From TEI to a CIDOC-CRM Conforming Model: Towards a Better Integration Between Text Collections and Other Sources of Cultural Historical Documentation

Øyvind Eide *(oyvind.eide@muspro.uio.no)*
Unit for Digital Documentation
University of Oslo
Christian-Emil Ore *(c.e.s.ore@edd.uio.no)*
Unit for Digital Documentation
University of Oslo

Introduction

In the last couple of years, there has been a growing interest towards including into TEI documents information about the world rather than information concerning the text of the document to be encoded only. We see examples of this through recent additions to the TEI standard, e.g. the person element (TEI P5, sec. 20.4.2), as well as through the work in the Ontologies SIG since it was established in 2004 (TEI Ontology SIG WIKI). In the SIG, the topic of discussion is how to organise this kind of information about the world according to specific ontologies.

One particularly promising ontology in this context is the CRM (CIDOC 2003). It has been used together with Topic Maps to organize information from TEI documents (Tuohy 2006) and as an attempt to find a solution to the so-called exhibition problem (Eide 2006). Further, attempts have been made to formalise a way to connect TEI and CRM documents (Ore 2006).

In this paper, we propose a method for automatic generation of CRM conforming models based on TEI documents. We will discuss limitations to this approach, as well as ways these may be overcome.

Our proposed method

The method we propose will include two important steps that should be possible to implement to any given TEI document: Mapping and model building.

Mapping

A mapping from the TEI document into a model conforming with CRM should be created. It will be based on a general mapping of TEI elements to CRM we are currently developing. But in TEI, many elements are defined quite loose, and depending on the way they are used, they may be modelled differently in CRM. According to the TEI guidelines, tag usage may be described in the TEI header. Such descriptions may help to decide which type of modelling is the most appropriate.

Ideally, such a specific mapping should be created based on an automatic reading of the TEI header. But an element description in a *tagUsage* element in the TEI header is in prose and will generally not be stringent enough to be understood by an automatic reading (TEI P5, sec. 5.3.4). Human interaction will be needed. It may be the case that use of the *equiv* element will make automatic creation of mappings possible, as a reference to a certain CRM class may be included as an external link (ibid, sec. 6.3.4).

Model building

A CRM conforming model based on the TEI document and populated with all instances of mapped elements should then be created. This model may be used as a query or a data mining system where the user looks for interesting structures in the CRM conforming model alone, as well as in combination with textual information collected from the TEI source document. But this model may also be used in connection with other CRM conforming models, such as museum databases. The connections will be based on regional or global object identification, such as authority lists of names and classification schema. The resulting "super model" may then be used as a data mining tool based on semantic integration between heterogeneous resources.

An example of a TEI input document

We are currently developing the building blocks for a system based on the method described above. The example text used in our work is taken from a manuscript describing examinations about geographical matters performed as court interviews in 1740s, printed in the 1960s (Schnitler 1962). The printed edition was digitised and marked up in a typical TEI way, with names of places and people as well as dates tagged. A short paragraph of this tagged version, based

on page 73 in the printed book, is translated to English and included below.

 Answ: Named <name type="person" xml:id="s1_24454"> Ole Nilsen</name>, is born in <name type="place" xml:id="s1_24457"> Tydals </name> mountains, Which is in
<name type="place" xml:id="s1_24460"> Norway </name>, of Sami parents, is 50 years old, married, and having one child; has mostly dwelled in <name type="place" xml:id="s1_24466" > Tydals </name> mountains, and now dwelling in the Norwegian <name type="place" xml:id="s1_24469"> Mærragers </name> mountains.

What is tagged — and what is not

Names and dates are tagged in our document. This means that many references to persons, places and times are included. But there are a lot of other references to similar real world entities that are not tagged, e.g. when words other than proper names are used to refer to them. In the example above, "one child" is not tagged, whereas other references to the same historical person in terms of his or her name are tagged as person names. Furthermore, events are not tagged in this text. Thus, whereas the name of a boy who is born and the place of his birth is tagged, the event connecting these together, i.e. the birth, is not marked up.

This is common to most TEI documents, and is based on the text centric tradition of the TEI community. There were good reasons why this tradition was established, but it may be the case that for some types of documents, textual references to e.g. events should be marked up.

A simple CRM model of parts of the paragraph above is included as Figure 1. The solid lines represent what can be directly read from XML elements in the TEI document, whereas the dotted lines shows the parts not based on the TEI markup. This shows that the information needed to model the connection between the person and the mountain - that he was born there - is not tagged in the TEI document.

