

1. Background

In the last few years, the TEI Ontologies SIG has been working on how world knowledge is expressed in TEI documents, in connection to other standards such as CIDOC-CRM and FRBR. This work has been reported in the meetings of the SIG as documented on the SIG Wiki (<http://wiki.tei-c.org/index.php/SIG:Ontologies>). In addition to the papers listed on the Wiki, an article titled “TEI and cultural heritage ontologies: Exchange of information?” was recently printed in *Literary and Linguistic Computing*.

As agreed at the SIG meeting in November 2007, an important next step will be to “start on the development of guidelines for how to create TEI documents that easily may be mapped to ontologies such as the CIDOC-CRM”. This document comprises a draft for such a set of guidelines.

2. Introduction

History of TEI

History of conceptual modelling/ontologies and of some models, e.g. CIDOC-CRM, FRBR, Dublin Core and literary ontologies.

All these standards have been developed in a context, for a purpose. The histories are important in order to understand that, which will help in understanding why they are as they are, and why things seem as shortcomings are there for a reason.

TEI has often been used to mark up texts without taking into consideration the world outside the text. In many cases, this is a good approach. But sometimes, one would like to add information that is related to an external world. One example is a historical document with many references to persons and places. A text internal approach would be to register all the names, using the appropriate TEI element types. But if one is interested in making an index of all the persons mentioned in the text, one is moving outside the text. The text contains names of the persons, but the reason to say that some of these names have a special connection is that they refer to the same physical or imaginary person.

Many of the modules in TEI can be said to have an implicit ontology. The same way as TEI makes textual features existing in the text explicit, such modelling will enable us to make implicit conceptual structures explicit.

If one wants to do this, a reference to the person will need to be included in the data connected to the document, typically in the markup. This can be done in different ways, as will be discussed below.

In the following, such an approach will be called conceptual modelling. There are many different reasons for wanting to do this, and many potential end results. One project will often want to pursue several results of this single process.

One result may be a CIDOC-CRM mapping of information from historical documents, used for import into a cultural heritage management system. Another may be an export into FRBR in order to connect the content in TEI documents to a library database. A third may be a mapping into Dublin Core in order to include information from TEI documents into a web publishing system, a fourth would be mapping to formats under development by Google or Yahoo.

Whatever the use may be, the method will open up for inclusion of TEI data in the semantic web. Not just the documents as items, but the world information described inside them. It opens for doing this on different levels of complexity, at different stages. The whole process can be pretty simple. Or one can make a complex mapping, but still export simpler versions from the mapping when that is requested. Converting from CIDOC-CRM to Dublin Core, for instance, is a well defined process. This also will enable mappings to future standards to come.

In sum, this will combine the strengths of TEI and conceptual models (ontologies) to a very robust and usable package.

2.1. What to map?

A good advice is to think about conceptual models from the outset. This should be part of the data analysis, in asking why to mark up, leading into what to mark up. What is the target ontology, what is the purpose of the markup?

If one is to produce a simple list of names, the needs may be different than if the result should be a thesaurus. The methods may also be different.

Types of information in the text to be mapped:

- person names and other referring strings denoting concepts of person-type.
- place names and other referring strings denoting concepts of place-type.

Referring strings ("he", "that place") is often equally interesting as names.

Types of information in the header: Personlists with person elements, Placelists with place elements. There will be a one-to-one relationship between a person element and a real life/fictionous person.

2.2. Types of texts

There are no principle differences between fiction and non-fiction, but the ontologies one will want to map to will differ.

If one want to record information about acting entities that are not persons (storms, animals, washing machines), e.g. in fiction or ethnographical texts, one may want to adjust TEI by adding an element agent similar to person.

3. Howto

The modules of TEI most important for this work is ... Here, we will list some elements that often will be important in mappings: name, person, place, event, ...

Name type elements will always be encoded in TEI. For the person type elements, they may be encoded in TEI, typically in the header or a separate section in the body, or they may not be encoded in TEI and only be stored in an external conceptual model. In the former case, links will go from name type elements via person type elements to the conceptual model (and backwards), in the latter case, the links will go directly between the name type elements and the conceptual model.

How the encoding of the TEI document is done will depend on the conceptual model chosen. Thus, it is good to start considering the model as soon as possible in the planning of textual encoding work.

Certain TEI elements may contain information that could be hard to map because some necessary information, e.g. reasons why something is asserted and who is responsible for the assertion, may be hidden. Some will also be based on a possible not formally available "point zero", such as age.

