Umanistica Digitale - ISSN:2532-8816 - n.3, 2018
F. Ciotti – Towards a Formal Ontology for the Text Encoding Initiative
DOI: http://doi.org/10.6092/issn.2532-8816/8174

Towards a Formal Ontology for the Text Encoding Initiative

Fabio Ciotti

Università di Roma "Tor Vergata"
fabio.ciotti@uniroma2.it

Abstract. In questo articolo viene presentata una proposta preliminare di ontologia per la rappresentazione dello schema Text Encoding Initiative (TEI). Le motivazioni e i benefici di una versione semantica e machine-readable della TEI sono molteplici: (1) l'interoperabilità semantica tra codifiche di testi comporterebbe la facilitazione dell'interrogazione cross-corpora e l'integrazione di Linked Open Data; (2) la formalizzazione del TEI *abstract model* apporterebbe migliorie significative in termini di consistenza e *soundness*. Data la complessità dello schema TEI, viene qui preso in considerazione un consistente ma ristretto sottoinsieme di elementi e attributi. In secondo luogo la meta-ontology EARMARK viene esaminata. In ultimo, vengono forniti alcuni dettagli implementativi.

This article presents the rationale and the proposal of a preliminary architecture of a formal ontology of the Text Encoding Initiative markup language. The reasons to have a formal and machine-readable semantics for TEI are manifold. In the first place, it would have a number of pragmatic and technical benefits, like better support for semantic interoperability in text encoding practices, easier cross-corpora query processing, seamless integration with Linked Open Data ecosystem. In second place, it would give a formalized account of the quasi-formal notion of the TEI *abstract model*, fostering the consistency and soundness of the TEI model. Given the complexity of the TEI encoding schema, specifying formally such an ontology will be a time consuming intellectual activity: in a first stage, we propose to limit its scope to a well-defined subdomain of the TEI, and to build it adopting pre-existing meta-ontology like EARMARK. The final part of the article gives some preliminary details of this design.

Introduction[1]

The Text Encoding Initiative markup language represents one of the most significant achievements of the Digital Humanities field and is now universally accepted as the standard formalism for the creation of textual digital resources in humanistic research and scholarship. One of the reasons of its success is the fact that it is based on the XML metalanguage, a sound and simple standard for data modeling and serialization.

There are many theoretical, pragmatic and social reasons for the wide and enduring acceptance of the TEI/XML couplet, notwithstanding the many criticisms and shortcomings. For example:

- XML is relatively easy to learn and use compared to other computer languages, especially if the complexity level of the encoding is low or medium;
- XML encoding affordances are similar to those of traditional textual annotation, a familiar practice to the average humanist;
- XML data format is portable (especially in the editing phase) between different platforms;[2]
- XML processing leaves to the user control on the editing process and on the resulting visualizations;
- XML introduces data quality control in text processing via its internal syntax and schema based parsing facilities;
- XML is flexible enough to accommodate a vast range of humanistic users requirements;
- XML has a good ecosystem of related standards and open source applications.

On the other hand, it is worth pointing out that, even if the TEI is an XML based language, its evolution has somewhat led to a certain level of abstraction from that language and, to some extent, from its underlying tree data model. We must remember, in fact, that it is possible to draw a neat distinction in the usage of XML language: it can be adopted as a full-fledged formal modeling language, in which case we accept the underlying tree data model as a good way to formally represent the object domain; but it can also be used as a mere syntax facility, a serialization language that is independent from the actual data model we are using to represent our domain (as it happens in the XML syntax of languages like RDF and OWL). The TEI, in

1 This article presents the results of a collaborative effort of the author with Francesca Tomasi, Fabio Vitali and Silvio Peroni, in order to develop an OWL 2 ontology to formally define the semantics of the Text Encoding Initiative. Some preliminary steps and the general context of this effort have already been presented at the TEI Conferences in 2014 and 2015 and in article published on the *Journal of the Text Encoding Initiative* (8.; see also 7.). The authorial responsibility of the present work is nonetheless to be attributed solely to Fabio Ciotti, with the exception of section "How: the architecture of the TEI ontology", that has seen the contribution of Silvio Peroni.

2 At least as far as this portability is limited to a purely syntactic level, since the limits of XML for semantic portability is precisely one of the reasons that motivates our proposal.

the course of its evolution, has moved from a modeling orientated usage of XML to a syntactic oriented usage of XML.

This progressive shift has been determined on the one hand by the need to represent many not hierarchical features of textuality, and on the other by the quest for a more semantic oriented way of modeling textual features. In fact, the common belief that XML markup expresses semantic information is technically flawed, in that XML *per se* is only a syntactic language to represent a tree based data model (4.; 12.).