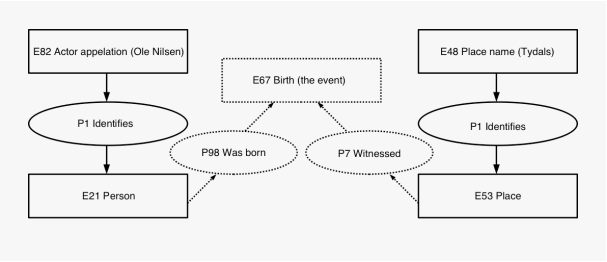

Figure 1

How to find the events

As the models created through the use of our proposed method is based on the XML tags alone, important information in the text is not considered when the model is created. Several possible ways to improve this exist. One is automatic event detection, as used e.g. in the Perseus project (Smith 2002). The use of this method too causes some problems. One problem is that fact that even if the system works quite good for English material, considerable work remains to be done to create a similar tool for 18th century Danish, even though the method is implemented also for smaller languages such as Finnish (Makkonen 2003). Named entity recognition systems are developed for modern Danish (Johannessen 2004), but this is of no use in the common situation where names are already tagged in the documents. Another problem is the fact that these kinds of methods will always be imperfect, resulting in either missed events, false positives, or both of these.

Another way to solve the problem would be to reread the text, identifying events and tagging them. This would be a reliable method, but time-consuming. Even a combination of these strategies, a semi automatic method, would mean quite a lot of work.

A possible way to solve the problem of event identification is to use the model. Any person or place in the CRM model has a link to the name of the person or place, an from the name in the CRM model to the TEI name element. This means that the textual distance to other person name elements, place name elements and date elements can be calculated. Further, it is possible to locate all words within a certain distance, and all words between two names.

This is similar to the first approach above. But in addition, we propose to connect the CRM model to external databases. An example of this would be if we have a CRM version of a database based on church book records. In this church book based CRM model a person may be found with a name similar to Ole Nilsen, a birth date in a possible range for being a witness in 1742, and a birth place with a name similar to a place name mentioned in the text in connection to the person's name. This external source may then help us to include the E67 Birth event in our model. This may turn out to be impossible without manual work, but we hope at least to make the manual work more effective.

Conclusion

A general observation from our work is that the more relevant information types is marked up in a TEI document, the easier it is to use automatic methods to generate CRM conforming models. But even a limited tagging with only

names and dates marked up do help. We will continue our work on the implementation of a system based on the method described in this abstract. We believe this will improve the usability of TEI documents as information sources as well as simplifying the process of adding more information, such as event elements, into such documents.

Bibliography

TEI Ontology SIG WIKI. Accessed 2006-11-12. <http://www.tei-c.org.uk/wiki/index.php/SIG:Ontologies>

Burnard, Lou, and Syd Bauman, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Ver. 0.5. Accessed 2006-11-13. <http://www.tei-c.org/release/doc/tei-p5-doc/html/>

CIDOC. "Definition of the CIDOC Conceptual Reference Model." ISO/DIS 21127. 2003. Accessed 2006-11-13. <http://cidoc.ics.forth.gr/definition_cidoc.html>

Eide, Øyvind. "The Exhibition Problem. A Real Life Example with a Suggested Solution." *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006. 58-61.

Johannessen, Janne Bondi, Eckhard Bick, Kristin Hagen, Dorte Haltrup, Åsne Haaland, Andra Björk Jónsdottir, Dimitrios Kokkinakis, Paul Meurer, and Anders Nøklestad. "The Nomen Noscio Project - Scandinavian Named Entity Recognition." *ALLC/ACH 2004 Conference Abstracts.* Göteborg: Göteborg University, 2004.

Makkonen, Juha, and Helena Ahonen-Myka. "Extraction of Temporal Expressions from Finnish Newsfeed." *Proceedings of 14th Nordic Conference of Computational Linguistics (NoDaLiDa 2003).* Reykjavik, 2003.

Ore, Christian-Emil, and Øyvind Eide. "TEI, CIDOC-CRM and a Possible Interface between the Two." *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006. 62-65.

Schnitler, Peter. *Major Peter Schnitlers grenseeksaminasjonsprotokoller 1742-1745. Bind 1 [Major Peter Schnitler's border examination protocols 1742-45. Volume 1].* Oslo, 1962.

Smith, David A. "Detecting Events with Date and Place Information in Unstructured Text." *International Conference on Digital Libraries. Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries.* Portland, 2002. 191-196.

Tuohy, Conal. "Topic Maps and TEI – using Topic Maps as a tool for presenting TEI documents." *TEI Day in Kyoto 2006.* 2006. 85-98.