Text Encoding Initiative semantic modeling. A conceptual workflow proposal

Fabio Ciotti[1], Marilena Daquino[2], and Francesca Tomasi[2]

[1] Department of Humanities, University of Roma, Italy
fabio.ciotti@uniroma2.it
[2] Department of Classical Philology and Italian Studies, University of Bologna, Italy
{marilena.daquino2, francesca.tomasi}@unibo.it

Abstract. In this paper we present a proposal for the XML TEI semantic enhancement, through an ontological modelization based on a three level approach: an ontological generalization of the TEI schema; an intensional semantics of TEI elements; an extensional semantics of the markup content. A possible TEI semantic enhancement will be the result of these three levels dialogue and combination. We conclude with the ontology mapping issue and a Linked Open Data suggestion for digital libraries based on XML TEI semantically enriched model.

Keywords: ontology, TEI, XML, interoperability, LOD, digital libraries.

1 Introduction

Digital libraries are moving towards a radical identity redefinition. Semantic Web technologies are, in particular, contributing to increase the expressivity of the digital library as an unexplored reservoir of raw data, on which keyword as ontologies, RDF, OWL, metadata and controlled vocabularies play a crucial role. “Yet semantic technologies offer a new level of flexibility, interoperability, and relationships for digital repositories” [1].

Typically XML is the meta-language used in order to populate libraries with documents aiming at ensuring the maximum syntactic interchange, even between technological heterogeneous repositories. In the ‘humanistic community’ XML is commonly used in combination with TEI (Text Encoding Initiative)[1], the ‘controlled vocabulary’ for declaring humanistic-domain-based interpretations. TEI is a flexible and customizable schema that represents a shared approach in the community, demonstrated by the fact that the most of existent digital libraries of national textual traditions are based on this standard[2].

The TEI project originated in 1987, following a conference organized by the ACH (Association for Computers and the Humanities). During the conference, the need to define a standard for the digitization of text inspired the ACH, along with the Association for Computational Linguistics (ALC) and the Association for Literary and Linguistic Computing (ALLC) to establish the first guidelines for the encoding and

[1] http://www.tei-c.org/

[2] http://www.tei-c.org/Activities/Projects/

exchange of texts in electronic format. In 1999, the TEI Consortium was founded with the aim of maintaining, developing and promoting the *Guidelines*[3]. Literary texts, whether in prose, verse, or drama, find in the TEI *Guidelines* a ready set of elements for the description of all necessary phenomena suitable for interpretation: from the definition of the elements of a document's logical structure to the specification of people's names, places and dates, from the description of a manuscript to the marking of phenomena peculiar to an edition (such as the apparatus of a critical edition), from linguistic analysis to rhetorical or narrative structures. TEI is also a project in continual evolution. From version P1 in July 1990 (an initial draft, 89 tags) it has evolved to P5 (2007-2014, version 2.7.0, about 550 tags).

The XML TEI approach lets two questions emerge, in a huge debate. From one hand XML is a topic largely addressed as a reflection on a meta-language that could be exploited not only for data interchange (the data community approach) but also for text representation (the document community approach). This issue reveals a shared widespread belief among researchers: XML is semantically poor and it needs strategies for data model improving.

From the other hand TEI, that is a real standard in the humanistic domain, could be refined in a semantic perspective starting from the translation of the TEI framework (i.e. schema, community personalization, real documents) in an ontology[4]. This process is not a simple task because the conversion of the schema in OWL is just one of the levels of conceptualization. In our proposal the semantic enhancement of the existent TEI framework requires a three levels approach.

The first level is the ontological generalization of the schema. This task could not be treated as a complete automatic process. It has to be the result of a deep analysis of the original model in a class/properties dimension. But it's not a simple 1:1 relation (tei:element=tei:class; tei:attribute=tei:property) it's even a conceptual analysis of scope and content of entities.

The second level regards the domain specific approach determined by a specific community trend. In particular the information managed by some attributes values are fundamental in order to define the semantic of a model. Possible specific local additions to the general TEI schema have to be considered in order to create an ontology that could be able to describe all the XML TEI based documents.

