Journal of the Text Encoding Initiative

Issue 9 | September 2016 - December 2017
Selected Papers from the 2014 TEI Conference

Formal Ontologies, Linked Data, and TEI Semantics

Fabio Ciotti and Francesca Tomasi

Electronic version
URL: http://journals.openedition.org/jtei/1480
DOI: 10.4000/jtei.1480
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Fabio Ciotti and Francesca Tomasi, « Formal Ontologies, Linked Data, and TEI Semantics », *Journal of the Text Encoding Initiative* [Online], Issue 9 | September 2016 - December 2017, Online since 24 September 2016, connection on 01 May 2019. URL : http://journals.openedition.org/jtei/1480 ; DOI : 10.4000/jtei.1480

Formal Ontologies, Linked Data, and TEI Semantics

Fabio Ciotti and Francesca Tomasi

1. The Problem of XML Semantics

1 The debate on the semantic role of markup languages has been quite lively during the last twenty years and the experience of the TEI community has played an active role in it. It is commonly acknowledged that markup conveys semantic information, ranging from local "interpretations" produced by a single scholar to general and shared visions of "what is text, really" (De Rose et al. 1997).

2 However, the widespread notion that markup and, in particular, XML markup has a semantic role or conveys semantic content, is flawed in a relevant sense. As Robin Cover observed some years ago:

> XML is a poor language for data modeling if the goal is to represent information objects in the problem domain such that they correspond transparently ('one-to-one') to the user's conceptual model of objects in this domain.
>
> (Cover 1998)

3 XML is a powerful formalism to define the syntactic aspects of the markup language. It is based on formal grammars, making it possible to automatically parse a document and validate its formal structure against a predefined schema. Furthermore, XML implies a (meta) data model, the well-known *ordered tree*. This enables the adoption of the language as a modeling device for some structural relationships of document-like information objects:

- hierarchy (A contains B)
- ordered adjacency (A followed by B)
- co-occurrence (if A then B)

4 Some recent *schema languages* (W3C Schema in particular) have also introduced strong data typing for element content and attribute values. Nonetheless, XML does not provide a computational semantics to markup (or to data). The fallacy of the belief in XML semantic awareness is hidden by the fact that XML markup is *human-readable* and usually by design XML vocabularies are meaningful lexical items in a natural language. But the meaning of those vocabularies is not accessible by an XML processor. As far as an XML parser is concerned, `<title>Finnegans Wake</title>` and `<blob>Finnegans Wake</blob>` are both acceptable and well-formed markup instances.

5 XML owes its semantic value and consistency almost entirely to human interpretation and control. Any markup restriction or semantic role, accordingly, needs to be expressed in natural language as instructions or documentation for human users. The TEI Guidelines (TEI Consortium 2015) are one epitomizing example of that kind of documentation. A more formal and controlled approach is the *literate programming*–oriented[1] ODD formalism (Burnard 2013), developed by TEI in order to combine in a single meta-XML document the definition of an XML schema and all its relevant documentation.

2. The XML Semantic Debate

6 Several proposals have been drawn up to provide XML with formalized and computable semantics. The work of Renear, Dubin, Sperberg-McQueen, and Huitfeldt (e.g., Sperberg-McQueen, Huitfeldt, and Renear 2000 and Renear, Sperberg-McQueen, and Huitfeldt 2002) constitutes the first, explicit contribution in this direction. These authors, from the observation that semantic markup

coincides with the set of inferences authorized by one of its constructs, propose a formal markup semantics based on Prolog clauses. More recent works on the topic have proposed an RDF-based model for text encoding (Tummarello, Morbidoni, and Pierazzo 2005; Tummarello et al. 2006); explored the potential of an OWL vocabulary that represents some core semiotic notions, in order to provide a better understanding of the semantics of markup (Peroni, Gangemi, and Vitali 2011); and revisited the idea of "transcriptional implicature" (Sperberg-McQueen and Huitfeldt 2008; Sperberg-McQueen, Marcoux, and Huitfeldt 2010 and 2014).

7 In these studies the range of application possibilities offered by the definition of a formal semantics for markup is widely recognized and justified:

> 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.
>
> (Sperberg-McQueen, Marcoux, and Huitfeldt 2010)

8 Nonetheless the same authors observe that if proposals for formal semantic approaches to markup have been very scarce, their practical applications are even fewer.

9 The reasons for this lack of interest from the wider encoding community are manifold and complex:

- theoretical complexity in a domain already hard to understand for the average humanist scholar;
- technical and practical difficulties in the application and exploitation of the approaches proposed;
- lack of tools and applications;
- excessively "revolutionary" scope of some proposals.

3. A Formal Ontology Approach for TEI Semantics

10 Since the first discussions of the semantics of markup languages in the early '90s and today's situation, the Semantic Web (or Web 3.0) paradigm and, more recently, linked data have developed and spread. This process has made available a number of syntactically rigorous and semantically well-founded languages and data-models, to represent (Resource Description Framework, RDF), logically define (RDF Schema; Ontology Web Language, OWL 2), and query (SPARQL) semantic data. In parallel with this diffusion of Web 3.0 architectural formalisms, a good number of systems and software components aimed at semantic data processing (storage, query, and inference) have appeared, mostly in the open source domain.

