

TEI and cultural heritage ontologies: Exchange of information?

(Draft, Accepted for the special issue of Journal of LLC on DH2008, Oulo Finland)

Christian-Emil Ore (c.e.s.ore@edd.uio.no), Øyvind Eide (oyvind.eide@edd.uio.no)

Unit for Digital Documentation, Humanities Faculty, University of Oslo, Norway

Introduction

Since the mid-1990s there has been an increase in the interest for the design and use of conceptual models (ontologies) in library science as well as in humanities computing. In text-oriented humanities computing, however, conceptual models and ontologies seem to be more connected to database development than to text research.

Reproducibility of results is a core concept in text-based research as in all research. The content in information systems and virtual reconstructions in the cultural heritage sector are to a large degree directly based on information deduced from text studies. In many cases the links from the information system back to the text are not available, and they may be complicated to re-establish. Even if it is possible to re-establish them, the process may be too expensive. These links are necessary to enable reproducibility of the deduction, since they document how the conclusions are based on the texts.

How should structured information, based on a reading of a text, be linked to the encoded text itself? It is important to base such linking on data standards evolved in the fields of text encoding and conceptual modelling. Thus, the understanding of text encoding represented by the TEI guidelines and the understanding of conceptual models represented by initiatives like the CIDOC CRM and FRBRoo should be combined.

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.

Archaeology, information extraction and encoding of texts

The discussion in this article is based on our work with archaeological collections and so called grey literature in the Norwegian University Museums, as well as with other electronic text collections established in the two large digitization projects, The Documentation Project (1992-1997) and the Museum Project (1998-2006).

Archaeologists have been using computer-based methods since the 1960s. Over the years, computer applications in archaeology have primarily been used for statistical analysis as well as to create inventory databases. Pattern recognition and simulation/AI were at their peaks around 1980 and 1990, respectively. In the 1990s GIS, 3D modelling and the Web were in focus (Scollar 1999). Most applications were designed to analyse information collected during fieldwork.

Like other scholars, archaeologists use texts, but text philology as such is not central to their discipline. They tend to have a simpler view of texts than the scholars researching texts per se. As a result of this, the task of creating a database on the basis of old reports is normally done through reading the text and keying into a database form the information considered to be essential in a normalized form. This method is fast, but the link to the original text is lost. Thus it is virtually impossible to check the correctness of the data at a later point in time and too costly to repeat the process with another focus.

A major part of the Norwegian Documentation/Museum Project was to create an information system for the archaeological museums in Norway. For almost 170 years the archaeological museums in Norway have published information on a yearly basis on their acquired artefacts in specially prepared acquisition catalogues. The description of finds in these catalogues are quite longwinded, including extensive information on the finds, the find contexts, its place and time, the finder or excavator, as well as detailed descriptions and classifications. The series of catalogues served for practical purposes as the main inventory for each museum. In general, old reports, catalogues and grey documents form the most important source of information about the museum collections. They also document the work of scholars in the field as well as in the museums. To be able to use the information available in these texts in an information system while keeping the integrity of the original texts, we used SGML/XML encoding as an intermediate step in the conversion process, see (Holmen & Uleberg 1996) and (Ore 1998).

There has been an increasing awareness of the need to include the information and content found in older archaeological and cultural-historically oriented documents into archaeological and cultural heritage systems. Our method of encoding and extracting information from electronic versions of old archaeological reports has been taken up by others (e.g. Crescioli, D'Andrea and Niccolucci 2002, Schloen 2001, Meckseper and Warwick 2003). Schloen suggests an XML formalism for storing and interchanging archaeological information. Meckseper and Warwick describe the situation in the UK and point out the usefulness of XML. Both address the question of how to use XML for writing and storing new archaeological documents.

In the course of the work in the two projects our main philosophy was to consider the old reports and catalogues as first-class objects in the information system. Thus the reports and catalogues were converted into electronic texts. In addition to traditional structural markup they were also given an extensive semantic SGML/XML markup. The encoding of the semantic content reflects our view on what constitutes important archaeological information and does not necessarily coincide with the concept of the 19th century author of the original material. The original author included information according to his concepts, from which we encode the subset that matches our conception or ontology. Thus it is important to include the original texts and make them available to future readers. This will enable them to judge for himself or herself. See (Holmen, Ore, and Eide 2004) for a more detailed presentation of the encoding process.

[1] The excavation in Wastland in 2005 was performed by Dr. Diggey. He had the misfortune of breaking the beautiful sword (C50435) into 30 pieces.