Starting from the groundbreaking work of Renear, Huitfeldt and Sperberg-McQueen 28., various efforts to formalize the semantic role of markup languages have been made in the past 20 years (17.; 33.; 30.; 24.; 29.).[3] However, none of them has reached a mature state and has produced an operational solution, mostly because of the lack of maturity of the enabling formalisms and technologies adopted, and of the lack of support from the community of users. Building on the achievements (and limits) of these previous efforts, we believe that the Semantic Web stack of languages and frameworks could provide a viable and balanced solution to the theoretical and pragmatic requirements for developing a formal semantic component for the TEI.

The rest of this article is devoted to an overall illustration of this proposal, and is divided in three parts that can be conveniently entitled "Why", "What" and "How". Each of them tries to answer to some basic questions that give shape to our proposal:

- Why: why do I think that the idea of giving TEI a formal semantics is a good idea, and how could it enhance its expressive power and hence its usefulness for the community?
- What: what in the TEI do I really think can conceivably be formalized in the form of an ontology? How far can we imagine going in this direction?
- How: which are the better technical and formal strategies to build a semantic model of the TEI subset we have identified in the step before?

Why *ontologize* TEI?

The reasons to have a formal and machine-readable semantics for TEI are manifold. We can divide them into two main categories: technical and pragmatic reasons, that are universally applicable to any XML markup language; and theoretical reasons, that are particularly relevant for the TEI and its domain of application.

3 Since this intellectual history has been satisfactorily described in various preceding works (in part. 24.; 8.) I am not going to dwell on it.

Technical and pragmatic reasons

The pragmatic and technical benefits of a formal definition of XML markup languages have been already pointed out in many previous works dedicated to this topic that dates back to the mid-90s (for a survey of the literature see 30.; 28.; 24.). Here is a brief summary of those arguments:

- enabling parsers to perform both syntactic and semantic validation of document markup;
- enabling the automatic inference of facts from documents by means of inference systems and reasoners;
- simplifying the federation, conversion and translation of documents marked up with different markup vocabularies;
- allowing users to query upon the structure of the document considering its semantics;
- creating visualizations of documents on the base of the semantics of their structure rather than the specific vocabulary in which they are marked up;
- increasing the accessibility of documents' content, even in the case of tag abuse, i.e., "using markup constructions in ways other than intended by the language designer";
- promoting a more flexible software design for those applications that use markup languages, guaranteeing a better maintainability even when markup language schema evolves.

The advantages envisioned in this list are neither TEI specific nor related to inter-markup languages relationships, but some of the issues have a special relevance for TEI and for the usage of TEI inside its reference community.

Let us take for instance the query issue. It is well known that, despite the recent efforts to give more constraints in the usage of the markup of the more common features, the *Guidelines* allow for many ways of expressing one and the same textual feature in TEI markup. As a consequence, the alleged interoperability enabling (or supporting) role of TEI (that is Text Encoding for *Interchange*), is rather undermined 1.. The possibility of having a set of ontological definitions of the XML markup expression, that is, a set of shared formal definitions of the textual features to which any single encoding project could bind idiosyncratic markup usage (keeping safe the need and the right to fine tune the encoding at local level), could help solve this problem. In fact, the end-user could express the search query in abstract terms (search for all sonnets that contain the word 'love') and on the base of the ontology an automatic engine could generate all the queries in one or more specific XML query language (for instance XQuery), as sketched in Figure 1.[4]

4 In this example the query "Search 'love' in sonnets" does not have to be taken necessarily as expressed in natural language: the point is searching in "sonnets" however they are encoded.

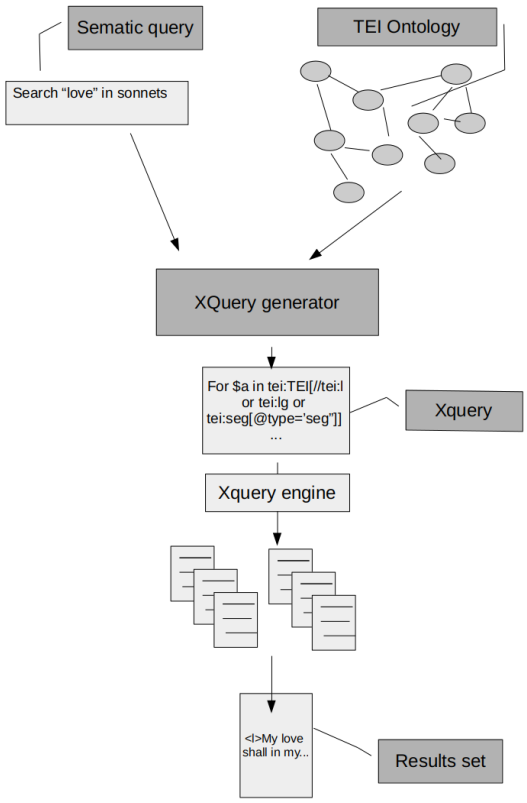

Figure 1: The schematic flow of a query generation based on a markup ontology

This result could be in principle attained adopting different technologies, but a semantic approach has some exclusive pros:

1) it is independent from the implementation;

2) it is more expressive than the average ad hoc solutions;

3) it can take advantage of inference engines' capabilities to extend or refine the query without previous knowledge of the details of encoding practices.