The third level regards the real markup, i.e. the analysis of a real domain described by real documents, in order to enrich a theoretical model. This means that the ontology will require a study of marked up documents in order to manage individuals and understand the potential relations that instances could manage with other external, and potentially different at the level of semantic declaration, information.

All the describe process has to be thought in a re-conceptualization of the text as a multi-levels entity. Following the FRBR approach we have to consider that each class and each property of the ontology reflect one of the levels of an analysed entity: the work, the expression, the manifestation, the item. The aim of each class and each property has to be considered in potentially each of these levels: the function of classes and properties naturally changes in relation to the level that class and properties want to refer to. We have to consider that a TEI document contains information regarding not only the text but a complex granularity of contextual information that have to be

[3] http://www.tei-c.org/Guidelines/

[4] http://www.tei-c.org/Activities/SIG/Ontologies/index.xml

likewise modeled.

The ontological approach, starting from an applied schema, let even to reflect on a deep issue: the passage from a strictly hierarchical model (the XML TEI markup) to a network model (the ontology). The data structure has a fundamental role in semantic definition of a modelled domain. The tree model has a representational potential that could be re-thought according to the graph expressivity, compelling a semantic re-definition of text features. In the conversion of the schema into ontology we have to keep in mind that elements and attributes organized as functional modules or classes in the schema (e.g. the class of attributes for the dates) could be reorganized on real entities or real relationships (e.g. date of birth, date of an event, date mentioned in a text).

This described method will help XML TEI based digital libraries to move from a static, flat dimension towards an inter-related, interoperable and interchangeable environment. Semantic Web technologies will provide tools and methods for migrating XML TEI files in the Linked Open Data cloud. This migration will involve two other issues: the mapping of the realized ontology on classes and properties of other pertinent ontologies; the compatibility of the semantic model proposal with the LOD specification. A detailed study of TEI real cases (i.e. TEI real documents) will be also strictly necessary in order to provide the ontology refinement towards a LOD compatibility.

The paper is then organized as follow: section 2 is devoted to describe the debate regarding the XML semantic enriching; section 3 focuses on the specific issue of ontology creation starting from the TEI framework; section 4 is devoted to ontology mapping and the LOD approach; section 5 opens to next research perspectives, in order to realize a real ontology based on these conceptual reflections.

2 The XML semantics debate

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

However, this markup and, in particular, the XML markup semantic role, clashes with the fact that, as [2] already observed some time 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". XML is a powerful formalism to define the syntactic markup aspects, and, through its data model, to model some limited structural features of information objects to that it is applied. Still, it owes its semantic value almost entirely to human interpretation. Any markup restriction or semantic role accordingly needs to be expressed in natural language as instructions for human users. This is the case for ODD formalism [3], which TEI developed with the aim to combine in a single meta-XML document the custom definition of its XML schema and all its relevant documentation.

Several proposals were drawn up to provide XML with formalized and computable semantics. The work [4, 5] constitutes the first, explicit contribution in this direction.

The 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 were focused on the proposal of a RDF based model for text encoding [6]; or by exploring the potentiality of LMM, an OWL vocabulary that represents some core semiotic notions, in order to provide a better understanding of the semantics of markup [7]; or also with the idea of "transcriptional implicature" [8, 9, 10].

In these last 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 [9].

Nonetheless the same authoritative authors of this last paper observed that if the proposals for formal semantic approaches to markup have been very scarce, their practical application are even less.

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;
- excessive "revolutionary" scope of some proposals.

In this paper, we propose a Semantic Web extension of the TEI infrastructure in order to formalize some of the semantic levels of the markup constructs it provides.

The rationales of our proposal are:

- to be based on well-established Semantic Web formalism and technologies;
- to be an extension not a replacement of current languages and practices;
- to provide a viable solution for some practical concerns that are relevant in the actual digital ecosystem in which TEI and XML live, especially interoperability and linked data.