11 Given this theoretical and technological context, we believe that a Semantic Web approach can represent the most viable solution, giving a formal definition to the implicit concepts underlying XML text encoding. In particular, in this paper we propose an *ontological* extension of the TEI framework to partially formalize the semantics of the markup constructs it provides.

12 The term ontology is used to designate a large and varied class of objects, ranging from controlled vocabularies and thesauruses to proper formal ontologies. In this paper we refer to the latter meaning: a formalized and shared account of a conceptual description of a domain (Gruber 2009).

13 In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.

14 In Semantic Web architecture, formal ontologies have the role of logically defining and constraining the terms of the descriptive languages adopted to state the semantic properties of information resources. This has an obvious parallel with the problem of assigning a semantics to markup languages like TEI.

15 The motivations for adopting a formal ontological approach in our proposal are both theoretical and practical. In the text-encoding domain (as in the entire field of Digital Humanities), models and modeling play a central methodological and foundational role. In the relevant literature we

commonly find assertions stating that text encoding is a kind of modeling. The very problem with models and modeling is that they are *umbrella terms*, relating to an ample and diverse set of conceptual objects and practices.

16 In our view the practice of ontological modeling is a good operationalist translation of the common definition of model. From the practical point of view, formal ontologies give access to a number of powerful computational tools and methods, like the application of inference and reasoning engines to analyze the textual data. Another relevant property of the Semantic Web ontology formalisms is the ability to compare and eventually merge different ontologies. This property has some desirable consequences in our context of application, especially in conjunction with another property of Semantic Web ontologies, the Open World Assumption (Patel-Schneider and Horrocks 2006). In brief, this assumption means that from the lack of an explicit assertion stating that some fact is true it does not follow that the fact is false (or its negation true), simply that it is not known. In a wider sense this means that knowledge can be incomplete and that it is possible that the attributes of a concept can be incomplete at any given moment. This property of a Semantic Web ontology is very helpful when the domain is very complex and subject to different points of view, and its modeling is conceived as a work in progress.

4. Rationales and Principles for a TEI Ontology

17 In order to avoid, as far as possible, the shortcomings of the earlier proposals for markup semantics, we have identified some guiding principles to drive the development of our ontology-driven approach.

1. To adopt well-established Semantic Web formalisms and technologies, in order to ensure that our approach is theoretically sound and practically implementable.
2. To extend, not attempt to replace, the current languages and practices; as long as Relax NG, W3C Schema, and ODD formalisms can do their job, there is no need to replace them. On the contrary, we believe that our ontological extension must cooperate and interoperate with the existing ecosystem of TEI technology.

3. To limit our ontological modeling *prima facie* to a subset of the TEI encoding scheme, the "core" markup facilities; it remains to be decided how to identify that core. The TEI community has historically tried to define subsets of the markup language to ease adoption and diffusion: TEI Lite (Burnard and Sperberg-McQueen 2012) and TEI Tite are well-known examples. We think that the best candidate for a test bed of the ideas presented in this paper is the TEI Simple customization of the TEI scheme (Turska, Cummings, and Rahtz 2016). The main reason for this choice is that this subset has been defined in a bottom-up approach from the analysis of the actual markup usage in some big text encoding projects. We are aware that this choice will leave aside some important and widely adopted parts of the TEI; nonetheless it is the best "ready-made" approximation of a subset of the TEI emerging from the wide and diverse pragmatic uses of the language in the community.
4. To provide a practical solution for some concerns that are relevant in the actual digital ecosystem in which TEI and XML live, especially interoperability and Linked Data exposition of XML TEI document content.
5. To appropriately distinguish between three different semantic levels expressed by the markup and its content:

    (a) general intensional semantics of TEI markup;
    (b) specific intensional semantics of TEI markup (defined by a particular user or community of practice);
    (c) extensional semantics of the markup elements.

18 The terms "intensional" and "extensional" semantics are to be taken in the meaning that was given to those terms by Carnap (1947): "intension" indicates the internal content (and structure) of a lexical term or concept, what constitutes its formal definition; and "extension" indicates the set of objects to which it can be applied.

19 Level (a) is defined by the maximal shared assumptions and usage practices of the TEI markup constructs, and it is translated into a general TEI ontology. For example, the meaning of the `<seg>` element is "a phrase level sequence of linguistic units." Level (b) relates to the particular interpretation of the markup adopted by a single or collective user. For example, in a given encoding project, the `<seg>` element might be used, for example, as a "linguistic manifestation

of a character's feature." Level (c) concerns the extensional semantics of the individual XML element's 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 items in a linked data set. The current TEI scheme already handles the case of simple extensional links with one or more external resources through the `@ref` attribute (whose value is one or more `xsd:anyURI`), or the `@key` attribute, as already evaluated (Ciula, Spence, and Vieira 2008). More complex relations with external semantic data could require complex stand-off markup structures (e.g., Huitfeld and Sperberg-McQueen 2001; Sperberg-McQueen and Huitfeld 2004; a general survey in Schmidt 2010).

