Two representations of the semantics of TEI Lite

Sperberg-McQueen, C. M.

cmsmcq@blackmesatech.com
Black Mesa Technologies LLC, USA

Marcoux, Yves

yves.marcoux@umontreal.ca
Université de Montréal, Canada

Huitfeldt, Claus

Claus.Huitfeldt@uib.no
Department of Philosophy, University of Bergen

Markup languages based on SGML and XML provide reasonably fine control over the syntax of markup used in documents. Schema languages (DTDs, Relax NG, XSD, etc.) provide mature, well understood mechanisms for specifying markup syntax which support validation, syntax-directed editing, and in some cases query optimization. We possess a much poorer set of tools for specifying the *meaning* of the markup in a vocabulary, and virtually no tools which could systematically exploit any semantic specification. Some observers claim, indeed, that XML and SGML are "just syntax", and that SGML/XML markup has no systematic semantics at all. Drawing on earlier work (Marcoux et al., 2009), this paper presents two alternative and complementary approaches to the formal representation of the semantics of TEI Lite: *Intertextual semantics* (IS) and *Formal tag-set descriptions* (FTSD).

RDF and Topic Maps may appear to address this problem (they are after all specifications for expressing "semantic relations," and they both have XML transfer syntaxes), but in reality their focus is on generic semantics — propositions about the real world — and not the semantics of markup languages.

In practice, the semantics of markup is most of the time specified only through human-readable documentation. Most existing colloquial markup languages are documented in prose, sometimes systematically and in detail, sometimes very sketchily. Often, written documentation is supplemented or replaced in practice by executable code: users will understand a given vocabulary (e.g., HTML, RSS, or the Atom syndication format) in terms of the behavior of software which supports or uses that vocabulary; the documentation for Docbook elevates this almost to a principle, consistently speaking not of the meaning of particular constructs, but of the "processing expectations" licensed by those constructs.

Yet a formal description of the semantics of a markup language can bring several benefits. One of them is the ability to develop provably correct mappings (conversions, translations) from one markup language to another. A second one is the possibility of automatically deriving facts from documents, and feeding them into various inferencing or reasoning systems. A third one is the possibility of automatically computing the semantics of part or whole of a document and presenting it to humans in an appropriate form to make the meaning of the document (or passage) precise and explicit.

There have been a few proposals for formal approaches to the specification of markup semantics. Two of them are *Intertextual Semantic Specifications*, and *Formal Tagset Descriptions*.

Intertextual semantics (IS) (Marcoux, 2006; Marcoux & Rizkallah, 2009) is a proposal to describe the meaning of markup constructs in natural language, by supplying an IS specification (ISS), which consists in a pre-text (or text-before) and a post-text (or text-after) for each element type in the vocabulary. When the vocabulary is used correctly, the contents of each element combine with the pre- and post-text to form a coherent natural-language text representing, to the desired level of detail, the information conveyed by the document. Although based on natural language, IS differs from the usual prose-documentation approach by the fact that the meaning of a construct is dynamically assembled and can be read sequentially, without the need to go back and forth between the documentation and the actual document.

Formal tag-set descriptions (FTSD) (Sperberg-McQueen et al., 2000) (Sperberg-McQueen & Miller, 2004) attempt to capture the meaning of markup constructs by means of "skeleton sentences": expressions in an arbitrary notation into which values from the document are inserted at locations indicated by blanks. FTSDs can, like ISSs, formulate the skeleton sentences in natural language prose. In that case, the main difference between FTSD and ISS is that an IS specification for an element is equivalent to a skeleton sentence with a single blank, to be filled in with the content of the element. In the general case, skeleton sentences in an FTSD can have multiple blanks, to be filled in with data selected from arbitrary locations in the document (Marcoux et al., 2009). It is more usual, however, for FTSDs to formulate their skeleton sentences in some logic notation: e.g., first-order predicate calculus or some subset of it.

Three other approaches, though not directly aimed at specifying markup semantics, use RDF to express document structure or *some* document semantics, and could probably be adapted or extended to serve as markup semantics specification formalisms. They are *RDF Textual Encoding Framework* (RDFTef) (Tummarello et al., 2005) (Tummarello et al., 2006), EARMARK (*Extreme Annotational RDF Markup*) (Di Iorio et al., 2009), and GRDDL (*Gleaning Resource Descriptions from Dialects of Languages*) (Connolly, 2007).

RDFTef and EARMARK both use RDF to represent complex text encoding. One of their key features is the ability to deal with non-hierarchical, overlapping structures. GRDDL is a method for trying to make parts of the meaning of documents explicit by means of an XSLT translation which transforms the document in question into a set of RDF triples. GRDDL is typically thought of as a method of extracting meaning from the markup and/or content in a particular document or set of documents, rather than as a method of specifying the meaning of a vocabulary; it is often deployed for HTML documents, where the information of most immediate concern is not the semantics of the HTML vocabulary in general, but the implications of the particular conventions used in a single document. However, there is no reason in principle that GRDDL could not be used to specify the meaning of a markup vocabulary apart from any additional conventions adopted in the use of that vocabulary by a given project or in a given document.