3 TEI and Semantic Web Technologies

During the time elapsed between the first approaches to the semantics of markup language and situation today the development and the relevant spread of the Semantic Web paradigm and, more recently, Linked Data occurred. This process has made available a number of syntactically rigorous and semantically well-founded languages and data models, such as RDF / RDFS, SPARQL and OWL 2, as well as systems and software components, aimed at the semantic data processing (storage, query, and inference). As stated [11], many research and evaluation projects in the Semantic Web technologies domain produced ontologies. From the LOD (Linked Open Data) perspective, i.e. a fundamental step in the direction of the Semantic Web realization, the ontology support would provide benefits in semantic expressivity power, data interchange and machine - but also users - intelligent consumption.

Starting from this context, we propose to develop an ontological approach for TEI, that will give a formal definition to the implicit concepts underlying XML TEI text encoding.

To this purpose, it is appropriate to distinguish between three different iteratives semantic levels expressed by the markup and its content:

1. Generalization on TEI Schema (in order to define a broad ontological description of entities involved in text encoding). A top-down approach: from the schema to the ontology (see 3.1).
2. Intensional TEI markup semantics (defined by a particular user of community of practice). A bottom-up approach: from the community to the ontology revision (see 3.2).
3. Extensional semantics of the markup content. A bottom-up approach: from real documents to ontology for refinement (see 3.3).

The rationales for this proposal are both theoretical and operational. In the DH community a great relevance has been given to the notion of model and modelling, so that we often can find assertion like “text encoding is a form of modeling”. The very problem with the model/modeling notions is that they are umbrella terms, relating to an ample and diverse sort of conceptual objects and practices. In general, we can summarize the roles assigned to modeling in scientific activity in three areas:

- representation/communication: models ensure that a community of practice shares the fundamental concepts of a domain;
- explanation/prediction: models relates facts and concepts providing explanations and possibly predictions of the behavior of a system;
- multiple views/perspectives mediation: models mediate between the different perspectives that can arise within a single community of practice and between different but proximal communities of practice.

Ontological modeling formalizes the common sense concept of model giving it a precise logical semantics a definite functional role in each of these areas. Creating formal models based on explicit conceptualization and logical foundation grants that all the discourses are firmly grounded to a common “setting” of the domain.

Formal ontologies license the application of computational inferences and reasoning to express explanation and make predictions. And finally Semantic Web modeling provides methods to compare and eventually merge different ontologies and, being based on the Open World Assumption, ensures the functionality of the model even if it is incomplete or conceived as a work in progress.

In order to formalize a complex taxonomy and a hierarchy for relations, following conceptual steps later described, more than one methodology have to be used iteratively in ontology engineering.

- Firstly when generalizing on entities involved in text encoding domain, a hybrid approach, which also take in account well-known models in humanistic communities, can be useful to define a shared controlled vocabulary, describing only entities of interest for the communities themselves, without too much detail of description in the first phase.
- When dealing with elements representing parts of a text, meaning both material

and abstract entities, and relations among them, a top-down approach in conceptualization is recommended, in order to clearly distinguish different semantic layers of a document, i.e. a work, for referencing its context information; the expression of a work (i.e. the text) as an entity with part of speech, abstract divisions, and as subject of interpretations; its manifestation (i.e. the material representation of text) for describing concrete features of material support and characters.

- Finally, revising hierarchy of concepts and formalizing properties among classes, also a bottom-up approach can be useful. Analysis of data structure and specific use cases of elements in a corpus of XML TEI documents helps in defining further specializations of main concepts in iterative and concentric development.

3.1 TEI schema ontological generalization

Transforming an XML Schema into an OWL ontology involves a general rethinking of its element set, its organization and its related hierarchy of concepts/relations – from a flat hierarchical structure into a multi-layers one – where more complex semantic relations among entities can be stated and where relations among strings of text and their abstract containers shall become relations among real entities.