20 The levels properly involved in our proposal are levels (a) and (b). Obviously the extensional semantics is not totally independent from the ontological and intensional levels, and we will need to find proper methods to formally link all those levels (cf. Section 7). This paper, though, will only sketch the main picture of our general proposal, since many technical details are still to be defined.

5. TEI General Ontology

21 The first level, TEI general ontology, formalizes the common notion that the TEI "is" or "expresses" a general text ontology. In several aspects, this statement overlaps with the well-known statement made by Sperberg-McQueen that the "markup reflects a theory of text" (1991). There is no general agreement that the TEI can actually be reduced to one ontology. We concede that this is true for the TEI as a whole, and that in principle there is no way to identify the "one true ontology" of a text. But in practice we are confident that under normal conditions all encoders share the same single interpretation for many common markup constructs, because of cultural and dispositional factors. On the other hand, as Guarino (2009) notes, a formal ontology always includes epistemological aspects as well as pure ontological assertions and therefore it is by nature prospective and amendable.

22 The underlying ontology of TEI is partially formalized in the set of TEI XML schema constraints, and for the most part conveyed by the natural-language prose of the Guidelines. One possible and rapid approach to producing a formal ontology for TEI would be the adaptation of the aforementioned constraints into a set of ontological statements in OWL. The conversion of XML schemas into OWL ontologies has been discussed in many papers in the last ten years, and many theoretical and

computational solutions have been proposed (e.g., Bohring and Auer 2005; Ghawi and Cullot 2009). We cannot address the technical details of these solutions here. Most of them are based on the mapping of W3C Schema primitives into OWL primitives.

23 The adoption of this kind of strategy could make the construction of a formal high-level TEI ontology a partially automated process. However, it is evident that the most relevant semantic properties of the markup, which cannot be expressed by common schema languages (and ODD), would not be modeled by this semi-mechanical translation.

24 The TEI itself has made an attempt to define explicitly its underlying conceptual model with the notions of abstract models and of element classes:

> 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....
>
> (TEI Consortium 2015, 23.4.3)

25 And specifically, regarding the design of the element classes and the naming conventions adopted to convey their typology, the Guidelines state:

> In fact, the nature of a given class of elements can be considered along two dimensions: as noted, it defines a set of places where the class members are permitted within the document hierarchy; it also implies a semantic grouping of some kind. For example, the very large class of elements which can appear within a paragraph comprises a number of other classes, all of which have the same structural property, but which differ in their field of application. Some are related to highlighting, while others relate to names or places, and so on.... If a model class has a name containing 'part,' ... then it is primarily defined in terms of its structural location. ... If, however, a model class has a name containing 'like,' ... the implication is that its members all have some additional semantic property in common....
>
> (TEI Consortium 2014, 1.3.2)

26 It is therefore possible to identify a set of strictly structural constraints, expressed in ODD or RelaxNG patterns, and a set of informal or semi-formal semantic/taxonomic directives. A proper ontological modeling should express both the abstract characterization of TEI elements' semantics and the ontological definition of their structural role. In addition, the ontology should define a precise semantics of the elements that are characterized unambiguously in the official TEI documentation (e.g., the element <p>), while it should relax the semantical constraints if the elements in consideration can be used with different semantic connotations depending on the context (e.g., the element <seg>). Finally, it should be possible to extend the ontology, reuse it, and define alternative characterizations of element semantics without compromising the consistency of the ontology itself.

27 In accordance with these overall principles we have decided to implement a complex ontological architecture, using some pre-existing meta-ontology frameworks to express the meaning of the TEI element set. The specification of markup semantics for the various TEI elements is done by means of LA-EARMARK classes and properties. The Extremely Annotational RDF Markup (EARMARK) is at the same time a markup metalanguage, which 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, providing a finer specification of their properties. LA-EARMARK is an extension of EARMARK with the Linguistic Act module of the Linguistic Meta-Model ontology, which allows the expression and assessment of facts, constraints, and rules about the markup structure as well as about the inherent semantics of the markup elements themselves (Peroni and Vitali 2009; Peroni, Gangemi, and Vitali 2011).

28 The general EARMARK class to define any markup element is `earmark:Element`. For instance, the <abbr> element is defined as follows (in Manchester Syntax):

```
Prefix earmark: <http://www.essepuntato.it/2008/12/earmark#>

Prefix tei: <http://www.tei-c.org/ns/1.0/>

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

29 We need to create appropriate restrictions of classes in order to identify and characterize possible subsets of elements described by the schema. This is achieved by adopting the classes and properties of the Collections Ontology (CO), which in OWL defines unordered and ordered collections (Ciccarese and Peroni 2014). For instance, the class of all the `<tei:p>` elements that occur inside `<tei:text>` and not inside `<tei:teiHeader>`:

```
<tei:teiHeader>:
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:pInText
 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 "body" and
     earmark:hasNamespace "http://www.tei-c.org/ns/1.0")
```

30 This class is disjoint from `tei:pInTeiHeader` class.

31 LA-EARMARK allows us to link particular classes of elements with the actual semantics they express. From our point of view, as we said, there are at least two semantic levels that we are interested in defining explicitly:

1. One level concerns the structural behavior of markup (e.g., the fact that an element is a block rather than an inline, a container rather than a field). This can be described by means of existing models such as the Pattern Ontology (Di Iorio et al. 2014).
2. The other relates to the intended semantics of an element (e.g., the fact that an element is a paragraph rather than a section, a personal name reference rather than a geographical reference). This can be described by a specialized TEI Semantics Ontology or by a combination (and/or an extension) of already existing ontologies.

32 The TEI Semantics Ontology component is the core layer in our architecture, and is actually under development. Its class hierarchy is defined on the base of the TEI abstract model classes, but a lot of work and refactoring is needed to achieve sound ontological structure.

33 The lower-level classes are the concepts expressed by TEI markup construct. It is worth noting, though, that there is not a one-to-one relation between elements and lower-level semantic classes, since we can identify at least three different markup "crystals" that can have a different ontological meaning:

1. one XML element on its own: e.g., `<abbr>`;
2. an XML element/attribute couple or a compound of so-called "Janus elements": e.g., `<corr resp=>` or `<choice><sic><corr>`
3. one XML element in a given context: e.g., `<p>` in `<text>` vs. `<p>` in `<teiHeader>`.

34 The link between the class describing kinds of elements and their related semantic characterization is expressed by means of the property `semiotics:expresses` provided by the Linguistic Act Ontology included in LA-EARMARK. The associations of semantics to markup elements can also be contextualized according to a particular agent's point of view, in order to provide provenance data pointing to the entity who was responsible for such specification. This is possible by means of properties that allow one to assign agency and responsibility to all these markup-to-semantics relations, as proper linguistic acts performed by someone.

6. Specific Intensional Semantics of TEI Markup

35 The specific intensional semantics is the level where local or even idiolectal semantic specifications of the markup can be defined. At this level we can find the specific structures of meaning that a markup term has for a specific user or community. For example, think of a specialization in the use of abstract container elements such as `<div>`, `<ab>`, and `<seg>` by way of the `@type` attribute that defines a restriction of the semantics compared to the one defined at the general ontology level.

36 These ontology specializations can be expressed as:

1. restrictions on properties and classes that extend the general ontology in OWL;
2. a set of inference rules expressed through Rule Language (like SRWL), which extend the general OWL ontology;
3. local contextual restrictions expressed as LA-EARMARK properties, as seen above.

37 How can a user possibly declare these local semantic extensions? The obvious place to declare these user- or project-specific ontological specifications for the TEI is the ODD document. One possible and straightforward solution is to adopt the `<constraint>` element already provided by the ODD language. Admittedly this choice would represent a sort of tag abuse, since this element is thought to bear restrictions on XML elements or attributes that can be validated by a parsing process, not semantic constraints. So the advisable solution will be to introduce a dedicated element in the ODD personalization that allows a user to declare the relevant ontological constraints in OWL. Those formulas could then be added to the general ontology during ODD processing.

7. Ontology Mapping, Matching, and Merging

38 Alignment of ontologies is a necessary step in the project. The creation of ontologies is the strongest example of knowledge conceptualization. This activity likewise requires a network dialogue in order to deal with interoperability issues and try to solve the problem of semantic heterogeneity in naming of classes and predicates. The TEI semantic enhancement proposed here has to compare the TEI general ontology with some selected, existent, and pertinent models already used in other domains and communities.

39 Ontological modeling is a complex and iterative process, and it requires a deep understanding of the role of classes and predicates, and how they are used in different domains. So creating correspondences is the most difficult issue: the ontology is a subjective representation of reality and therefore it is natural that differences between points of view occur. Any attempt to classify the mismatches between ontologies must recognize possible issues at both the linguistic and the semantic levels. The most common issue derives from the different use of the same concepts in different domains. But the use of distinct names for the same concept also has to be considered. Problems could also emerge from other modeling features: the different scope of some classes or predicates; the different focus in classes and predicates definition; the possibility of using different constructs; different modeling conventions; and different levels of granularity.

40 Mapping, matching, and merging are the common methods used in order to address issues related to semantic integration (Noy 2004; Choi, Song, and Han 2006). These processes require on the one hand the comparison between models and on the other the possible integration of single elements of different vocabularies into one single model.

41 First of all it is necessary to understand that ontology alignment requires the creation of binary relationships between the vocabularies of two ontologies. This problem could be solved with an ontology-to-ontology mapping, but it could also make use of a common reference ontology (upper-level ontology) onto which the different models are mapped. A study of the existing tools and methods (Kalfoglou and Schorlemmer 2003) will be necessary in order to face the problem (from machine learning, concept lattices, and formal theories to heuristics, database schemas, and linguistics).