Point 3) can explain why we could use the notion *sonnet* in our imagined semantic query, even if it is not explicitly defined and encoded inside the TEI document as such: provided the existence of a TEI ontology defining the notion of poem it would be "sufficient" to extend that ontology with a sonnet class, that has the restriction of being composed of two quatrains and two triplets.[5]

Another useful application of a sound and well-defined semantics for TEI XML markup is the

5 The reality would be more complex than this, but the advantage is that once defined the new class it could be used in each successive query.

possibility to define interoperability relationship between TEI data sets and other data and metadata models and languages working directly at the abstraction level of the ontology and not at the level of the documents, or of the XML schema, where lots of difficulties arise.[6]

Theoretical reasons

In the previous section we have identified some of the pragmatic benefits that the availability of a computational semantics of TEI markup (*inter alia*) could provide. But we envision also some deeper theoretical and foundational advantages in the idea of a semantic model for TEI.

It is a commonly acknowledged notion that the very core of digital methods' application in humanities research is the notion of model/modeling. The terms couple *model/modeling* are understood in many different ways in the community (23.; 18.; 10.; 9.). In this context, TEI is not only a markup facility but first and foremost a conceptual model of textuality. In fact, in the *Guidelines* we can even find an explicit statement asserting this, when we find the definition of the important concept of *TEI abstract model* (32.: chap. 23):

> The TEI Abstract Model is the conceptual schema instantiated by the TEI Guidelines. These Guidelines define, both formally and informally, a set of abstract concepts such as 'paragraph' or 'heading', and their structural relationships, for example stating that 'paragraph's do not contain 'heading's. These Guidelines also define classes of elements, which have both semantic and structural properties in common. Those semantic and structural properties are also a part of the TEI Abstract Model…

The notion of an *abstract model* is used in many formal procedures: for instance, in the assessment of TEI conformance, or in the definition of *Schematron* rules that constrain the usage of some elements (32.: chap. 23):

> It is an important condition of TEI conformance that elements defined in the TEI Guidelines as having one specific meaning should not be used with another… The semantics of elements defined in the TEI Guidelines are conveyed in a number of ways, ranging from formally verifiable data types to informal descriptive prose

The problem here is that this very notion, although used extensively in many formal procedures related to TEI definition and usage, is *not* formally defined. This ends up in a lot of problems and circularities. As is well known, Alan Turing, in the quest of a solution for the *Entscheidungsproblem*,[7] had to find a formal equivalent of the intuitive notion of *algorithm* or *effective procedure* or *calculus*, since it is not possible to mix formal and non- or quasi-formal

6 In this sense, this ontology should play the same role that CIDOC CRM framework aims to play in the context of Cultural Heritage metadata interoperability 16..

7 The *Decision problem* is the problem to find an algorithmic procedure to asses mechanically the validity (or non validity) of a first order logical formula, proposed by Hilbert in 1928, to solve which Turing developed the notion of *algorithmic machine*, later known as *Turing machine*, and the rationales of the theory of computation 13..

notions in a formal argument (and from this emerged the concept of what we call a *Turing machine*). We need to make the same conceptual move for the quasi-formal notion of the *TEI abstract model*, if it has to be of any use other than a sort of regulatory principle. Of course, this means that we have to accept that something will "be lost in translation": this formal model is not necessarily what text really is, but how in the TEI we model some core aspects of the notion of text for the purposes of computation.

The adoption of Semantic Web formalisms to define this abstract conceptual model 8. gives us the possibility of expressing the TEI in a well-defined data model that can accommodate, at least to some extent, the *plurality* of textuality. This formal ontology, independent from any serialization format, could in perspective become the 'real TEI', from which a set of serializations can be derived in any language of choice (14.; 15.).[8]

What part of the TEI can be ontologized?

TEI scheme as whole is very complex and diverse, the result of decades of work, refinements, extensions, additions; it covers in details many different areas of application. We acknowledge that it is impossible to reduce this fuzzy cloud to a unique formal semantic definition.

Moreover, TEI real usage in the community is largely influenced by pragmatic factors. That is, the intended meaning of the markup in concrete markup acts is determined by the circumstances of usage, the context in which the markup happens, the presuppositions of the encoder himself. This has produced a loose conglomerate of applications, mutually related by a sort of "family resemblance" (in Wittgensteinian sense). The *Guidelines* 32. themselves, in spite of the efforts, present errors and inconsistencies partly because of intrinsic difficulties in editing a huge reference book, partly because of some design principles that informed its very development.[9]

It is impossible to reduce to a unique formal semantic definition this cloud. However, we can identify a subset of shared assumptions, a common ground of notions about the role and meaning of TEI markup and the nature of document like objects: this common set constitutes what we could call the *core ontology*, and can be the object of an ontological formalization.[10]

8 If we accept a slightly weakened notion of interoperability this formal semantics could also give an operational solution to that problem.

