Øyvind EIDE
Christian-Emil ORE

TEI, CIDOC-CRM and a Possible Interface between the Two

In the work of the TEI Ontologies SIG there have been an interest in finding practical ways of combining TEI encoded documents with CIDOC-CRM models. One way of doing so is including CIDOC-CRM information in a TEI document and linking CIDOC-CRM elements to TEI elements where appropriate. In this paper, this method is described through an example, together with an outline of the additional elements necessary in the TEI DTD used.

Background

In projects at the Unit for Digital Documentation, University of Oslo, we have created SGML and later XML encoded versions of printed and hand-written museum documents, such as acquisition catalogues, for more than ten years (Holmen 1996). To be able to store such documents in a standard format, we are planning to use TEI. Much of our material are archaeological documents, and there have been a growing interest in the use of XML in general and TEI in specific in archaeological community the last few years (Falkingham 2005, sec. 3.3, cf. sec. 4.3 and 5.2.3).
We also use CIDOC-CRM as a tool for modelling the contents of such tagged documents as they are read by museum professionals. We use this method to be able to include information from XML encoded documents in our museum inventory databases, with references back to the encoded documents (Holmen forthcoming). We would like to store CIDOC-CRM models in close relation to the TEI encoded document. This paper describes an example of how we try to define a syntax in which to store such datasets.

Extension of a TEI DTD

There are two different ways in which to extend a TEI DTD for inclusion of CIDOC-CRM models. We may include an element for each and every entity and property used in the model, or we may just include one TEI element for CIDOC-CRM entities and one for properties. We have chosen the latter method. This gives a limited and rather simple expansion of the DTD. This is similar to the way the XML version of the bibliographic standard MARC is designed (MARCXML).
This method will make it possible to create one document storing both textual markup and semantic interpretations of a text, while keeping the two parts of the document separate, except for links between specific elements in the two parts. This means that the document can be published as a text as well as form the base of an import to a database of records based on the interpretation, keeping the links back to the original text.
In this paper, we use a DTD fragment to show an outline of the extensions we need. The extensions is composed of a root **crm** element including a number of **crmEntity** elements and a number of **crmProperty** elements.

The root CIDOC-CRM element

```
<!ELEMENT   crm          (crmEntity*, crmProperty*)>
<!ATTLIST   crm
            id           #ID>
```

The entity element

```
<!ELEMENT   crmEntity    #PCDATA
<!ATTLIST   crmEntity
            id           #ID
            typeNumber   #NUMBER>
```

The property element

```
<!ELEMENT   crmProperty  #EMPTY
<!ATTLIST   crmEntity
            id           #ID
```

typeNumber #NUMBER
from #IDREF
to #IDREF>

Example of use

A typical situation in which this approach could be used is in archaeological documents. We have created a short dummy document containing some of the informations types commonly existing in our museum documents, as shown in Example 1.

> The excavation in Wasteland in 2005 was performed by Dr. Diggey. He had the misfortune of breaking the beautiful sword (C2343B) in 30 pieces.

Example 1

A tagging of this could be made as in Example 2.

```
<p id="p1">The
  <ab id="e1">excavation in
    <name type="place" id="n1">Wasteland</name>
  </ab> in
  <date id="d1">2005</date>
  was performed by
  <name type="person" id="n2">Dr. Diggey</name>.
  He had the misfortune of
  <ab id="e2">breaking
    <ab id="o1">the beautiful sword</ab>
    in 30 pieces
  </ab>.
</p>
```

Example 2

There are many objects and relations of interest when modelling the archaeological world described in this text. A typical museum curator reading could include the elements shown in Table 1.

1. A place identified by a name documented in n1.
2. A person identified by a name documented by n2.
3. A time identified by a date documented in d1.
4. An event (the excavation) documented in e1.
5. An event (the breaking) documented in e2.
6. An object (sword) documented in o1.
7. Dr. Diggey participated in the excavation
8. Dr. Diggey and the sword participated in the breaking
9. The excavation took place at the place identified by a name documented in n1 and at a time identified by a date documented in d1.

Table 1

A possible CIDOC-CRM representation of one of the entities in Table 1, the excavation in line 4, is shown in Example 3. Included are also references to lines 2, 3, 7 and 9.
Note that Example 3 is only showing part of a model that would represent a normal archaeological reading of the paragraph above. E.g., the date should have a "is documented in" property such as the ones for the activity and the person, and the place (Wasteland) should be documented in a way similar to the person Dr. Diggey.