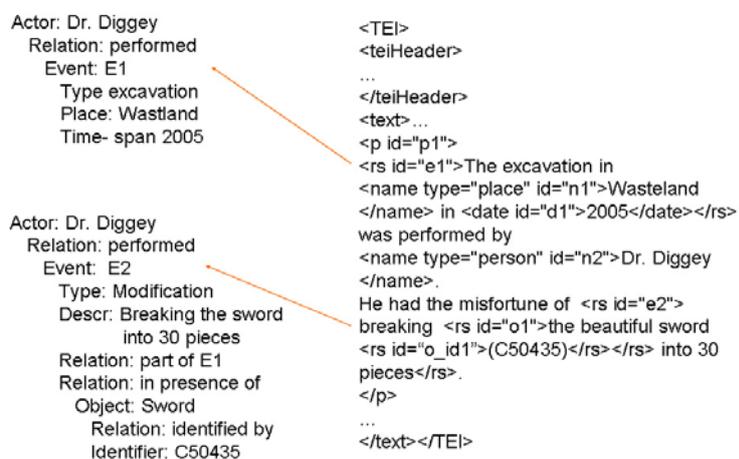


Figure 1: Information markup and extraction from the fragment in [1].

The example in Figure 1 illustrates the method. The small text found in [1] is a fragment of an imaginary archaeological excavation report and contains many of the problems encountered in actual archaeological grey literature. The left column in the example shows some of the information that can be read out of the text, expressed in a simplified CIDOC CRM form. The column to the right is the text fragment with a simple TEI markup. The elements indicate which parts of the text correspond to the information objects on the left. Two main events are identified. The reading assumes that the breaking of the sword occurred during the excavation and that the pronoun *he* point to appellation *Dr. Diggey*. This is not the only possible reading, but in this example, clearly the most probable. However, the uncertainty even in this simple case is an indication of why it is important to link extracted information to the source.

CIDOC CRM

CIDOC CRM is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM).

The CIDOC CRM is event-centric core ontology in the sense that the model does not have classes for all particulars like, for example, the Art and Architecture Thesaurus with thousands of concepts. The central idea is that the notion of historical context can be abstracted as things, people and ideas meeting in space-time. The model contains in addition identification of real world items by real world names (appellations), a generalized classification mechanism (types), part-decomposition of immaterial and physical things, temporal entities, groups of people (actors), places and time (time span), location of temporal entities in space-time and physical things in space, reference of information objects to any real world item (aboutness), and intellectual influence of things and events on human activities..

CIDOC CRM is defined in an object-oriented formalism which allows for a compact definition with abstraction and generalization through the inheritance mechanisms (ISA hierarchy). CIDOC CRM has 86 classes and 137 properties. The most central classes and properties for data interchange are shown in Figure 2.

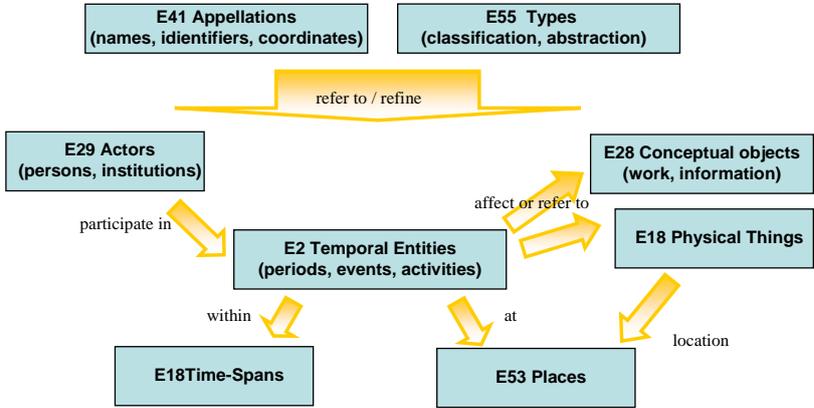


Figure 2: The CIDOC CRM, top classes for data integration

As an illustration of how events can be modelled in CRM, a traditional English wedding is used as an example. As shown in Figure 3, the event is at the core of the model, both conceptually and visually. Through this the event occurring at a specific location, the persons, the groom, the bride and the groom’s best man are connected. Their roles and the event itself are classified by the types (e.g. selected from a thesaurus).