3.1. Definitions

Define ontology and conceptual models:

"The term ontology means literally the study of being and was until recently the name of a branch of philosophy and a term used in the singular only. During the last ten years the term has been adopted by computer and information sciences and the scope of the term has been expanded significantly. Today, it may denote everything from data models to classification

systems and explanatory models in natural sciences. In this article we use ontology in the meaning conceptual model. That is, a formally defined model resulting from an analysis of a specific domain and not necessarily a data model in the computer science sense." (Ore & Eide: "TEI and cultural heritage ontologies: Exchange of information?" LLC 24(2) 2009)

3.2. Where to store the ontologies

Three options:

- Store the information as an external XML- document, e.g. RDF (RDF, 2008) or CRM-Core (Sinclair et al., 2006). This external document can then be stored with the TEI document, e.g. by using METS.
- Store the information in the TEI-header using an external XML name-space, e.g. RDF or CRM-Core.
- Store the information in the TEI-header using the existing elements in TEI P5.

What formalism should it be stored in, TEI or the external conceptual model?

3.3. Examples

3-4 examples in here: TEI-->CRM, TEI-->FRBR?, TEI-->literary ontologies, TEI-->DC. Mappings to systems developed by Yahoo and Google as well?

3.4. Different approaches to integration

Two different approaches for information integration (ontology connection):

- Only work with strings (e.g. place and person names) in the TEI document, everything else outside
- External world information in TEI header, possibly in a separate TEI document if that is more convenient.

"Real" world module in TEI? Not module, but elements could be divided into yes/no relation to text-external world. The YES ones can be grouped, e.g. actors, events, ...

Maybe we should make suggestions, not guidelines?

Need high level of explicitness in order to interpret encoded texts so that the information they express can be modelled in a conceptual model. This explicitness can be added to the markup, or it can be in the extraction algorithm what is the best trade-off?

3.5. Local ontologies in TEI

Special ontologies already in TEI:

- TEI header: Library ontology
- Oral corpus
- Manuscript description

3.6. Should events be marked up in TEI? HOW

In running texts, this could be done by `<rs type="event">` or `<milestone type="event"/>`.

Events are more difficult than persons and places, all and everything, hard to define what to include.

Possible criteria for marking up an event:

- Makes material changes in a person or a place.
- The existence of a date element.

- Person-place-date all have possible connection between them - any event that functions to attach two or more of (person, place, date) should be marked up.

If one is working with/looking for events in TEI documents, one should be aware of the events to be found in manuscript descriptions and transcriptions of oral sources.

3.7. Relations HOW

Need more than nesting in order to connect values as in marriage example. In very simple cases nesting may do, but soon cases will appear when the relationships are too complex, such as "this also applies to the persons discussed in the last paragraph".

Will relations always be between place/person type elements, and not name type elements?

No: Cannot put relation between person element, relations have to be connected to the context, often the name, commonly the event (e.g. or marriage) This does not mean that two strings of characters marry, but it means that a marriage cannot be seen as a relationship taken out of time and place - it has to be connected to a place in the text.

Discuss the marriage example (TEI P5 sec. 13.3.2.3 Personal Events) in details.

3.8. Persons and perspective HOW

Perspective of a person is important in fictions, but also in types of non-fiction: Who is the speaker/thinker/creator of views inside the text.

3.9. Applications

Any project who are going to use these methods will need some sort of application to do the actual extraction of information from the TEI documents. Such applications are often written in XSLT, but any scripting or programming language could be used, e.g. PERL or PYTHON.

Even if it is impossible to make tools for mapping of all possible TEI documents, it would still be a good idea to develop applications that can be used to extract conceptual models from specified groups of TEI documents. Such applications could be used as is by some users, whereas others can use them as a base for developing their tailor-suited systems. This would be similar to the XSLT stylesheets available for transformation into HTML and PDF.

It is also important to store mappings already done, including the ODD documents describing the TEI source documents. This could be in a form of a library handled by the TEI Ontologies SIG.

This may also overlap with the TEI Tools SIG.

4. Conclusion

4.1. Different levels

Three levels of document collections:

- All TEI documents
- A group of TEI documents
- A single TEI document

Conclusion/suggestions: Ontology mapping cannot be defined for 1, but could be defined for 2 and 3. If mappings are done on these levels, publish them! Including the ODD defining the TEI version being the source of the mapping. Building up a library in connection to these guidelines?

Conceptual models are also a good starting point for more or less complex indexes and search structures.

One of the good things with this approach is that once your data is stored in a conceptual model, converting to simpler structures (Dublin Core, google-friendly models) will be easy. Should be suggest that toole would be developed for this?