E7 Activity	--> P2 Has type	--> E55 Type[1]				
	--> P14 Carried out by	--> E21 Person	--> P131 Is identified by	--> E82 Actor appelation[2]	--> P70 Is documented in	--> E31 Document[3]
	--> P4 Has time-span	--> E52 Time-span	--> P78 Is identified by	--> E50 Date[4]		
	--> P70 Is documented in	--> E31 Document[5]				

1) archaeological excavation
2) Dr. Diggey
3) the element identified by the id "n2" in the text of Example 2 above
4) 2005
5) the element identified by the id "e1" in the text of Example 2 above

Example 3

Example 4 shows this using the TEI-CRM syntax outlined in the DTD addition above. The **crm** element holds the small CIDOC-CRM model we have expressed in a TEI syntax, while the **link** element holds connections between the CIDOC-CRM model and the TEI text from Example 2. In this example we see that although all the CIDOC-CRM information may be expressed in such a syntax, an XML validation of the document will only validate a part of the information. It will not check whether the model adheres to the rules for e.g. which CIDOC-CRM properties may be used in connection to which entities.

```
<crm id="crm-mod1">
  <crmEntity id="ent1" typeNumber="7"></crmEntity>
  <crmEntity id="ent2" typeNumber="55">archaeological
    excavation</crmEntity>
  <crmEntity id="ent3" typeNumber="21"></crmEntity>
  <crmEntity id="ent4" typeNumber="82">Dr. Diggey</crmEntity>
  <crmEntity id="ent5" typeNumber="31"></crmEntity>
  <crmEntity id="ent6" typeNumber="52"></crmEntity>
  <crmEntity id="ent7" typeNumber="50">2005</crmEntity>
  <crmEntity id="ent8" typeNumber="31"></crmEntity>
  <crmProperty id="prop1" typeNumber="2" from="ent1" to="ent2"/>
  <crmProperty id="prop2" typeNumber="14" from="ent1" to="ent3"/>
  <crmProperty id="prop3" typeNumber="131" from="ent3" to="ent4"/>
  <crmProperty id="prop4" typeNumber="70" from="ent4" to="ent5"/>
  <crmProperty id="prop5" typeNumber="4" from="ent1" to="ent6"/>
  <crmProperty id="prop6" typeNumber="78" from="ent6" to="ent7"/>
  <crmProperty id="prop7" typeNumber="70" from="ent1" to="ent8"/>
</crm>
<linkGrp type="TEI-CRM interface">
  <link targets="#ent5 #n2"/>
  <link targets="#ent8 #e1"/>
</linkGrp>
```

Example 4

Conclusion and further research

While different uses of ontological models in connection to TEI documents will differ in their technical solutions, e.g. whether the ontological model rests in a separate document or not, and which syntax is chosen for the model, the three main elements shown here have to be present:

- a TEI document

- an ontological model expressed in some XML syntax
- link elements to connect the specific elements from the two together

We have described a way of expanding TEI that gives us the tools we need to include a CIDOC-CRM model in a TEI document, and connect specific CIDOC-CRM entities to specific TEI elements in the non-CRM part of the document. We would like to see research into similar methods of connecting informations in other ontological systems to TEI documents, to discover whether a similar method is feasible. It would also be interesting to see if it is possible to make a general addition to TEI for this use, or if each ontological system needs its own tag set.
In our own research, we will write out an ODD to test this method on samples of our own data, and then continue to implement this model on real data, so that the usability of this method for complete documents and CIDOC-CRM models can be examined.

References

CIDOC-CRM. *ISO/FDIS 21127. Information and documentation -- A reference ontology for the interchange of cultural heritage information* [Definition of the CIDOC Conceptual Reference Model].
Falkingham, G. (2005) *A Whiter Shade of Grey: a new approach to archaeological grey literature using the XML version of the TEI Guidelines.* Internet Archaeology, issue 17. URL: http://intarch.ac.uk/journal/issue17/falkingham_toc.html (as of 2005-11-14).
Holmen, J.; Uleberg, E. (1996) "Getting the most out of it - SGML-encoding of archaeological texts." Paper at the IAAC'96 Iasi, Romania. URL: http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html (as of 2005-11-14).
Holmen, J.; Jordal, E.K.A; Olsen, S.A.; Ore, C.E. (forthcoming) "From XML encoded text to objects and events in a CRM compatible database. A case study". In: *Beyond the Artifact. Proceedings of CAA 2004, Computer Applications and Quantitative Methods in Archaeology.*
MARCXML. *MARC 21 XML Schema.* URL: http://www.loc.gov/standards/marcxml/ (as of 2005-11-14).
TEI Ontologies SIG. URL: http://www.tei-c.org/Activities/SIG/Ontologies/ http://www.tei-c.org/wiki/index.php/SIG:Ontologies (as of 2005-11-13).
TEI P5 (2005) *Guidelines for Electronic Text Encoding and Interchange. [draft] Version 0.2.1.* TEI Consortium, 2005.