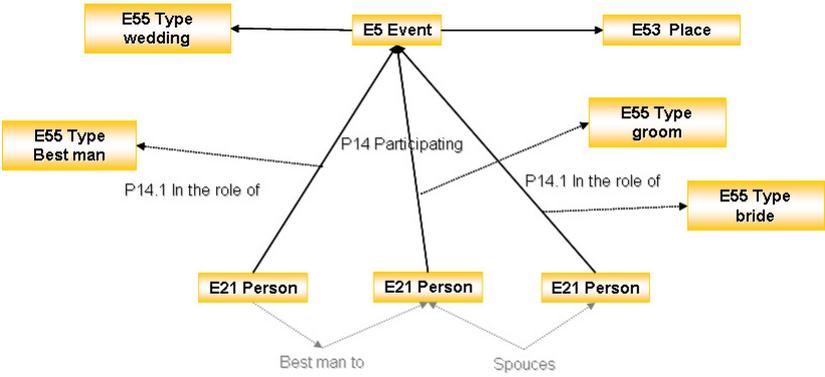


Figure 3: A CRM diagram for a traditional English wedding

In an information system the formal relation between the spouses has to be deduced by checking whether the two have participated in a wedding event in the role of bride and groom. Alternatively one may introduce short-cut relations indicated in grey at the bottom.

One of the basic principles in the development of the CIDOC CRM has been to have empirical confirmation for the concepts in the model. That is, for each concept there must be evidence from actual data structures widely used. Such evidence is found in mapping database schemata from different domains to the CIDOC CRM and from actual harmonization efforts

with competitive proposals. Even though the model was initially based on data structures in museum applications, most of the classes and relationships are surprisingly generic. The model was accepted by ISO in 2006 as ISO21127.

The archival standard EAD has already been mapped without any problems to the CIDOC CRM (Theodoridou and Doerr 2001). A more challenging task has been the harmonization of CIDOC CRM and IFLA's library ontology FRBR (FRBR 1998). The harmonization work was started in 2003 and has recently been completed. In the process it became clear that the CIDOC CRM didn't make sufficient distinctions on the level of concepts and symbolic representations. Some minor changes had to be introduced, see (FRBR_{oo} 2008) and (Doerr and LeBoeuf 2007). This was an example of how harmonization work can improve the models and encoding schemas. It is our hope that the study presented in this article can help reach a deeper understanding of the TEI P5 ontological modules and formal ontologies exemplified by the CIDOC CRM.

Text encoding and conceptual models

The Text Encoding Initiative (TEI) is a consortium which according to its website "collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics."

The TEI has published five major editions of its recommendations (P1-P5). The markup tool has shifted from SGML to XML. The most recent version of the guidelines, TEI P5, is completely XML-based and comprises XML schemata for adding XML-based markup to electronic texts. Using these guidelines and schemata, the texts' formal and logical structure, as well as inline annotations in the texts can be made available for further analysis and presentation. The TEI XML schemata are intentionally defined in a weak manner to cover most genres and text types. The TEI guidelines are focused on how to annotate texts and do not prescribe any specific conceptual model. However, central parts of the TEI recommendations, e.g. the TEI header, presuppose an implicit underlying conceptual model.

TEI has in its 20 years of history concentrated on the markup of functional aspects of texts and their parts. That is, a person's name is marked, but linking to information about the real-world person denoted by that name was not in the main scope. The TEI has not focused on the markup of the semantic content based on possible readings of a text, and there are not many examples of (TEI-based) markup of semantic content of texts. There may be several reasons for this situation. We see the following three as most important: Firstly, it is very time-consuming to do such encoding. Secondly, it is less objective and tends to be more contested than functional markup. Thirdly and perhaps more important, there has not been any well-developed uniform formalism or guidelines for the encoding of semantic content.

The scope of TEI has gradually broadened, however, to include more real world information external to the texts in question. The Master project (Master 2001) is an early example of this change. In TEI P5 a series of new elements for marking up real world information is introduced and several such elements from the P4 are adjusted. The TEI guidelines are intended to cover encoding of a large variety of texts in many cultural contexts and genres. For this reason one may argue that TEI should not have ontological elements at all or they should be independent of any specific ontology. However, this ideal is not possible and perhaps not feasible to follow completely. For example, the elements in the TEI header and the bibliographical module are quite naturally based on the traditional(implicit), conceptual model used by most librarians. The new and revised modules described in TEI P5, Chapter 13

Names, Dates, People, and Places, are defined without any explicit references to any specific ontology and are designed to cover a wide variety of real world descriptions. Here the authors of the guidelines apparently intend to follow an ontological neutrality ideal. It is however not possible to define this set of elements without having (several) implicit conceptual models in mind. This part of the TEI P5 may have been clearer if these mental models had been harmonized and made explicit in the guidelines. It is not our intention in this article to suggest an internal harmonization of TEI P5. Instead we have mapped the central part of the well-defined ontology CIDOC CRM into the set of these elements and thereby tested the expressive power of these elements.