The conversion of XML Schemas into ontologies is an issue discussed in many papers in the last ten years, and for which many theoretical and computational solutions have been proposed. We cannot get into the technical details of these solutions here. Most of them are based on the mapping of W3C Schemas primitives into OWL primitives [e.g. 12].

TEI has made explicit its conceptual model with the notion of element class in the design of its literate schema language ODD:

> The TEI scheme distinguishes about five hundred different elements. To aid comprehension, modularity, and modification, the majority of these elements are formally classified in some way. Classes are used to express two distinct kinds of commonality among elements.[…] A class is known as an attribute class if its members share attributes, and as a model class if its members appear in the same locations. In either case, an element is said to inherit properties from any classes of which it is a member [13, 1.3].

And later, specifically about model classes:

> 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. In some cases, the 'set of places where class members are permitted' is very constrained: it may just be within one specific element, or one class of element, for example. In other cases, elements may be permitted to appear in very many places, or in more than one such set of places. [13, 1.3.2]

Guidelines state that the distinction between those two kind of model classes is epitomized by the naming conventions adopted:

> 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. [13, 1.3.2]

We can try to identify a proper structural constraints set and an informal semantic/taxonomic directives set from such explanations, but a well-formalized model has to reorganize such functional and pragmatic approach into a more balanced one which take in account formal logic constraints and rules for creation of a taxonomy. Drawing from this analysis of the TEI schema architecture, as a first approximation, we can formulate OWL constructs through a conceptual workflow following a few of minimal required steps.

- General analysis and recognition of entities in the Schema, i.e. all TEI elements that can be converted into OWL classes. This compels a wide conceptualization and study of entities involved in the wide domain of text encoding – and that will create the basic taxonomy of the ontology – which encompasses a description of the context of a document of interest (people and events related to the life cycle of a document), the document as an object itself (through FRBR conceptual model) and information that can be extracted from the content of the document (people, relations, events described in the text). We will then have well-known entities such as *Agent*, meaning a Person, an Organization or a Group somehow involved in the life cycle of the object of interest or simply cited in text); *Document,* identifying the document of interest and all related similar objects); *Time* and *Place* as entities related to documents, agents and events or simply described in the text; *Event* and *Situation* as broad entities for defining any sort of action, situation and specific issue related to the document life cycle, or also as described in the text; finally, we will have different entities for defining document elements, in order to identify both material – concrete- and conceptual elements related to the text. These entities don't cover the wide range of specific concepts needed in a complete description of the domain, but are minimal required entities in order to define a shared conceptualization, according with some of most known ontologies (see section 4).
- Analysis of TEI Model Classes of type "Like" and "Part". Generally, most of -Like type elements can peacefully be converted into OWL classes: this entails that an "automatic" transformation is allowed here for the formalization of such entities. E.g. elements of TEI Model Class `model.persEventLike` – birth, death, event, listEvent – can be transformed in OWL classes with the same name. However they have to be reanalysed in iterative controls for a correct hierarchical characterization without redundancy, wrong or badly conceptualization, following OntoClean methodology [14] as a correct way for creation of a taxonomy. E.g. considering previous example, *listEvent* have to be deleted as a wrong, unnecessary entity; *birth* and *death* have to be correctly declared in taxonomy as kind of events. Indeed, here "automatic"

doesn't mean strictly automated, because a general rule for such transformation is unpredictable. Furthermore, we noted that transformation of elements members of `-Like` type Classes into OWL Classes is not generalizable when these ones are also member of TEI Model Class of type `-Part`. When this situation occurs, in most cases `-Like` type elements, and also `-Part` type elements, can be converted in object and data properties. E.g. `model.nameLike` elements – like *name*, *orgName*, *persName* – when members of `model.addrPart`, can better be converted into data properties.

- Generalization on elements of `-Att` type Model Classes: these ones are also converted into OWL classes, restricting in an iterative way the scope of elements to be transformed, but neither here a right generalization is possible.
- Generalization on attributes: these ones can be converted into OWL datatype properties, whose domain is the union of all Classes (derived from TEI Model Classes) they apply to, and whose range is the datatype assigned to each of them in the XML Schema. However, they also have to be revised in order to define which attributes point to "real" entities and then have to be declared as object properties (e.g. `@who` attribute, a pointer to a person reference; `@source` attribute, a pointer to a bibliographical source).

