Formal ontologies, Linked Data and TEI semantics

Fabio Ciotti1, Francesca Tomasi2
1Università Roma Tor Vergata, Italy; 2Università di Bologna, Italy;
fabio.ciotti@uniroma2.it

The problem of the XML semantics

The debate on the Markup Language semantic role has been quite lively during the last twenty years and the experience of TEI practice community has played an active role in this context. It is commonly acknowledged that the markup conveys semantic aspects, whether they are local "interpretations" produced by a single scholar, or rather the expression of a general text theory.

However, this markup and, in particular, the XML markup semantic role, clashes with the fact that, as Robin Cover (1998) already observed some time ago, XML is a poor language for semantic data modelling. XML is powerful in defining the markup syntax, and, through its data model, to model some structural features of information objects. Still, it owes its semantic value almost entirely to human interpretation. Any markup restriction or semantic role accordingly needs to be expressed in natural language as instructions for human users, as in ODD formalism (Burnard 2012), developed to combine in a single meta-XML document the custom definition of XML schema and all its relevant documentation.

Several proposals were drawn up to provide XML with formalized and computable semantics. The work of Renear, Dubin, Sperberg-McQueen and Huitfeldt (2000 and 2002, and others) constitutes the first, explicit contribution in this direction. The Authors, from the observation that semantic markup coincides with the set of inferences authorized by its constructs, propose a formal markup semantics based on Prolog clauses. More recent works on the topic were performed by Peroni Gangemi and Vitali (2011) and Sperberg-McQueen, Marcoux, and Huitfeldt (2009).

In these years the development and the spread of the Semantic Web and Linked Data paradigm has made available a number of syntactically rigorous and semantically well-founded languages and data models (RDF / RDFS and OWL 2), and application for semantic data processing. In our opinion, the range of possibilities offered by these tools, justify the opportunity to build a semantic and ontological extension to TEI.

A semantics for TEI based on OWL 2 DL

We propose an extension of TEI infrastructure in order to formalize the semantic levels of the several markup constructs it provides. It is appropriate to distinguish between different semantic levels expressed by the markup and its content:

1. TEI general ontology

2. Local ontology and intensional TEI markup semantics (defined by a particular community of practice)

3. Extensional semantics of the markup content

TEI General Ontology

The first level consists in the common notion that the TEI is or expresses a text ontology. or as stated by Sperberg-McQueen that the "markup reflects a theory of text" (1991). Actually as Guarino (2009) notes, like a theory a formal ontology conflates epistemological aspects and ontological commitements.

The TEI conceptual domain can be translated into a formal ontology in OWL DL, where it can express the hierarchical relationships, three classes and TEI elements as well as a limited set of contextual constraints and restrictions in addition to those imposed on elementary data types. As a first approximation, we can formulate this construction scheme as follows:

1. TEI model class and element class are converted into (macro) OWL classes

2. The elements are converted into sub-classes

3. The attributes are converted into OWL property (DataProperty if their value is a literal or a data, or ObjectProperty if they specify relationships between elements)

4. A couple of hasChilds and hasFather reverse properties to explicitly formalize the XML tree hierarchical relationships

In many respects, the construction of a formal high-level TEI ontology could be a partially automated process starting from the implicit semantics in the schema. However, it is predictable that many semantics restrictions, which cannot be expressed by common Schema Languages (and ODD), should be explicitly expressed.

Local ontologies and intensional semantics of the elements

At this level we place the formal definitions of the semantics intended for a certain XML / TEI constructs subset by a particular users' community. For example, think of a specialization in the use of abstract elements such as , <ab>, <seg> or their attributes that define an intensional, more specific and restricted semantics compared to that described at general ontology level.

These ontology specializations can be expressed as:

1. Restrictions on properties and classes that extend the general ontology in OWL DL

2. A set of inference rules expressed through Rule Language.

3. Semantic definitions through specialized formalisms such as EARMARK (see Peroni, S., Vitali, F. 2009).

How can we possibly express these local semantic extensions? The most obvious method is to introduce a dedicated construct in the ODD language that allows a user to declare the relevant formulas. However, this strategy does not cover the need to define semantically specific instances of a markup element.

Extensional semantics of the markup content

The last semantic level concerns the extensional semantics of the individual XML elements content within a document. We adopt the term 'extensional' because, in general, it is suitable for fixing the referent of a linguistic expression identified by the markup through its reference to resources (information entities) via URI, or the connection to the linked data set.

The case of the extensional bond with a single external resource is already managed by the current TEI scheme through the @ref attribute. The possibility to express multiple referential connections can be simply expressed by specifying a URI sequence, separated by a space. More complex markup structures may allow the expression of more complex, logical architectures through stand-off elements.