9 Take for example the choice of having very abstract generic identifiers for structural elements, like or <lg>, instead of more determined ones like *<chapters> or *<stanzas>. Here the difference is certainly a matter of degree of specification, as one of the reviewer of this article has rightly observed; nonetheless it is hard to deny that the term 'stanza' is less open to interpretation than 'lines group'.

10 It is worth noting that it's hard to assert in a definitive way that the TEI abstract model is co-extensive with this core semantics; there are probably some areas that fall outside

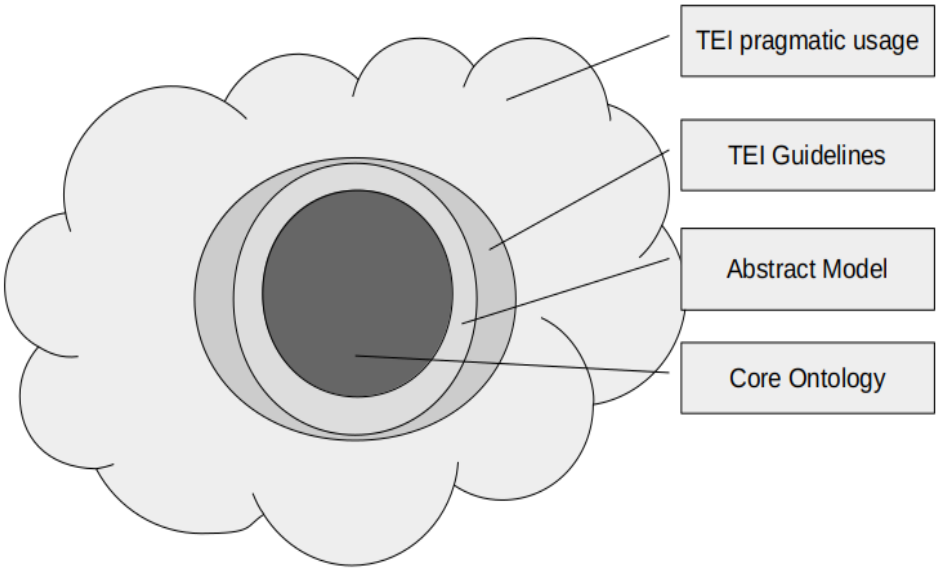

Figure 2: The TEI cloud

For many diverse reasons we suggest that *prima facie* we can assume the *TEI simplePrint* customization element set as a satisfying approximation of this common ontology (34.; 32.). The choice to build our formalization effort on the foundations of *TEI Simple*, is obviously dictated also by practical reasons, but it is not an opportunistic *ad hoc* choice, as it may seem.

One of the most common definition of computational ontology is: 'An ontology is a formal, explicit specification of a shared conceptualization' (31.: 184). This definition differs from the original and most known one given by Gruber 20. in that it stresses the 'shared' aspect of the conceptualization, that is fundamental for its successful adoption, as Nicola Guarino notes (21.: 14):

> For practical usage of ontologies, it turned out very quickly that without at least such minimal shared ontological commitment from ontology stakeholders, the benefits of having an ontology are limited. The reason is that an ontology formally specifies a domain structure under the limitation that its stakeholder understand the primitive terms in the appropriate way. In other words, the ontology may turn out useless if it is used in a way that runs counter to the shared ontological commitment. In conclusion, any ontology will always be less complete and less formal than it would be desirable in theory.

TEI Simple has been defined by a group of domain experts that have analyzed the actual usage of markup in some big textual repositories, and have selected and partially organized a set of

one hundred or so elements, that can describe all the textual features represented by TEI markup in those documents 5.:

> The TEI simplePrint schema […] sought to define a new *highly-constrained and prescriptive subset* of the Text Encoding Initiative (TEI) Guidelines suited to the representation of early modern print materials, a formally-defined set of processing rules which permit modern web applications to easily present and analyze the encoded texts, mapping to other ontologies, and processes to describe the encoding status and richness of a TEI digital text. Its choice of elements reflected the practices followed in the encoding of large-scale literary archives, notably those produced by the Text Creation Partnership. Practice of other comparable archives such as the German Text Archive was also taken into account.

This process fits perfectly in the refined definition of ontology. Another relevant aspect of the TEI simplePrint definition is that the markup selection process has tried to assign one and only one textual feature to each markup item. This unequivocal definition is useful to assure the fulfilling of the ontological commitment to an intensional specification of the domain required by ontological design (21.: 8–9).