As mentioned earlier, it is important for the reproducibility of scientific deduction to document on the basis of which part of the source text conclusions are made. Text fragments about and names denoting places, persons, events, etc., should be marked up in the running text and the corresponding structure deduced or added real world information in a separate structure in, for example, the header of an TEI-encoded text. Pointers should be used to connect the part of the texts that are used as the source for the objects and properties (e.g. names) to the structure in the header. In a CIDOC CRM compliant setting, for example, a TEI name element with the attribute type set to person should be interpreted as an instance of the CIDOC CRM Class E82 Actor Appellation and linked to the corresponding element in the CIDOC CRM XML-encoded structure. To be able to do this one generally needs to go through the TEI elements and define a correspondence between the TEI elements and the classes and properties of the conceptual models. Then a CIDOC CRM compliant structure can be extracted semi-automatically from a properly marked-up text. This is similar to the work in the Norwegian Documentation/Museum Project described above, where the complete archaeological database was extracted in a corresponding way from densely tagged archaeological yearbooks and acquisition catalogues.

The TEI XML schema is defined in such a way that one can “plug in” other XML schemata by the use of external name spaces. The archival standard EAD is defined by a XML schema. The conceptual models CIDOC CRM and FRBRoo are defined in an object-oriented formalism, but can alternatively be defined by XML schemata. Thus information compliant with the conceptual models EAD, CIDOC CRM and FRBRoo can easily be stored in an extended TEI header by the use the name-space technique. Alternatively the real world or ontological information can be stored in separate XML-documents and packed with the TEI-encoded texts into a whole by the use of METS (METS 2008).

We may divide these methods into three different strategies for integrating the TEI-encoded text with a conceptual model:

1. 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
2. Store the information in the TEI-header using an external XML name-space, e.g. RDF or CRM-Core
3. Store the information in the TEI-header using the existing elements in TEI P5.

Depending on the work at hand, all of these three methods may be useful. The objective of our limited/little investigation is, however, to study the expressive force of the ontological

elements introduced in recently published TEI P5. Thus, the solution we discuss in this article is number 3.

In the rest of this article we will study the ontological elements in the TEI P5 compared with the CIDOC CRM. We have been working with the CIDOC CRM for many years and in our opinion it is the best conceptual model or ontology for the cultural heritage sector. Thus, it is a good tool for such a comparison.

Both TEI P5 and CIDOC CRM are the result of ongoing community based research. CIDOC CRM was accepted as an ISO standard in 2006 with possible amendments in 2009. The TEI guidelines are a set of recommendations and not a fixed standard, and details may be changed more easily. This study is based on TEI P5 as it was in November 2008. In this paper we have focused on the basic ontological parts and not on the details.

TEI P5 ontology elements in the light of CIDOC CRM

In TEI P5 most of the ontologically oriented elements are defined in the modules described in chapter 2, *The TEI Header*, in chapter 10, *Manuscript Description*, and in chapter 13, *Names, Dates, People*. Most of the new elements stem from the work with the encoding of manuscripts and manuscript catalogues (e.g. the Master project). In the TEI P5 the existing ontological elements from P4 are explained in a more systematic way together with the new elements. Even though the TEI consortium presents itself as “maintaining a *standard* for the representation of texts in digital form” the TEI P5 is meant to be a set of recommendations and not a fixed standard with respect to the use of the elements. The set of ontological elements in chapter 13 is not intended to be a (part of a) definition of a formal ontology. Other parts, especially the elements for bibliographical information found in the TEI header, are very closely related to the conceptual model underlying current library cataloguing standards. The implicit model in the TEI header and the one described in chapter 13 overlap. For example, `publPlace` mark up the name of the place for publication of a book, while `placeName` can contain any place name. A `publicationStmt` element contains information about an event, the publication, while `event` is in chapter 13, but is not intended to be used in connection with bibliographical information.

The elements in chapter 13 are the most general ontological elements. In the following these elements will be compared with the corresponding elements in the CIDOC CRM. In Figure 4 the main elements in chapter 13 of the TEI P5 are listed with the closest corresponding element in the CIDOC CRM. There is not always a complete match as the scope notes indicates.