42 If the mapping requires the selection of a method (Yang, Steele, and Lo 2007), choosing which ontologies are potentially in conflict or in agreement with our semantic TEI proposal is not merely a secondary issue to face. A particularly relevant aspect of the conceptual model definition process will then be to check the pertinent existing ontologies in order to ensure maximum portability in all contexts. The TEI Ontologies SIG has already done some relevant work in this area, especially the work of Ore and Eide with CIDOC-CRM (2009). In particular, the last paper by Eide (2015) describes some mapping solutions. However, besides the most common ontologies devoted, for example, to cultural heritage (CIDOC-CRM), archives (EAD and EAC-CPF), and metadata (DC, DC terms, and EDM), other ontologies, developed in other domains, provide new forms of conceptualization. For example, an ontology such as FABIO or CiTO could be an interesting application case (Peroni and Shotton 2012). FABIO is based on the FRBR approach to the document as a complex entity. The stratification of analysis levels enriches the description of cultural entities. CiTO is useful in order to manage the citation process, towards the definition of multiple relationships and cross-relationships between the data related to attribution statements. In particular, CiTO makes it possible to define the relationships between an interpretation and the source supporting that interpretation. But other common ontologies and models, such as FOAF, SKOS, Bibframe, or Schema.org, are a necessary reference for attempting semantic interoperability. But the analysis of the state of the art has to be combined with another fundamental step: integrating into the ontology information already conceptualized in other domains. This means that the TEI model will be refined by acquiring the vocabulary used in the domains mentioned above: already shared and public-domain classes and properties could be encapsulated in the TEI conceptual model. At the same time the TEI model could contribute to populating other models: specific classes and properties could be reused by other communities as a result of the TEI semantic extension.

In both cases, the process must be done with the awareness of the objective impossibility of a complete overlapping of models. By adopting this hybrid approach we approximate semantic interoperability: interconnection, integration, and semantic dialogue are vital ingredients in the process. We could even try to approach the deepest interchange possible between conceptual models (Bauman 2011). The greater the sharing of principles inside the community of humanities, the better the process will work.

8. LOD Conversion and Dissemination

43 In order to test the outlined model the TEI ontology must be verified in a linked open data dimension (i.e., tested in Linked Open Vocabularies)[2] and, likewise, TEI documents must be analyzed in order to attempt a semantic enrichment. The TEI ontology proposal is thus helpful in exposing both the TEI semantic model and individual TEI documents as linked data. TEI conversion into LOD consists of a sequence of steps that covers two levels of analysis: the work on the TEI model at the schema level, and the work on TEI XML files in order to produce open and linked data using the modeled ontology.

44 The first level is related to the above description of the TEI model formalization for a first macro-modelization. This level also encompasses mapping of the TEI ontological model onto selected ontologies in order to guarantee interchange and expressivity of the model from a reuse perspective. The second level is instead related to our definition of "extensional semantics": adding identification mechanisms to in-line markup, eventually pointing to stand-off description, and refining the TEI markup in order to be LOD-compliant. This will mean going beyond data silos, opening towards the "web of data" in an integrated environment. In detail:

- the user experience will improve through a machine-interpretable semantics (TEI will be more "understandable" by the machine);
- the recall, i.e., the fraction of the documents that are relevant to the query that are successfully retrieved, will be more precise (the specification of the TEI predicates' role will refine the results);

- disambiguation, by using a conceptual model, will be more precise, especially in case of homography of classes and predicates (TEI classes and predicates will be mapped onto the same concepts in other ontologies);
- mashup practices (smart applications) will improve the knowledge base through new connections, even between heterogeneous domains (the TEI model will dialogue with other communities);
- inferences will be available, giving the possibility of new knowledge discovery.

45 As Berners-Lee (2006) states, many research and evaluation projects in the domain of Semantic Web technologies have produced ontologies. From the LOD perspective, i.e., a fundamental step in the direction of the Semantic Web realization, the ontology support would provide benefits in semantic expressivity, data interchange, and machine—but also users'—real exploitation (Heath and Bizer 2011). In the "LOD 5 Stars"[3] idea, the use of ontology is not compulsory. In our model the ontological support will be the key to enriching data in order to augment the semantic of TEI documents, by transforming the information managed by the schema in knowledge defined by the ontology. The idea of a strong interconnection between ontologies and the possibility of creating relationships between vocabularies, from a reuse-oriented perspective, will be fundamental to making the TEI semantic model compliant with the cloud: "link your data to other data to provide context" (the five-star level in the "5-star open data plan");[4] that is, the context is determined by the relationships. Siloed data are not able to express their semantic power. Information is the result of mutual connections between data. The knowledge graph approach, coined by Google in 2012 (Singhal 2012) is demonstrating that one of the natural consequences of the migration from document to data, inaugurated by the Semantic Web, enriches the user experience.

46 In order to finalize the model from an LOD cloud perspective—as regards the collection of TEI-based documents—various methods will have to be explored, beginning with the creation of an RDF triple store by converting some pertinent elements of the refined TEI XML files into RDF through XSLT. Experiments in converting XML files into RDF have already been undertaken: "a transformation to RDF has to create the URIs of its resources and connect them through the RDF triple structure consisting of subject, predicate and object" (Breitling 2009).

47 The problem is double: how to define dereferenceable URIs for elements (concepts and/or real web documents) and how to work on a specific XML dialect, that is, the TEI vocabulary?

48 These questions are not trivial because they require reflection on the TEI markup model. The topic is difficult and we are now trying to address this complexity. A first approach we are attempting is the following. In general we can assert that: TEI elements are `rdf:description` about a node id (e.g., through an `@ref`) that we could manage in XSLT for transforming the `@ref` value into a URI. This approach yields:

```
SUBJECT = rdf:description about a TEI element (the @ref value)
PREDICATE = an attribute of the element in the subject (for managing cross-
references) or, simplest, the child element
OBJECT = literal (the content of the element)
```

49 An example of an XML TEI markup in a document:

```
<person ref="#persona01">
 <persname>Vespasiano</persname>
</person>
```

That is: an entity (person) with a value (a fragment) referring to a `@xml:id` (`"persona01"`) to be converted in dereferenceable URI (e.g., `http://www.person.it/about#persona01`) through XSLT, a predicate corresponding to the child element (`persname`) and a literal as the object (Vespasiano). Another fundamental issue is the identification of pertinent authorities for the data matching (e.g., VIAF, Geonames, Worldcat, SNAP, or DBpedia). In order for the datasets to be able to exchange information in a network dimension, sharing of authorities is a crucial point. Mechanisms such as "see-also" or even "same-as" could manage relationships between named entities.

50 Likewise, the need for a really "linked" environment requires the discovery of links in the cloud by using semi-automatic methods of entity recognition (NER).

51 Finally, populating the LOD cloud Data Catalog and the Datahub[5] with both the semantic vocabulary and the converted datasets will improve visibility and interchange.

9. Conclusions and Perspectives

52 In our opinion, the possibility of providing a TEI-formalized semantics using Semantic Web standard technology constitutes a good opportunity to achieve these objectives:

1. Strictly set out the general semantics of the markup language in order to facilitate 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 include any XML TEI repository in the open linked data (LOD) environment (Isaac et al. 2011). TEI could be redefined as a "linked open vocabulary" able to exchange information with other LOV datasets at either vocabulary or element level.
3. Ease the conversion of existent TEI-based digital libraries into open and linked datasets able to be shared in the LOD cloud.
4. Provide users with advanced formal tools to define their interpretations of the texts to which they apply the markup, and allow innovative computational processing based on semantics, using tools such as reasoners and semantic query engines.

53 However, the cost and the practical complexity of such an extension are significant, and several theoretical problems, format choices, and implementation details remain to be defined.

BIBLIOGRAPHY

Bauman, Syd. 2011. "Interchange vs. Interoperability." In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, vol. 7. doi:10.4242/BalisageVol7.Bauman01.

Berners-Lee, Tim. 2006. "Linked Data." Personal Note, Last modified June 18, 2009. https://www.w3.org/DesignIssues/LinkedData.html.

Bohring, Hannes, and Sören Auer. 2005. "Mapping XML to OWL Ontologies." In *Marktplatz Internet: Von e-Learning bis e-Payment. 13. Leipziger Informatik-Tage (LIT 2005)* Lecture Notes in Informatics P-72, edited by Klaus P. Jantke, Klaus-Peter Fähnrich, and Wolfgang S. Wittig, 147–56. http://subs.emis.de/LNI/Proceedings/Proceedings72/GI-Proceedings.72-11.pdf.

Breitling, Frank. 2009. "A Standard Transformation from XML to RDF via XSLT." *Astronomische Nachrichten* 330(7): 755–60. http://arxiv.org/abs/0906.2291. doi:10.1002/asna.200811233.

Burnard, Lou. 2013. "Resolving the Durand Conundrum." *Journal of the Text Encoding Initiative* 6. http://jtei.revues.org/842. doi:10.4000/jtei.842.

Burnard, Lou, and C. M. Sperberg-McQueen. 2012. "TEI Lite: Encoding for Interchange: An Introduction to the TEI." Final revised edition for TEI P5, August. http://www.tei-c.org/Guidelines/Customization/Lite/index.xml.

Carnap, Rudolf. 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

Choi, Namyoun, Il-Yeol Song, and Hyoil Han. 2006. "A Survey on Ontology Mapping." *ACM SIGMOD Record* 35(3): 34–41. doi:10.1145/1168092.1168097.

Ciccarese, Paolo, and Silvio Peroni. 2014. "The Collections Ontology: Creating and Handling Collections in OWL 2 DL Frameworks." *Semantic Web* 5(6): 515–29. doi:10.3233/SW-130121.

Ciula, Arianna, Paul Spence, and José Miguel Vieira. 2008. "Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project." *Literary and Linguistic Computing* 23(3): 311–25. doi:10.1093/llc/fqn018.

Cover, Robin. 1998. "XML and Semantic Transparency." Technology report. In *Cover Pages*. Last modified November 24, 1998. http://xml.coverpages.org/xmlAndSemantics.html.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1997. "What Is Text, Really?." *ACM SIGDOC Asterisk: The Journal of Computer Documentation* 21(3): 1–24. doi:10.1145/264842.264843.

Di Iorio, Angelo, Silvio Peroni, Francesco Poggi, and Fabio Vitali. 2014. "Dealing with Structural Patterns of XML Documents." *Journal of the Association for Information Science and Technology* 65(9): 1884–1900. doi:10.1002/asi.23088.

Eide, Øyvind. 2015. "Ontologies, Data Modeling, and TEI." *Journal of the Text Encoding Initiative* 8. http://jtei.revues.org/1191. doi:10.4000/jtei.1191.