How: the architecture of the TEI ontology

The last part of this article describes the overall design and methodological rationales of our ontological modeling of the TEI simplePrint. The design requirements for building such ontology are the following:

- the ontology should express at the same time an abstract characterization of TEI simplePrint elements' and attributes semantics and intended meaning and an ontological definition of their structural role in the document;
- the ontology should define a precise semantics of the elements having a clear characterization in the official TEI documentation (e.g., the element "<p>"), while it should relax the semantic constraints if the elements in consideration can be used with different connotations depending on the context (e.g., the element "<seg>"), allowing for further specification by users;
- it should be possible to extend the ontology, reuse it and define alternative characterizations of elements' semantics without compromising the consistency of the ontology itself;
- where possible existing ontologies or meta-ontologies should be reused.

In accordance with these overall principles, the TEI ontology can be implemented as an OWL 2 ontology based on preexisting and *ad hoc* defined ontology modules. The specification of

markup semantics for the various TEI simplePrint elements is done by means of EARMARK class and properties. The Extremely Annotational RDF Markup (25.; 14.) is at the same time a markup metalanguage, that can express both the syntax and the semantics of markup as OWL assertions, and an ontology of markup that makes explicit the implicit assumptions of markup languages (and, in particular, of the hierarchy of XML-based languages), providing a finer specification of the properties of markup, up to and including the possibility of toggling on and off the strict hierarchy of XML instantiations.

EARMARK is suitable for expressing markup semantics straightforwardly. However, we want to associate coherent semantics to markup items following precise and theoretically-founded principles. LA-EARMARK 24. is an extension of EARMARK with the *Linguistic Act* ontology, a module of the *Linguistic Meta-Model* 26.. LA attempts to provide a formal representation of the fundamental structure of a linguistic act according to the classical semiotic triangle model and providing the OWL 2 formalizations of the notions of:

- *linguistic act*: any communicative situation including information entities, agents, meanings, references, and a possible spatiotemporal context;
- *information entity*: any symbol that has a meaning, or denotes one or more references. They can be natural language terms, sentences or texts, symbols in formal languages, icons, or whatever device can be used as a vector for communication;
- *meaning*: any (meta-level) object that explains something, or is intended by something, such as linguistic definitions, topic descriptions, lexical entries, logical constraints, etc. They can be *interpretants* for information entities, and *conceptualizations* for individuals and facts;
- *reference*: any (set of) individual or fact from the world we are describing. They can have interpretations (creating meanings) and can be denoted by information entities.[11]

In this context, a markup construct can be seen as a kind of information entity with its own expression, and it becomes possible to express and assess facts, constraints and rules about the markup structure as well as about the inherent semantics of the markup elements themselves.

In our ontology any TEI XML element is expressed as an Earmark class *earmark:Element*. For instance, the TEI <p> element is defined as follows:[12]

```
Prefix earmark: <http://www.essepuntato.it/2008/12/earmark#>
Prefix co: <http://purl.org/co/>
Prefix tei: <http://www.tei-c.org/ns/1.0/>

Class: tei:p a
earmark:Element that
       earmark:hasGeneralIdentifier "p" and
       earmark:hasNamespace "http://www.tei-c.org/ns/1.0"
```

11 This description is taken and slightly adapted from 24..

12 All ontology examples are expressed in OWL 2 *Manchester Syntax* (Horridge and Patel-Schneider, 2009)

If and when we need to identify and characterize semantically some subsets of one element type defined by the schema we introduce appropriate restrictions of its EARMARK class by the way of the properties provided by the Collections Ontology (CO), which defines as OWL classes unordered and ordered collections 6.. The need for this characterization is determined by the fact that in the TEI schema there are many elements that can appear in different regions of the document tree, with distinct functional and semantic roles. The most apparent example is the element <p>, that is used in its 'proper' role of designating a block of prose in the textual content of the XML document, as in every context where there is the need to include some free form textual description or metadata – be it inside the subcomponents of the TEI Header or not. In our ontological model we have differentiated those usages of the XML element. For instance, the class of all the elements <p> that occur inside the <text> element and not inside the <teiHeader> is expressed as follows:

```
Class: tei:pText
EquivalentTo:
        earmark:Element that
        earmark:hasGeneralIdentifier "p" and
        earmark:hasNamespace "http://www.tei-c.org/ns/1.0" and
        co:elementOf some (
                        earmark:Element that
                        earmark:hasGeneralIdentifier "text" and
                        earmark:hasNamespace "http://www.tei-c.org/ns/1.0")
```

The assignment of semantic properties to the XML elements classes defined in EARMARK is expressed by the mean of *punning* meta-modeling facility introduced in OWL 2 19.. *Punning* is a mechanism introduced in order to assign properties to classes, retaining the decidability and tractability of the formalism 3.. This is achieved introducing the possibility to use an Internationalized Resource Identifier (IRI) inside the ontology both as a class name and an individual name, on the base of the context. Adopting punning, the semantic role assigned to one element can be expressed by the way of classes defined in other external ontologies or ontological modules.