TEI	CIDOC CRM
	E39 Actor This class comprises people, either individually or in groups, who have the potential to perform intentional actions for which they can be held responsible
<person> provides information about an identifiable individual, for example a participant in a language interaction, or a person referred to in a historical source	E21 Person This class is a subclass of E39 Actor and comprises real persons who live or are assumed to have lived. Legendary figures who may have existed [...] fall into this class if the documentation refers to them as historical figures.

<org> (organization) provides information about an identifiable organization such as a business, a tribe, or any other grouping of people	E74 Group This class is a subclass of E39 Actor and comprises any gatherings or organizations of two or more people that act collectively or in a similar way due to any form of unifying relationship.
<place> contains data about a geographic location	E53 Place This class comprises sizes/measures/extents in space, in particular on the surface of the earth, in the pure sense of physics: independent from temporal phenomena and matter. The instances are usually determined by reference to the position of “immobile” objects such as buildings, cities, mountains [...] and may be identified by one or more instances of E44 Place Appellation.
<event> contains data relating to any kind of significant event associated with a person, place, or organization.	E5 Event This class comprises changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological or legal phenomena.
<relation> (relationship) describes any kind of relationship or linkage amongst a specified group of participants	Property The properties of CRM serve to define a relationship of a specific kind between two classes. A property is defined with reference to both its domain and range.
<name> (name, proper noun) contains a proper noun or noun phrase	E41 Appellation This class comprises all sequences of signs of any nature, either meaningful or not, that are used or can be used to refer to and identify a specific instance of some class within a certain context.

Figure 4: Some central ontological elements in the TEI P5 and corresponding classes and mechanisms in the CIDOC CRM.

In TEI P5 the core ontological elements are *place*, *person* and *org*. To each element there is a corresponding name element labelled *placeName*, *personName* and *orgName*, respectively, which can be used to mark up strings in the running text denoting the real world objects. There is a well-developed specialization hierarchy of names with respect to the type of real world object they denote: (geopolitical) block, country, region, settlement. The element *country* is equivalent to ‘<placeName type=”country”>’, *settlement* to ‘<placeName type=”settlement”>’, etc. The elements are meant to be used to mark up strings used denoting a country, a settlement and so on. Similar hierarchies exist for *personName* and *orgName*. On the top of the hierarchies is the general *name* element. This hierarchy is handy and, at the first glance, intuitive. It constitutes, however, a fragment of a specific conceptual model implicitly underlying this part of the TEI P5. In ordinary language the term ‘country’ has several meanings. In this setting it can denote: an organization (in the political science sense), a physical territory or a place. The CIDOC CRM makes a clear distinction between a place and what is located or connected to it. The model underlying TEI P5 is unclear and underspecified at this point.

The two core ontological elements *person* and *org* represent individual humans and groups/legal bodies, respectively. The TEI has corresponding name elements, but no detailed hierarchy as is the case for geographically oriented names. The distinction between individuals and groups such as legal bodies is commonly used, but TEI does not have a super element “actor”. Without this general actor element, one always either has to specify a group or an individual.

In conceptual models the meaning is usually expressed by properties or relations between instances of the classes (e.g. *person*, *events* and *place*). The wedding example in Figure 3 illustrates this. The *relation* element in TEI P5 is an open schema for encoding all relations between instances of *place*, *person* and *org*. Relations can be established on the basis of a reading of a text. However, the relations are rarely identified by a name or a string in the text. Therefore a *relationName* element is unnecessary and is not included in the TEI P5.

TEI P5 uses the type attribute in *relationGrp* to group relations. The name attribute of the *relation* element is to identify the kind of relation (*spouse*, *best man*, *employed by*), as shown in the TEI P5 example reproduced in Figure 5. The values of the name attributes identify the meaning of the relation. In general the values of the name attribute should be taken from the fixed set of the relation names of the underlying ontology, e.g. the CIDOC CRM..

```
<listPerson>
<person xml:id="p1">
<!-- data about person p1 -->
</person>
<!-- more person elements here -->
</listPerson>
<relationGrp type="personal">
<relation name="parent" active="#p1 #p2" passive="#p3 #p4"/>
<relation name="spouse" mutual="#p1 #p2"/>
</relationGrp>
<relationGrp type="social">
<relation name="employer" active="#p1" passive="#p3 #p5 #p6 #p7"/>
</relationGrp>
```

The persons with identifiers p1 and p2 are the parents of p3 and p4; they are also married to each other; p1 is the employer of p3, p5, p6, and p7.

