE 2013).
[15] Sanderson, R., Ciccarese, P., de Sompel, H. V. (2013). Open annotation data model. W3C Community draft, 08 February 2013. http: //www.openannotation.org/spec/core/20130208/

[16] Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T. (2011). An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics, 2 (2): 1–24. DOI: 10.1186/2041-1480-2-S2-S4
[17] ISO (2012). ISO 24612:2012 Language resource management — Linguistic annotation framework (LAF). ISO.
[18] Ciccarese, P., Ocana, M., Clark, T. (2012). Open semantic annotation of scientific publications using DOMEO. Journal of Biomedical Semantics, 3 (1): 1–14. DOI: 10.1186/2041-1480-3-S1-S1