Ghawi, Raji, and Nadine Cullot. 2009. "Building Ontologies from XML Data Sources." *20th International Workshop on Database and Expert Systems Applications*, 480–84. doi:10.1109/DEXA.2009.68.

Gruber, Tom. 2009. "Ontology." In *Encyclopedia of Database Systems*, edited by Ling Liu and M. Tamer Özsu. New York: Springer.

Guarino, Nicola. 2009. "The Ontological Level: Revisiting 30 Years of Knowledge Representation." In *Conceptual Modeling: Foundations and Applications. Essays in Honor of John Mylopoulos*, edited by Alexander T. Borgida, Vinay K. Chaudhri, Paolo Giorgini, and Eric S. Yu. Lecture Notes in Computer Science 5600, 52–67. Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-02463-4_4.

Heath, Tom, and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, lecture 1. San Rafael, CA: Morgan & Claypool. doi:10.2200/S00334ED1V01Y201102WBE001.

Huitfeldt, Claus, and C. M. Sperberg-McQueen. 2001. "TexMECS: An Experimental Markup Meta-language for Complex Documents." Working paper of the project Markup Languages for Complex Documents (MLCD). University of Bergen. Revised October 5, 2003. http://mlcd.blackmesatech.com/mlcd/2003/Papers/texmecs.html.

Isaac, Antoine, William Waites, Jeff Young, and Marcia Zeng. 2011. "Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets." W3C Incubator Group Report (25 October). http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/.

Kalfoglou, Yannis, and Marco Schorlemmer. 2003. "Ontology Mapping: The State of the Art." *The Knowledge Engineering Review* 18(1): 1–31. http://eprints.soton.ac.uk/260519/1/ker02-ontomap.pdf. doi:10.1017/S0269888903000651.

Knuth, Donald E. 1984. "Literate Programming." *The Computer Journal* 27(2), 97–111. doi:10.1093/comjnl/27.2.97.

Noy, Natalya F. 2004. "Semantic Integration: A Survey of Ontology-based Approaches." *ACM Sigmod Record* 33(4), 65–70. doi:10.1145/1041410.1041421.

Ore, Christian-Emil, and Øyvind Eide. 2009. "TEI and Cultural Heritage Ontologies: Exchange of Information?." *Literary and Linguistic Computing* 24(2): 161–72. doi:10.1093/llc/fqp010.

Patel-Schneider, Peter F., and Ian Horrocks. 2006. "Position Paper: A Comparison of Two Modelling Paradigms in the Semantic Web." In *Proceedings of the 15th International World Wide Web Conference*, 3–12. New York: ACM. doi:10.1145/1135777.1135784.

Peroni, Silvio, Aldo Gangemi, and Fabio Vitali. 2011. "Dealing with Markup Semantics." In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics 2011)*, edited by Ciara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie Lindstaedt, and Tassilo Pellegrini, 111–18. New York: ACM. doi:10.1145/2063518.2063533.

Peroni, Silvio, and David Shotton. 2012. "FaBiO and CiTO: Ontologies for Describing Bibliographic Resources and Citations." *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 17:33–43. doi:10.1016/j.websem.2012.08.001.

Peroni, Silvio, and Fabio Vitali. 2009. "Annotations with EARMARK for Arbitrary, Overlapping and Out-of-order Markup." In *Proceedings of the 9th ACM Symposium on Document Engineering (DocEng 2009)*, 171–80. New York: ACM. doi:10.1145/1600193.1600232.

Renear, Allen, David Dubin, C. M. Sperberg-McQueen, and Claus Huitfeldt. 2002. "Towards a Semantics for XML Markup." In *DocEng'02: Proceedings of the 2002 ACM Symposium on Document Engineering*, edited by Ethan Munson, Richard Furuta, and Jonathan I. Maletic, 119–26. New York: ACM Press. doi:10.1145/585058.585081.

Schmidt, Desmond. 2010. “The Inadequacy of Embedded Markup for Cultural Heritage Texts.” *Literary and Linguistic Computing* 25(3): 337–56. doi:10.1093/llc/fqq007.

Singhal, Amit (May 16, 2012). “Introducing the Knowledge Graph: Things, Not Strings.” *Official Blog* (of Google). https://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html.

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): 34–46. doi:10.1093/llc/6.1.34

Sperberg-McQueen, C. M., and Claus Huitfeldt. 2004. “GODDAG: A Data Structure for Overlapping Hierarchies.” In *Digital Documents: Systems and Principles* [Revised papers from the 8th International Conference on Digital Documents and Electronic Publishing (DDEP 2000) and the 5th International Workshop on the Principles of Digital Document Processing (PODDP 2000)], edited by Peter King and Ethan V. Munson, 139–160. Lecture Notes in Computer Science 2023. Berlin: Springer. doi:10.1007/978-3-540-39916-2_12.

———. 2008. “Markup Discontinued: Discontinuity in TexMecs, Goddag structures, and Rabbit/Duck Grammars.” In *Proceedings of Balisage: The Markup Conference 2008*. Balisage Series on Markup Technologies, vol. 1. doi:10.4242/BalisageVol1.Sperberg-McQueen01.

Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen Renear. 2000. “Meaning and Interpretation of Markup.” *Markup Languages: Theory & Practice* 2(3): 215–34. http://cmsmcq.com/2000/mim.html.

Sperberg-McQueen, C. M., Yves Marcoux, and Claus Huitfeldt. 2010. “Two Representations of the Semantics of TEI Lite.” Paper presented at Digital Humanities 2010, London, July 7–10. Abstract available at http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-663.html.