Conclusions and perspectives

In our opinion, the possibility of providing a TEI-formalized semantics through the use of Semantic Web standard technology constitutes a manifold opportunity to:

1. Strictly set out the general semantics of the markup language in order to facilitate the management and research in open and multi-standard contexts, such as large-scale general libraries and large institutional repositories

2. Facilitate interoperability with other standards relevant in the Digital Cultural Heritage context and the inclusion of any XML / TEI repository in the Open Linked Data environment.

3. Provide users with advanced formal tools to semantically define their interpretations of the texts they apply the markup to and give, in this way, the possibility of innovative computational processing based on semantics (reasoner and semantic query engines).

However, the cost and the practical complexity of such an extension are notable and several theoretical problems, format choices and implementation details are still to be defined. A possible candidate for a test-bed of the ideas presented in this paper

could be the forthcoming “TEI Simple” (formerly known as “TEI Nudge”, Mueller 2013) customization of the TEI scheme.

References

Burnard l. (2013). « Resolving the Durand Conundrum », Journal of the Text Encoding Initiative, 6. DOI: 10.4000/jtei.842

Barabucci, G., Di Iorio, A., Peroni, S., Poggi, F., Vitali, F. (2013). «Annotations with EARMARK in practice: a fairy tale». In Tomasi, F., Vitali, F. (Eds.), Proceedings of the 2013 Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities (DH-CASE 2013). New York, ACM. DOI: 10.1145/2517978.2517990

Cover, R, (1998). XML and semantic transparency. Technology report, Cover Pages. <http://www.oasisoprn.org/cover/xmlAndSemantics.html>.

Guarino, N. (2009). The Ontological Level: Revisiting 30 Years of Knowledge Representation. In Conceptual Modeling: Foundations and Applications. Lecture Notes In Computer Science, Vol. 5600. Springer-Verlag, Berlin, Heidelberg 52-67. <http://dx.doi.org/10.1007/978-3-642-02463-4_4>

Huitfeldt, C. & Sperberg McQueen, C. M. (2001). TexMECS: An experimental markup meta-language for complex documents. Working paper of the project Markup Languages for Complex Documents (MLCD), University of Bergen

Mueller, M. (2013). TEI-Nudge or Libraries and the TEI, Center for Scholarly Communication & Digital Curation Blog. <http://cscdc.northwestern.edu/blog/?p=872>

Peroni, S., Vitali, F. (2009). Annotations with EARMARK for arbitrary, overlapping and out-of order markup. In Proceedings of the 2009 ACM Symposium on Document Engineering (DocEng 2009): 171-180. New York, USA: ACM. DOI: 10.1145/1600193.1600232

Peroni, S., Gangemi, A., Vitali, F. (2011). Dealing with Markup Semantics. In Ghidini, C., Ngonga Ngomo, A., Lindstaedt, S., Pellegrini, T. (Eds.), Proceedings the 7th International Conference on Semantic Systems. New York, USA: ACM. DOI: 10.1145/2063518.2063533

Renear, A., M. Sperberg-McQueen, and C. Huitfeldt (2002). Towards a semantics for XML markup. In R. Furuta, J. I. Maletic, and E. Munson (eds.), DocEng’02: Proceedings of the 2002 ACM Symposium on Document Engineering, McLean, VA, New York, NY: ACM Press

Schmidt, Desmond. (2010). The inadequacy of embedded markup for cultural heritage texts, Literary and Linguistic Computing 25.2

Sperberg-McQueen, C. M. (1991), 'Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts', Literary and Linguistic Computing, 6.1.

Sperberg-McQueen C. M., Claus Huitfeldt, and Allen Renear. (2000). Meaning and interpretation of markup. In Markup Languages: Theory & Practice, 2 (3), MIT Press

Sperberg McQueen, C. M. & Huitfeldt, C. (2004). GODDAG: A Data Structure for Overlapping Hierarchies. In the Proceedings of the 8th International Conference on Digital Documents and Electronic Publishing, DDEP 2000 Munich, Germany

Sperberg McQueen, C. M. & Huitfeldt, C. (2008). Markup Discontinued: Discontinuity in TexMecs, Goddag structures, and rabbit/duck grammars. In the Proceedings of Balisage: The Markup Conference 2008. Montreal, Canada

Sperberg-McQueen, C. M., Marcoux, Y., Huitfeldt, C. (2009). Two representations of the semantics of TEI Lite. In Proceedings of Digital Humanities 2010. London, UK.