This basic set of rules does not cope with the modeling into OWL of structural XML content models, for which a mix of OWL objectProperties and restrictions can be used as proposed in [15].

This mapping capture the basic semantic of the TEI XML schema as whole. Some more ontological axioms can be added to specify other semantic assumptions. In fact, OWL allows multiple inheritance of classes.

In many respects, the construction of a formal high-level TEI ontology could be a partially automated process starting from the implicit semantics in the schema. However, the most of semantic restrictions, which cannot be expressed by common Schema Languages (and ODD), should be explicitly and manually stated, as the most of issues related to the creation of a correct taxonomy itself.

3.2 Intensional semantics of the elements

We adopt the term intensional semantics since 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>`, `<seg>` or of the `@type` attribute that define an intensional, more specific and restricted semantics compared to that described at general ontology level (e.g. `<div>` could have a value associated to `@type` choosen from a controlled vocabulary suggested by the TEI model [`act|scene|chapter|part`] but it could be manage also values defined by a local community).

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. Semantic definitions through specialized formalisms such as EARMARK (see [16] and [17] with TEI-based samples).

How can a user possibly declare these local semantic extensions? The most obvious method is to adopt `<constraint>` element - or 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.

Once verified these situations they will have to be provided in the ontology, originally created starting from the TEI schema, in an iterative process (from documents to ontology and vice versa).

3.3 Extensional semantics of the markup content

However, this strategy does not cover the need to define semantically specific instances of a markup element. For example, assume that in a given markup application `<seg>` element is used as "manifestation of a character's feature". You may need to qualify a single instance of the element, for example, to indicate what particular feature you are encoding.

The last semantic level concerns the extensional semantics of the individual XML elements content within a document. We adopt the term 'extensional' because, in general, it is suitable for fixing the referent of a linguistic expression identified by the markup through its reference to resources (information entities) via URI, or the connection to items in Linked Data Set. This is a case already widely addressed in several projects.

The current TEI scheme already handles the case of simple extensional link with one or more external resource through the `@ref` attribute (whose value is one or more xsd: anyURI). More complex relations with external semantic data could require as complex standoff markup structures.

Three samples from a digital editon of a collection of letters by Vespasiano da Bisticci[5] shows three references `@ref` where the string in natural language could be treated as specific identified entity or resource through URI and it's connected with more complex formal description: proposopraphy form the person `<persname>`; codicology for the manuscript `<bibl>`; lexicography for the vocabulary `<term>`.

```
<persname ref=”http://vespasianoletters.it/people.xml#PS”>
      Piero Strozi
</persname>

<bibl ref=”http://vespasianoletters.it/manuscripts.xml#P_SN>
       <author>Prinio</author>
</bibl>

<term type=”binding” ref=” http://vespasianoletters.it/lexicon.xml#leg”>
          legaranno
</term>
```

4 Ontology alignment and Linked Open Data

As we said above, a particularly relevant aspect of the conceptual model definition process will be the check of the existing ontologies in order to ensure maximum portability in all contexts, in a hybrid approach to ontology development. The TEI

[5] Vespasiano da Bisticci, *Letters*, ed. by Francesca Tomasi, Bologna, 2013, http://vespasianodabisticciletters.unibo.it

ontologies Special Interest Group has already done some relevant work in this area, especially thanks to the work of Ore and Eide with CIDOC-CRM [18]. However, beside the most common existent ontologies devoted e.g. to cultural heritage (CIDOC-CRM also in FRBRoo version[6]), archives (EAD[7] and EAC-CPF[8]), metadata exposure (DC[9] and DC terms[10]), other ontologies, developed in other different domains, provide new form of conceptualization. For example, ontologies as FABIO and CITO could be an interesting application case [19]. FABIO is based on the FRBR approach to the document as a complex entity. The stratification of the levels of analyis, as we said above, enrich the description of cultural entities. CITO is useful in order to manage all the citation process, towards the definition of multiple relationships and cross-relationships between data.