If proposals for formal semantics of markup are scarce, their application to colloquial markup vocabularies are even scarcer. Most examples found in the literature are toy examples. A larger-scale implementation of RDFTef for a subset of the TEI has been realized by Kepler (Kepler, 2005). However, as far as we know, no complete formal semantics has ever been defined for a real-life and commonly used colloquial vocabulary. This paper reports on experiments in applying ISSs and FTSDs to an existing and widely-used colloquial markup vocabulary: TEI Lite.

Developing an ISS and an FTSD in parallel for the same vocabulary is interesting for at least two reasons. First, it is an opportunity to verify the intuition expressed in Marcoux et al., 2009 that working out ISSs and FTSDs involves much the same type of intellectual effort. Second, it can give insight into the relative merits and challenges of natural-language vs logic-based approaches to semantics specification.

The full paper will focus on the technical and substantive challenges encountered along the way and will describe the solutions adopted.

An example of a challenge is the fact that TEI Lite documents can be either autonomous or transcriptions of existing exemplars. Both cases are treated with the same markup vocabulary, but ultimately, the meaning of the markup is quite different: in one case, it licences inferences about the marked-up document itself, while in the other, it licences inferences about the exemplar. The work reported in Sperberg-McQueen et al., 2009 on the formal nature of transcription is useful here to decide how to represent statements about the exemplar, when it exists. However, the problems of determining whether any particular document is a transcription or not, and of putting that fact into action in the generation of the semantics remain. One possible solution is to consider as external knowledge the fact that the document is a transcription. In the FTSD case, that external knowledge would be represented as a formal statement that could then trigger inferences about an exemplar. In the ISS case, it would show up as a preamble in the pre-text of the document element. Another solution is to consider the transcription and autonomous cases as two different application contexts of the vocabulary, and define two different specifications. The benefits and disadvantages of the two solutions will be discussed.

Follow-on work will include developing a GRDDL specification of TEI Lite, and comparing it to the ISS and FTSD. It will also include the elaboration of tools to read TEI Lite-encoded documents and generate from them either a prose representation of the meaning of the markup (from the ISS) or a set of sentences in a formal symbolic logic (from the FTSD). We also expect to induce a formal ontology of the basic concepts appealed to by the three formalisms and attempt to make explicit some of the essential relations among the concepts in the ontology: What kinds of things exist in the world described by TEI Lite markup? How are they related to each other?

References

Connolly, Dan. *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*. http://www.w3.org/TR/grddl/.

Di Iorio, A., Peroni, S., Vitali, F. (2009). 'Towards markup support for full GODDAGs and beyond: the EARMARK approach'. *Balisage: The Markup Conference*. Montréal, Canada, August 2009. Balisage Series on Markup Technologies. 3 vols. doi:10.4242/BalisageVol3.Peroni01.

Kepler, F. N. *RDF Textual Encoding Framework*. http://sourceforge.net/projects/rdftef/.

Marcoux, Y. (2006). 'A natural-language approach to modeling: Why is some XML so difficult to write?'. *Extreme Markup Languages*. Montréal, Canada, August 2006. http://conferences.idealliance.org/extreme/html/2006/Marcoux01/EML2006Marcoux01.htm.

Marcoux, Y., Rizkallah, É. (2009). 'Intertextual semantics: A semantics for information design'. *Journal of the American Society for Information Science & Technology*. **Volume 60, Issue 9**: 1895-1906. doi:10.1002/asi.21134.

Marcoux, Y., Sperberg-McQueen, C. M., Huitfeldt, C. (2009). 'Formal and informal meaning from documents through skeleton sentences: Complementing formal tag-set descriptions with intertextual semantics and vice-versa'. *Balisage: The Markup Conference*. Montréal, Canada, August 2009. doi:10.4242/BalisageVol3.Sperberg-McQueen01.

Sperberg-McQueen, C. M., Huitfeldt, C., Marcoux, Y. (2009). 'What is transcription? (part 2) (abstract)'. *Digital Humanities 2009 Conference Abstracts, June 2009*. Pp. 257-260. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.

Sperberg-McQueen, C. M., Huitfeldt, C., Renear, A. (2000). 'Meaning and Interpretation of Markup: Not as Simple as You Think'. *Extreme Markup Languages 2000*. Montréal, Canada, August 2000.

Sperberg-McQueen, C. M., Miller, E. (2004). 'On mapping from colloquial XML to RDF using XSLT'. *Extreme Markup Languages 2004*. Montréal, Canada, August 2004. http://conferences.idealliance.org/extreme/html/2004/Sperberg-McQueen01/EML2004Sperberg-McQueen01.html.

Tummarello, G., Morbidoni, C., Kepler, F., Piazza, F., Puliti, P. (2006). 'A novel Textual Encoding paradigm based on Semantic Web tools and semantics'. *5th edition of the International Conference on Language Resources and Evaluation*. Genoa, Italy, May 2006. http://www.sdjt.si/bib/lrec06/pdf/225_pdf.pdf.

Tummarello, G., Morbidoni, C., Pierazzo, E. (2005). 'Toward Textual Encoding Based on RDF'. *9th ICCC International Conference on Electronic Publishing*. Leuven-Heverlee, Belgium, June 2005. http://elpub.scix.net/data/works/att/206elpub2005.content.pdf.