The *TEI Semantics Ontology* (TSO) is the ontological component that specifies the general intended semantics of the TEI elements (e.g., the fact that an element is a paragraph rather than a section, a personal name reference rather than a geographical reference).[13] This component is the core of our modeling effort and its development is still underway. Its definition is based on a typological categorization of the TEI simplePrint elements set, based on their logical function and intended meaning and independently from their structural position in the tree data model underlying the markup language. For the moment, the role of XML attribute has not been taken into account for the definition of this ontological component. The

13 The design principles that has governed its development have been influenced by the definition of the Document Component Ontology 11.. The Document Component Ontology is part of the SPAR Ontologies, a set of ontologies devoted to the semantic description and processing of document like objects: http://www.sparontologies.net. In a further step some classes from DCO will be imported inside TSO, or will be mapped via an equivalence relationship.

possibility to use EARMARK facilities to specialize element classes partially fulfills the necessity to map specific element/attribute pairs to their intended meaning, but this issue surely needs a deeper analysis and exploration.

The main class of TSO is *tso:documentEntity* which defines the class of all TEI documents. All the remaining classes are defined as subclasses of this class and are concepts expressed by TEI elements, although there is not a one to one relationship between XML elements and OWL semantic classes.

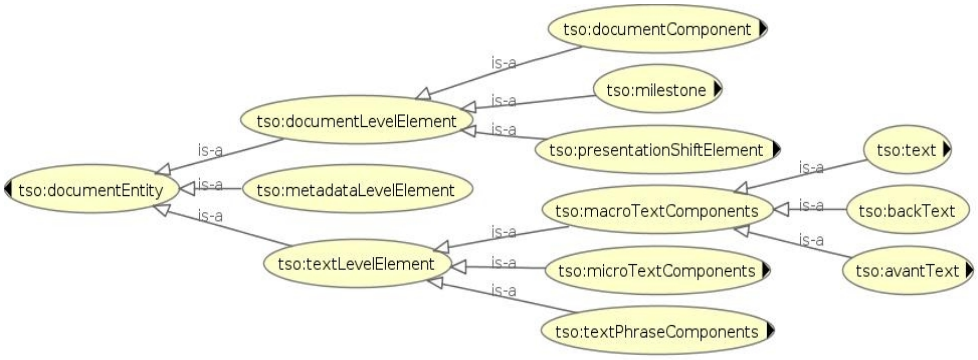

Figure 3: The graph of the higher levels of TSO

As is shown in Fig. 3, the overall class hierarchy of this ontology component is divided into three main subclasses:

- *tso:documentLevelElement*: the class of the elements related to the material document structure. Its subclasses define the features and characteristic pertaining to the text carrier that can be expressed by the XML markup
- *tso:metadataLevelElement*: the class of the elements related to the metadata of the document
- *tso:textLevelElement*: the class of the elements related to the textual structure and features modeled by the XML markup. It is farther subdivided into macro textual divisions (peritextual sections, chapters, sections etc.), micro textual blocks (paragraphs, verse, stanzas, epigraphs, abstract, quotations, lists, etc.) and phrase level components (linguistic features, editorial features, emphasis and other kind of distinctive phrases, etc.).

The link between the EARMARK class that identifies a specific element and its related semantic characterization is expressed by applying the punning to that class (so that it becomes an *Individual*) and then adopting the LA property *semiotics:expresses* that is used to specify the intensional meaning attributed to a linguistic (and a markup, *inter alia*) construct. For instance, we can use it to say that the <p> class defined above expresses the structural semantics of being a paragraph in TSO, as follows:

```
Individual: tei:pText
       Facts:
       semiotics:expresses tso:paragraph
```

If necessary, the associations of semantics to markup elements can be contextualized according to a particular agent's point of view in order to provide provenance data pointing to the entity that was responsible for such specification. This is possible by means of the properties provided by the Linguistic Act Ontology included in LA-EARMARK that allow to assign agency and responsibility to all these markup-to-semantics relations, as proper linguistic acts done by someone. This feature of the meta-ontology framework provides the possibility of project or even user defined customizations and refinements of the ontology.