Figure 5: Example from TEI P5 guidelines for *relation* and *relationGrp*

In TEI P5 the events are seen as nameless secondary ontological elements. This is surprising. Firstly, events are often identified by an appellation. Strings like “the Second World War” or “Digital Humanities 2008” clearly denotes events. The introduction of an *eventName* element can easily be defended although it is not strictly necessary since the general *name* element with a proper type can be used instead. Secondly, events will very often be the cause why relations between persons are established. As previously mentioned, the wedding is another example. The *best man* relation is relative to a wedding as is the *spouse* relation.

In TEI P5 the use of the *relation* element is restricted to relations between persons, actors and places. This restriction should be weakened so that the *relation* element can be used to encode relations between places, actors (persons and organizations) and events. As the wedding example in Figure 3 illustrates, we would then need both a *type* attribute and a *name* attribute in the *relation* elements. The value of the name attribute indicates the category of relation, eg. “participation” and the value of the type attribute the finer sub categorisation, eg. “*spouse*”, “*bride*”. The values of the type attribute should be taken from a well-defined thesaurus in an external name space or defined by the use of the *taxonomy* element in the TEI header of the document

One may however argue that the connection between the relation and an event can be deduced by the fact that the name elements (*persName*) denoting the related persons occur textually inside the corresponding *Event* element. That is, if the names of the groom and the best man occur inside a text embraced by event tags and describing the wedding, then one can clearly deduce the relation (best man of, spouse) between the persons is directly linked to the event. However, in most real texts such a neat textual embedding will not be the case and the connection has to be stated explicitly.

Classification and type systems in the TEI

In the previous sections we discussed the ontological core classes: places, actors (person and org in the TEI), objects, abstracts, events, appellations (names) and the generalized relation schema. In a conceptual model one also needs mechanisms to describe and classify the instance of the classes. In general descriptions are given as free-text notes. Classification is usually done by selecting terms from a closed vocabulary or taxonomy expressed as a thesaurus. In the CIDOC CRM an instance of a class can be given a description (a string) via the typed *P3 has note* property and a type via the *P2 has type* property. The CIDOC CRM is a core ontology and does not include any elaborate classification systems. The class *E55 Type* is meant to be a hook for the actual classification systems.

TEI P5 has a surprisingly rich set of elements and attributes that can be used to describe and classify real world entities of the “classes” *person*, *org* and *place*, see the table in Figure 6. In addition to the common elements like *sex* and *education*, one finds the two general classification elements *trait* and *state*. In the TEI P5 module system the other elements are grouped as trait-like or state-like properties. It is not evident that this is a useful distinction. The two elements *trait* and *state* can be difficult to use in a meaningful way. It may also be so that the distinction between ‘trait’ and ‘state’ is language-dependent. One may also ask if it is necessary to have such a large collection of very specific elements for the characterization of persons and places. However, many of these elements have been added during the long history of TEI for specific purposes including, for example, the description of the social background of speakers in transcribed oral material. Thus the set of elements is ad hoc and does not represent any underlying conceptual model. A simple and sound clean-up will be to extend the scope of the *desc* (description) element to be a general description of the denotation of its parent element and not only a description of the use of the element as it is today. In addition the *type* attribute should be used to specify the type of description, that is ‘<desc type=”age”>’ should be equivalent to ‘<age>’ and so on.

Most of the elements in Figure 6 can be given attributes for temporal information: *period*, *when*, *notBefore*, *notAfter*, *from*, *to* and editorial information: *cert*, *resp*, *evidence*, *source*, *precision*. The “editorial attributes” are used for information on a meta-level and thus not a part of the real world description. The “temporal” attributes are more problematic. Firstly, it is

hardly useful to equip an *age* element with an attribute *notBefore*. We believe such peculiarities stem from the wish to make a compact definition of TEI P5. Secondly, the use of the temporal attributes transforms a property into a temporal entity or a state. Thirdly, the temporal attributes may hide facts that should be explicit. It is a very powerful mechanism for expressing synoptic information based on extensive hidden scholarly investigation about real world events. As long as the justification for the values in these elements is not present, however, it is hard to map this information onto an event centric conceptual model like the CIDOC CRM or to exchange information with any system based on other semantic contractions.

<**state**> contains a description of some status or quality attributed to a person, place, or organization at a some specific point in time.

<**trait**> contains a description of some culturally-determined, and in principle unchanging, characteristics attributed to a person or place .