This means that an early mapping stage between potential relevant ontologies will be necessary to align the TEI ontology to the most popular conceptual models[11]. We can assert that: once a conceptual model for TEI is defined the next step is the identification of all the pertinent existent ontologies.

Then the alignment process[12] will contribute to refine the model: already shared classes and properties could be encapsulated in the TEI conceptual model and specific classes and properties as a result of the TEI semantic extension could contribute to populate the cloud. In addition to possibility of exchange among models and then communities, ontology alignment is yet another step to ensure validity of conceptualization: indeed, as in an iterative workflow, first releases of the ontology for TEI have to be managed as feasibility studies, that can’t be immediately opened into Linked Data cloud.

Such approach, already used by other Schema conversions into ontologies like EAC-CPF [20], grants a granularity of description that is able to satisfy different needs at different times – firstly taxonomic consistency and then specific issues related to various approaches in markup semantics.

The project of TEI conversion into a LOD compliant version consists then in a sequence of steps that could be described as:

a. formalization of the TEI model by converting the schema into OWL classes and properties for a first macro-modelization;
b. revision of the resulted ontology by working on different corpora of XML TEI in order to refine specifications;
c. TEI ontological model mapping onto selected ontologies in order to guarantee interchange but also expressivity of the model in a reuse perspective;
d. adding URI to in-line markup, when needed, in order to be LOD-compliant.

To finalize then the model in a LOD perspective the following methods have to be explored:

e. creation of the RDF triple store by converting the refined XML TEI files

[6] http://www.cidoc-crm.org/frbr_inro.html

[7] http://www.loc.gov/ead/

[8] http://eac.staatsbibliothek-berlin.de/

[9] http://dublincore.org/

[10] http://dublincore.org/documents/dcmi-terms/

[11] See e.g. the Europeana effort in the EDM (Europeana Data Model) proposal: http://pro.europeana.eu/edm-documentation

[12] A first general discussion and set of tools in: http://www.ontologymatching.org

and then populating the LOD Cloud;

f. discover of links in the cloud by using semi-automatic methods of entity recognition in other datasets (e.g. `Dbpedia.org`).

5 Conclusions and perspectives

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 the management and research in open and multi-standard contexts, such as large-scale general libraries and large institutional repositories;
2. facilitate interoperability with other standards relevant in the Digital Cultural Heritage (CIDOC-CRM, EAD / EAC-CPF, METS, EDM) context and the inclusion of any XML / TEI repository in the Open Linked Data environment. TEI could be redefined as a Linked Open Vocabulary able to dialog with other LOV datasets either at vocabulary or element level (PREFIX `tei:`)
3. ease the conversion existent TEI based digital libraries in open and linked datasets able to share the LOD cloud. In an aggregator dimension, as the Archive Hub Linked Data[13], the TEI triplestore could benefit from the relationships with pertinent datasets at all the level of features' description
4. provide users with advanced formal tools to define their interpretations of the texts they apply the markup to and give, in this way, the possibility of innovative computational processing based on semantics intended as a reasoner and semantic query engines.

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

A possible candidate for a test-bed of the ideas presented in this paper could be the forthcoming "TEI Simple" (formerly known as "TEI Nudge" [21]) customization of the TEI scheme. We are looking forward for the first results of the project to start a practical experimentation.