Conclusions and further steps

The proposal to adopt a formal ontology approach to define formally the TEI that we have presented here is admittedly a proof of concept. As we have clearly stated, it is probably not possible to provide a complete and sound formalization of the whole TEI, as is pragmatically used by the community. This objective can be attained for smaller or locally defined subsets, as is the TEI simplePrint. Even with this restriction in order to develop a semantic framework that can express all the constraints and conditions that, some substantial refinements and improvements are required, of which the most relevant are:

- introducing in our modeling a complete (as far as possible) definition of the XML attributes (this is a rather complex area, since most attribute in TEI have ambiguous, multiple or undetermined definitions);
- refining and factorizing the TEI Semantics Ontology component;
- extending the ontological modeling to some other subsets of the TEI that are suitable for formalization. Simple is not all, and with appropriate time and work force the ontology can be extended to some other area of the TEI;
- devising an elegant and easy to use formalism to allow for project specific modifications and specializations of the element ontology – to be expressed via ODD – and for document specific idiosyncratic uses of markup.

These improvements are actually out of the scope and even of the possibility of our small research team, and they need a larger and organized research program, that should involve a much larger part of the TEI community.

In the long term, the formalisms we have adopted in our ontological modeling could evolve to become the principal formalization of the Text Encoding Initiative, independent of any serialization. EARMARK allows for expressing structural constraints in an ontologically precise definition, and can even instantiates the markup of a text document as an independent OWL document outside of the text strings it annotates, permitting a native stand-off markup strategy.

Through appropriate OWL restrictions it can define structures such as trees or graphs and can be used to generate validity constraints (including contextual constraints currently unavailable in most validation languages). The assessment of ontological properties in the semantic domain is in many ways comparable to validation in the XML domain. Moving from a syntactical perspective to a semantic one – as proposed by EARMARK – opens new perspectives for a general approach to assessment as well.

A key point of such approach is the translation of many markup properties from a syntactical to an ontological level. In the case of XML schema validation, for instance, this means expressing schema definitions as ontology classes and properties; and schema documents as ontology instances and assertions that expresses hierarchies as semantic relations. Starting from an ontological *TBox* representing the schema and an *ABox* 2. representing the document, we can then conclude that the document is valid according to the schema if and only if the ABox is consistent with the Tbox 15..

At this moment, XML is still probably the better strategy to encode digital texts in real word projects for many practical reasons. However, there is no reason for the TEI to be strictly based on it, as it is *de facto* now. Technical or pragmatic issues should not determine the choice of a formalization. Elena Pierazzo, at the time chair of the TEI Board, in her paper at the TEI conference 2015 in Lyon entitled "TEI: XML and Beyond" said: "The next few years will be crucial for the survival and expansion of the TEI: in order to survive and overcome the new challenges that come with the fast-evolving world of data representation it will have to part ways with XML as a sole technological implementation and while becoming more abstract offer concrete solutions for those problems that have accompanied its whole life".[14] The proposal drawn in this article can represent a small contribution to the TEI to envision the shape of its own future.

References

1. Bauman, S. (2011). Interchange vs. Interoperability. Proceedings of Balisage: The Markup Conference 2011 doi:10.4242/BalisageVol7.Bauman01.
2. Bergman, M. K. (2009). The Fundamental Importance of Keeping an ABox and TBox Split *AI³ Adaptive Information* http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/.
3. Bergman, M. K. (2010). Metamodeling in Domain Ontologies *AI³ Adaptive Information* http://www.mkbergman.com/913/metamodeling-in-domain-ontologies/#mm1.
4. Buzzetti, D. (2002). Digital Representation and the Text Model. *New Literary History*,

14 The small abstract of the talk is available at http://tei2015.huma-num.fr/en/papers/. Unfortunately, the talk has not been published and is not available in any more complete format.

33(1): 61–88.

5. Burnard, L. Mueller, M., Rahtz, S. Cummings, J., Turska, M. (2017). *An Introduction to TEI simplePrint*. http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_simplePrint.doc.html.

6. Ciccarese, P. and Peroni, S. (2014). The Collections Ontology: creating and handling collections in OWL 2 DL frameworks. *Semantic Web*, 5(6): 515–29 doi:10.3233/SW-130121.

7. Ciotti, F., Daquino, M. and Tomasi, F. (2016). Text Encoding Initiative Semantic Modeling. A Conceptual Workflow Proposal. In Calvanese, D., De Nart, D. and Tasso, C. (eds), *Digital Libraries on the Move: 11th Italian Research Conference on Digital Libraries, IRCDL 2015, Bolzano, Italy, January 29-30, 2015, Revised Selected Papers*. Cham: Springer International Publishing, pp. 48–60 http://dx.doi.org/10.1007/978-3-319-41938-1_5.

8. Ciotti, F. and Tomasi, F. (2016). Formal Ontologies, Linked Data, and TEI Semantics. *Journal of the Text Encoding Initiative,*9 doi:10.4000/jtei.1480. http://jtei.revues.org/1480.

9. Ciula, A. and Eide, Ø. (2016). Modelling in digital humanities: Signs in context. *Digital Scholarship in the Humanities* doi:10.1093/llc/fqw045.