	Person	Org(anization)	Place
Trait-like	age faith langKnowledge nationality sex socecStatus trait		climate location population terrain trait
State-like	affiliation education floruit occupation persName residence state	state	bloc country district geogName placeName region settlement state

Figure 6: Some of the description and classification element in the TEI P5.

The TEI P5 currently has no element sets for general classification. As for other topics like events and relations, the TEI has the necessary elements for classification according to a formal taxonomy, but they are confined to bibliographical classification, as stated in the scope notes shown in Figure 7. A simple extension of the scope of the elements *classDecl*, *taxonomy* and *catRef* to all kinds of classifications will give TEI a general mechanism for classification according to formal taxonomies.

Objects and abstracts

The CIDOC CRM, as an original model for museums, has a well-developed hierarchy of classes for physical things and objects. In addition the CRM has a class hierarchy for abstracts

(conceptual objects). Examples are the motif of a painting or the content of a text. This hierarchy has been considerably extended as a result of the harmonization with FRBR. FRBR has a four-level, hierarchical work expression, manifestation and item. The first three represent abstracts and the fourth, physical objects (e.g. books). In the FRBR_{oo} the work, expression and manifestation hierarchy is plugged in as subclasses in the conceptual objects hierarchy of CIDOC CRM, as an item in the physical object hierarchy.

The TEI P5 has no general elements for encoding information about concrete and abstract objects. It has, however, two specialized element sets, one for bibliographical information and one for information about manuscripts as physical objects. The *bibl* element can contain information about manifestations and expressions in the FRBR sense. The *msDesc* and *objectDesc* elements can contain information about a physical manuscript: size, material, owner, and so on. The elements *msDesc* and *objectDesc* do not correspond to item in the FRBR or to a subclass of physical objects in the CIDOC CRM, but are descriptions or classifications of such entities. If the scope of desc is extended the '<msDesc>' will be equal to '<desc type="ms">'. A necessary extension to the TEI will be to introduce the new elements *conceptualObject* and *physicalObject*.

<classDecl> (classification declarations) contains one or more taxonomies defining any classificatory codes used elsewhere in the text.

<taxonomy> defines a typology used to classify texts either implicitly, by means of a bibliographic citation, or explicitly by a structured taxonomy

<catRef/> (category reference) specifies one or more defined categories within some taxonomy or text typology

Figure 7: The elements available for encoding taxonomies.

Conclusions

In this article we have studied the expressive power of the real world descriptions TEI P5 by mapping central parts of the CIDOC CRM onto TEI P5. It is clear that the TEI P5 has moved a great step in the direction of an event-oriented model compared with TEI P4. Our use of CIDOC CRM as a yardstick has shown that the expressiveness can be greatly improved by extending the scope of very restricted elements like the *relation* element. On the other hand it is also clear that some parts of the real world description in TEI, like the *persTraitLike* and *persStateLike* modules, represent an ad hoc selection of types of real world descriptors.

In (Ore & Eide 2006) we suggested that the real world information should be stored in a separate formalism outside or within the TEI header. We still believe this is the best solution in many cases. However, the result of this limited study ensure that with the extensions and adjustments suggested in this article and summarized below, it will be possible to express information that confirms with the CIDOC CRM by the use of the tag sets in TEI P5. Thus the information can be stored as an integrated part of the TEI document as well.

Suggested extensions and adjustments:

- Introduce an element *conceptualObject* for conceptual/abstract objects.
- Introduce an element *physicalObject* for physical objects.

- Extend the scope of *relation* to the object elements and to the event element and add a type attribute.
- Extend the scope of *taxonomy* to non-textual entities
- Extend the scope of *desc* to all ontological elements and let desc be a super element of the classification elements, e.g. ‘<age>’ will be equal to ‘<desc type=”age”>’
- Consider to state explicitly other equivalences like ‘<publisher>’ and ‘<name type=”publisher”>’

To continue this research, an extended TEI tagset should be developed with elements for abstracts corresponding to the ones in FRBR and CRM. This will not change the ontological structure of TEI significantly. However, these adjustments will make the ontological information in a TEI document compliant with the other cultural heritage models including, for example, EAD, FRBR/FRBR_{oo}, CIDOC CRM, MUSEUMDAT and CDWA-Lite. It is important that TEI is a part in the ongoing harmonization process between all these initiatives, and thus will enable wide information integration and exchange beyond TEI encoded data

Acknowledgments

We would like to thank our colleagues at the Unit for Digital Documentation and in the Museum and Documentation Project for the many interesting discussions about data integration and text encoding over the years. The working group meeting in the CRM SIG, the FRBR_{oo} and the meetings in TEI ontology SIG and the many good colleagues in the digital humanities have also been important sources of inspiration.