6 References

1. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer, Berlin Heidelberg (2009)
2. Cover, R.: XML and semantic transparency. Technology report, CoverPages. `http://www.oasisoprn.org/cover/xmlAndSemantics.html` (1998)
3. Burnard, L.: Resolving the Durand Conundrum. Journal of the Text Encoding Initiative 6. DOI: 10.4000/jtei.842 (2013)
4. Renear, A., Sperberg-McQueen C.M., Huitfeldt C.: Towards a semantics for XML markup. In: Furuta, R., Maletic, J. I., Munson, E. (eds.), DocEng'02. Proceedings of the 2002 ACM Symposium on Document Engineering. ACM Press, McLean, VA, New York, (2002)
5. Renear, A., Sperberg-McQueen, C.M., Huitfeldt, C.: Meaning and interpretation of markup. In: Markup Languages: Theory & Practice 2 (3). MIT Press, Cambridge, MA (2000)
6. Tummarello G., Morbidoni C., Pierazzo E.: Toward textual encoding based on RDF.

[13] `http://datahub.io/it/dataset/archiveshub-linkeddata`

In: ELPUB2005. Challenges for the Digital Content Chain: Proceedings of the 9th ICCC International Conference on Electronic Publishing. Peeters Publishing, Leuven (2005)

7. Peroni S., Gangemi A., Vitali F.: Dealing with Markup Semantics. In: Ghidini C., Ngonga Ngomo, A., Lindstaedt, S., Pellegrini, T. (eds.), Proceedings the 7th International Conference on Semantic Systems. ACM, New York. DOI: 10.1145/2063518.2063533 (2011)
8. Sperberg-McQueen, C.M., Huitfeldt C.: What is transcription? Literary & Linguistic Computing 23.3, 295–310 (2008)
9. Sperberg-McQueen, C.M., Huitfeldt C., Marcoux Y.: What is transcription? Part 2. Talk given at Digital Humanities 2009, College Park, Maryland. Slides on the Web at `http://blackmesatech.com/2009/06/dh2009/` (2009)
10. Sperberg-McQueen, C. M., Huitfeldt C., Marcoux Y.: Transcriptional implicature. A contribution to markup semantics. Paper given at Digital Humanities 2014, Lausanne, Switzerland (2014)
11. Shadbolt N., Hall W., Berners-Lee T.: The Semantic Web Revisited. IEEE Intelligent Systems Journal, May/June 2006, 96-101 (2006)
12. Ivezic N., Marjanovic Z.: Mapping XML Schema to OWL. In: Enterprise Interoperability. Spinger, Berlin Heidelberg, 243-252 (2007)
13. TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, `http://www.tei-c.org/Guidelines/P5/`
14. Guarino, N., Welty, C.A.: An overview of OntoClean. In: Handbook on ontologies. Springer, Berlin Heidelberg, 201-220 (2009)
15. Bedini, I, Gardarin G., Nguyen B.: Transforming XML Schema to OWL Using Patterns. In: 5th IEEE International Conference on Semantic Computing (ICSC), Palo Alto (2011)
16. Peroni S., Vitali F.: Annotations with EARMARK for arbitrary, overlapping and out-of order markup. In: Proceedings of the 2009 ACM Symposium on Document Engineering (DocEng 2009), 171-180. ACM, New York. DOI: 10.1145/1600193.1600232 (2009)
17. Barabucci G., Di Iorio A., Peroni S., Poggi F., Vitali F.: Annotations with EARMARK in practice: a fairy tale. In: Tomasi F., Vitali F. (eds.), Proceedings of the first Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities (DH-CASE 2013). ACM, New York. DOI: 10.1145/2517978.2517990 (2013)
18. Eide Ø., Ore C.E.: TEI, CIDOC - CRM and a Possible Interface between the Two. Digital Humanities 2006. First ADHO International Conference, 62-65 (2006)
19. Peroni S., Shotton D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web 17, 33-34. DOI:10.1016/j.websem.2012.08.001 (2012)
20. Mazzini S., Ricci F.: EAC-CPF Ontology and Linked Archival Data. In: Proceedings of the 1st International Workshop on Semantic Digital Archives (SDA), `http://ceur-ws.org/Vol-801/` (2011).
21. Mueller M.: TEI-Nudge or Libraries and the TEI, Center for Scholarly Communication & Digital Curation Blog, `http://cscdc.northwestern.edu/blog/?p=872` (2013)

*All web sites were last visited on 5 December 2014