———. 2014. “Transcriptional Implicature: A Contribution to Markup Semantics.” Paper presented at Digital Humanities 2014, Lausanne, Switzerland, July 7–12. Abstract available at http://dharchive.org/paper/DH2014/Paper-61.xml.

TEI Consortium. 2015. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.8.0. Last updated April 6. http://www.tei-c.org/Vault/P5/2.8.0/doc/tei-p5-doc/en/html/.

Tummarello, Giovanni, Christian Morbidoni, Fabio Kepler, Francesco Piazza, and Paolo Puliti. 2006. “A Novel Textual Encoding Paradigm Based on Semantic Web Tools and Semantics.” In *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation*, 247–52. Paris: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2006/pdf/225_pdf.pdf.

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

Turska, Magdalena, James Cummings, and Sebastian Rahtz. 2016. “Challenging the Myth of Presentation in Digital Editions.” *Journal of the Text Encoding Initiative* 9. https://jtei.revues.org/1453. doi:10.4000/jtei.1453.

Yang, Kai, Robert Steele, and Amanda Lo. 2007. "An Ontology for XML Schema Ontology Mapping Representation." In *Proceedings of the 9th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2007)*, edited by Gabriele Kotsis, David Taniar, Eric Pardede, and Ismail Khalil Ibrahim. Vienna: Österreichische Computer Gesellschaft.

NOTES

1 See Knuth 1984.

2 Pierre-Yves Vandenbussche and Bernard Vatant, Linked Open Vocabularies, http://lov.okfn.org/dataset/lov/.

3 http://5stardata.info/

4 http://5stardata.info/

5 Open Knowledge Foundation, Datahub, https://datahub.io/.

ABSTRACT

The debate on the semantic role of markup languages has been quite lively and the TEI community has played an active part in it. It is commonly acknowledged that markup conveys semantic information. However, XML is a poor language for semantic data modeling. Several proposals have previously been drawn up in the past to provide XML with formalized and computable semantics. In our opinion, the formalisms offered by the Semantic Web paradigm are mature enough to build a workable semantic extension of the TEI.

Our model distinguishes three semantic layers in the TEI: one general and shared intensional semantic layer; one idiolectal specialized layer; and finally an extensional semantics. Our proposal is directed toward the first two layers. We propose to build such semantic layers by adopting a set of OWL formal ontologies.

Furnishing the TEI with a semantics based on a formal ontology could have interesting outcomes: facilitating the management of and research using document collections in open and multi-standard contexts; aiding interoperability with other relevant standards in the digital cultural heritage context; and providing users with advanced formal tools to semantically define their interpretations of the texts and enable innovative computational processing. In order to allow a semantic interoperability between standards, the TEI ontology has to be aligned to other models; likewise mapping and merging procedures have to be evaluated. Finally, the idea of migrating XML/TEI documents following this semantic model into a linked open data dimension requires that we face important issues in order to facilitate the data interchange in the cloud.

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.

INDEX

Keywords: ontology, linked data, XML semantics, modeling, Semantic Web, text encoding theory

AUTHORS

FABIO CIOTTI

Fabio Ciotti is a tenured assistant professor at the University of Roma Tor Vergata, where he teaches digital literary studies and theory of literature. He is president of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD, the Italian Digital Humanities association), a former elected member in the TEI Consortium Technical Council, and a current member of the EADH (European Association of Digital Humanities) Executive Board. His scientific and research work covers various aspects and themes of Digital Humanities and literary studies, from both the theoretical and the practical point of view. Fabio Ciotti has been a scientific consultant for text encoding and technological infrastructures in several digital libraries and archives projects, most notably Biblioteca Italiana (Italian literary tradition, http://www.bibliotecaitaliana.it) and DigilibLT (Late Latin tradition, http://www.digiliblt.unipmn.it/).

FRANCESCA TOMASI

Francesca Tomasi is assistant professor in Library and Archival Science in the School of Italian Studies at the University of Bologna, where she teaches courses on archives and computer, multimedia production, and Digital Humanities. Her research is focused in particular on digital editions, digital libraries, and digital historical archives. She is President of the Library of the School of Humanities in the University of Bologna. She is the editor of a scholarly digital edition (Vespasiano da Bisticci, Letters: http://vespasianodabisticciletters.unibo.it, 2013). Among her last publications she counts the monograph *The Digital Humanist. A Critical Inquiry*, with D. Fiormonte and T. Numerico, Punctum books, Brooklyn, NY 2015 and several articles dedicated to data modeling and extraction in the context of Cultural Heritage objects.