References

Crescioli, M., D'Andrea, A., and Niccolucci, F. (2002), : *XML Encoding of Archaeological Unstructured Data*. Pp. 267-275 in: *Archaeological Informatics: Pushing the Envelope*. Proceedings of CAA 2001. Oxford.

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff M. (eds.) (2008): *Definition of the CIDOC Conceptual Reference Model*. URL:
http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.5.pdf (checked 2008-09-27)

CDWA Lite www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html
 (checked 2008-09-27)

Doerr, M., LeBoeuf, P. (2007) *Modelling Intellectual Processes: The FRBR - CRM Harmonization in Digital Libraries: Research and Development*, LNSC 4877 Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-77087-9

EAD *Encoding Archival Description*, <http://www.loc.gov/ead/>, (Checked 2008-09-27)

FRBR (1998). *Functional Requirement for Bibliographic Records*. Final Report. International Federation of Library Associations. URL: <http://www.ifla.org/VII/s13/frbr/frbr.pdf> (checked 2008-09-27)

FRBR_{oo}, *FRBR object-oriented definition and mapping to FRBR_{ER}* (version 0.9 draft)
http://www.ifla.org/VII/s13/wgfrbr/FRBRoo_V9.1_PR.pdf (checked 2008-09-27)

- FRBR Review Group** (2008), Working Group on FRBR/CRM Dialogue
http://www.ifla.org/VII/s13/wgfrbr/FRBR-CRMdialogue_wg.htm (checked 2008-09-27)
- Holmen J., Ore C-E and Eide Ø.** (2004), *Documenting two histories at once*, Magistrat der Stadt Wien, Referat Kulturelles Erbe, Stadtarchäologie Wien (ed) *The E-way into the Four Dimensions of Cultural Heritage*, Procs. CAA2003, Bar International Series 1227, 2004, Oxford, UK
- Holmen J. and Uleberg E.**, (1996) *Getting the most out of it - SGML-encoding of archaeological texts*. Paper at the IAAC'96 Iasi: Romania.
http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html (checked 2008-09-27).
- MASTER** (2001). "Manuscript Access through Standards for Electronic Records (MASTER)." Cover Pages: Technology Reports. URL: <http://xml.coverpages.org/master.html> (checked 2008-09-28)
- METS** (2008), *Metadata Encoding & Transmission Standard*
<http://www.loc.gov/standards/mets/> (checked 2008-12-27)
- Meckseper, C. and Warwick, C.L.H.**,(2003). The Publication of Archaeological Excavation Reports Using XML. *Literary and Linguistic Computing* 18(1), 63-75. ISSN: 0268-1145
- MUSEUMDAT** www.museumdat.org/, (checked 2008-09-27)
- Ore, C.E.** (1998) *Making multidisciplinary resources*, in: *The Digital Demotic, A Selection of Papers from Digital Resources in the Humanities 1997*, ed. by L. Burnard, M. Deegan and H. Short. (Office for Humanities Communication, King's College, London, Publication 10, 1998
- Ore, C.E., and Eide, E.** (2006), *TEI, CIDOC CRM and a Possible Interface between the Two*. P. 62-65 in *Digital Humanities 2006. Conference Abstracts*. Paris, 2006.
- RDF** (2008), Resource Description Framework, <http://www.w3.org/RDF/>
- Schloen, D.**(2001), *Archaeological Data Models and Web Publication Using XML*. *Computers and the Humanities* 35:123-152, 2001.
- Scollar, I.** (1999), *Twenty-five years of Computer Applications to Archaeology*, in *Archaeology in the Age of the Internet*. BAR International Series 750, 1999.
- Sinclair, P., Addis, M, Choi, F., Doerr, M., Lewis, P. and Martinez, K.** (2006), *The use of CRM Core in Multimedia Annotation in Proceedings of (the) First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006)*. URL:
<http://cidoc.ics.forth.gr/docs/paper16.pdf> (checked 2008-09-27)
- TEI P5** (2008). Guidelines for Electronic Text Encoding and Interchange. URL:
<http://www.tei-c.org/Guidelines/P5/> (checked 2008-09-27)
- Theodoridou, M. and Doerr, M** (2001). *Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM*, Technical Report FORTH-ICS/TR-289. URL:
<http://cidoc.ics.forth.gr/docs/ead.pdf> (checked 2008-09-27).