10. Ciula, A. and Marras, C. (2016). Circling around texts and language: towards' pragmatic modelling' in Digital Humanities. *Digital Humanities Quarterly*, 10(3) http://www.digitalhumanities.org/dhq/vol/10/3/000258/000258.html.

11. Constantin, A., Peroni, S., Pettifer, S., Shotton, D. and Vitali, F. (2016). The Document Components Ontology (DoCO). *Semantic Web*, 7(2): 167–81.

12. Cover, R. (1998). XML and semantic transparency. *Cover Pages* http://xml.coverpages.org/xmlAndSemantics.html.

13. Davis, M. (2011). *The Universal Computer*. Natick: Taylor & Francis Inc.

14. Di Iorio, A., Peroni, S. and Vitali, F. (2010). Handling Markup Overlaps Using OWL. In Cimiano, P. and Pinto, H. S. (eds), *Knowledge Engineering and Management by the Masses*, 6317. (Lecture Notes in Computer Science). Springer Berlin Heidelberg, pp. 391–400 http://dx.doi.org/10.1007/978-3-642-16438-5_29.

15. Di Iorio, A., Peroni, S. and Vitali, F. (2011). A Semantic Web approach to everyday overlapping markup. *J. Am. Soc. Inf. Sci. Journal of the American Society for Information Science and Technology*, 62(9): 1696–716.

16. Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3): 75.

17. Dubin, D., Renear, A., Sperberg-McQueen, C. and Huitfeldt, C. (2003). A logic programming environment for document semantics and inference. *Literary and*

Linguistic Computing, 18(1): 39–47.

18. Eide, Ø. (2014). Ontologies, data modeling, and TEI. *Journal of the Text Encoding Initiative* doi:doi:10.4000/jtei.1191. https://jtei.revues.org/1191.
19. Golbreich, C., Wallace, E. K. and Patel-Schneider, P. (2009). OWL 2 Web Ontology Language new features and rationale. *W3C Proposed Recommendation*.
20. Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2): 199–220.
21. Guarino, N., Oberle, D. and Staab, S. (2009). What is an Ontology?. *Handbook on Ontologies*. Springer, pp. 1–17.
22. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F. and Rudolph, S. (2009). OWL 2 web ontology language primer. *W3C Recommendation*, 27(1): 123.
23. McCarty, W. (2005). *Humanities Computing*. London: Palgrave Macmillan.
24. Peroni, S., Gangemi, A. and Vitali, F. (2011). Dealing with markup semantics. Proceedings of the 7th International Conference on Semantic Systems. ACM, pp. 111–18 doi:10.1145/2063518.2063533.
25. Peroni, S. and Vitali, F. (2009). Annotations with EARMARK for arbitrary, overlapping and out-of order markup. Proceedings of the 9th ACM symposium on Document engineering. ACM, pp. 171–80 doi:10.1145/1600193.1600232.
26. Picca, D., Gliozzo, A. M. and Gangemi, A. (2008). LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge. Proceedings of the 6th Language Resource and Evaluation Conference.
27. Rahtz, S., Mueller, M., M., Pytlik-Zillig, B., Turska, M. and Cummings, J. (2015). TEI Simple Processing Model Specification. http://htmlpreview.github.io/? https://github.com/TEIC/TEI-Simple/blob/master/tei-pm.html.
28. Renear, A., Dubin, D. and Sperberg-McQueen, C. M. (2002). Towards a semantics for XML markup. Proceedings of the 2002 ACM symposium on Document engineering. ACM, pp. 119–26.
29. Sperberg-McQueen, C. M., Marcoux, Y. and Huitfeldt, C. (2014). Transcriptional Implicature: A Contribution to Markup Semantics. Digital Humanities 2014. Book of abstracts. pp. 360–62 https://dh2014.files.wordpress.com/2014/07/dh2014_abstracts_proceedings_07-11.pdf.
30. Sperberg-McQueen, C., Marcoux, Y. and Huitfeldt, C. (2010). Two representations of the semantics of TEI Lite. *Proceedings of Digital Humanities*: 7–10.
31. Studer, R., Benjamins, V. R. and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1): 161–97 doi:10.1016/S0169-023X(97)00056-6.

32. TEI Consortium (2016a). *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.0.0, Last Updated on 29th March 2016*. http://www.tei-c.org/Guidelines/P5.

33. Tummarello, G., Morbidoni, C. and Pierazzo, E. (2005). Toward Textual Encoding Based on RDF. From Author to Reader: Proceedings of the 9th ICCC International Conference on Electronic Publishing (ELPUB 2005). Leuven: Peeters, pp. 57–63 http://elpub.scix.net/data/works/att/206elpub2005.content.pdf.

34. Turska, M., Cummings, J. and Rahtz, S. (2016). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative,* 9.

Last consultation URLs: 28/11/2018