From XML-tagged Acquisition Catalogues to an Event-based Relational Database

Ellen Jordal, Jon Holmen, Stein A. Olsen and Christian-Emil Ore

The Museum Project for Norwegian Universities,
University of Oslo, Norway
{jon.holmen, e.k.a.jordal, s.a.olsen, c.e.s.ore}@muspro.uio.no

Abstract. We have worked with electronic preservation of old museum catalogues for more than ten years using SGML and later XML to give the text a semantic content markup. The question is whether these tools are suitable for converting the content of the catalogues into our newly developed CRM-based data model. In this article we outline the model and how we map the information in full text museum catalogues into a database according to this model.

1. Background

Old reports, acquisition catalogues and grey documents form an important source of information about the museum collections and document the work of scholars in the field and at the museums. The Museum Project for the Norwegian Universities has been working on the digitization of this kind of material since 1992 with the aim of creating a shared information bank for the Norwegian University Museums In this work our main philosophy is to consider the old reports and catalogues as first class in the information system. Thus the reports and catalogues have been converted into electronic text and in addition to structural markup been given an extensive semantic content SGML/XML-markup. See Holmen, Ore, and Eide 2004 for a more detailed presentation of of the need to include the information and content found in older archaeological and cultural historical oriented documents into archaeological and cultural heritage systems.of the In this article we focus mostly on the information extraction

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) and been suggested independently. Schloen 2001 suggests an XML formalism for storing and interchanging archaeological information. Meckseper 2001 describes the situation in the UK and points out the usefulness of XML.

To find and mark up information in a text is similar to do an excavation. In both case it is important to have an initial idea of what one is looking for, although this is usually changed several times during the work. However a text can be "excavated" many times.

We first designed a simple event based data model for the "world" described in the catalogues and reports. The DTD (Document Type Definition) was based on an analysis of both the structure, the content of the text and the data model The first DTDs had rather strict content models (right hand sides). The reports and catalogues span, however, more than 200 years. Thus the content models had to be weakened to allow for the great variety in the structure of the documents. The DTDs used for the archaeological reports and catalogues, this was achieved by writing complex alternative content models and then joining them by the use of the '|' (or) operator. We created these complex models because we believed that the order of the elements in the running text and their relative position in a strict hierarchical model could be used to extract important information .

The lessons learned from the work especially with the catalogues and reports from the Ethnographical museum, are that the elements relative hierarchical position is important locally in the text but the order inside paragraphs are not important. Thus the final DTD for the Ethnographical museum has almost no restrictions on the sequence of the elements in the content model. If the elements (tags) represents a more refined ontology (data model), say the CRM, our current believe is that the markup is almost reduced to milestone elements with very simple DTD, (D'Andrea, Holmen, Niccolucci, Ore and Ryan 2004).

On the basis of the markup we have designed several mappings into various database structures. In this article we will discuss how the markup can be used to extract information from the marked-up text into an event-based CRM-compatible database. As a case study we use an example from the acquisition catalogues at the Ethnographical Museum in Oslo.

2. An Event-based Model for Museum Material

Generally speaking a (scientific) model is a formal structure giving a schematic and simplified description of certain phenomena. Our original model was developed in the first half of the 1990s and inspired by the work of Lene Rold and her colleagues at the National Museum of Denmark (Roll 1992). The model used then, was event oriented but rather specialised (see Holmen and Uleberg 1996; Holmen and Ore 1996) The SGML/XML mark-up schema for the encoding of the text was based on this model, see figure 6 for an example from the encoded texts.

A couple of years ago we decided to find a shared model for archaeology, ethnography and natural history. In our museum model it was necessary for us to be able to store information about physical objects and their contexts: Whether objects are man-made, such as a stone axe, or created naturally, such as

flowers, they share several features: They exist in nature or are used within a culture in a certain area within a specific period of time. Cultural objects are produced, they are used and then perhaps they are destroyed. Natural objects are born, they "live" and they die. As museum objects they have been collected by somebody, at a place within a specific period of time using known or unknown methods. At one point in time the objects arrive at the museum where they go through preservation, classification and measurements before they become part of an exhibition or they are just stored in a museum magazine.

The CIDOC Conceptual Reference Model (CRM) is the result of more than 10 years of interdisciplinary work. The CRM is developed with the aim of creating a model to express, store and exchange information concerning museum objects and their context. The CRM is event oriented and the events are used in order to store information about who did what where and when, see figure 1.

Thus the CRM was a good candidate for a museum model. However, we decided first to develop an event-oriented model based on the data found in the documents and then consider if it was possible to express or map this model into the CRM. This was also meant to be a test of the CRM and the test was successful.

The model is designed for preserving the "full" story in terms of objects, relations, activities and events for all museum objects. The main entities have been sub-typed in order to refine the model in accordance with the data that has been entered into it. All entities and relations in the model have a CRM equivalent as shown in parenthesis. This is also the case with the subtypes.

2.1 Identification of Information Pieces in a Text

It is hard if not impossible to find the exact meaning expressed by a piece of text. In this article we do not pretend to give a contribution to the study of meaning as such. Our aim has been to give a solution to a practical problem: How to extract, in a scientific and on a later point in time reproducible way, information from catalogues and reports into a museum database. The last 12 years work with the conversion of the acquisition catalogues and reports has shown us that it is not so difficult to identify pieces of information in a text in accordance with an existing model or ontology. That is, as long as one has defined a model or "toy world ", it is possible to instruct non-specialists to identify information elements in a text in accordance with this model. However, the model should not be too complex. In the beginning of the 1990s we developed a pretty specialized, event oriented model which then was extended to the present general model. This model can be mapped into the CRM. As an example we use a sentence from the acquisition of the Ethnographic Museum for 1854 (translated into English): "A long Chinese pipe of bone. Donated 12th May 1853 by Kildal-Lund". The identification of information pieces in accordance with our model with CRM class numbers given in parenthesis, is shown in figure 2.

The relevant pieces of information in the sample sentence are mapped into the classes of objects, events, actors, time spans, places and relations between these.

Man made object (E22): A pipe
  Object attributes
    material: bone
    shape: long
Relation (P12->P108): was produced
Event (E5->E12): Production
  Place (E53): China
Event (E5->E10): Donated
  Time-span (E52): 12th May 1853
Relation(P28): by
Actor (E39): Kildal-Lund.

Fig. 2. Information elements in the text.

If this information came from the content of a table, one only had to do this mapping once for each information column, and extract the information by program, but since it is a running text, each sentence has to be read and information extracted manually.

It is of course interesting to try to automate the process, but we stress that in our case almost everything has been done manually.

At some (trivial) level the information in all the wide variety of catalogue texts can be identified in accordance to the model. The problem lies in the amount of work needed to identify the important information and mark up the text to enable fast references from the extracted data back into the original text.

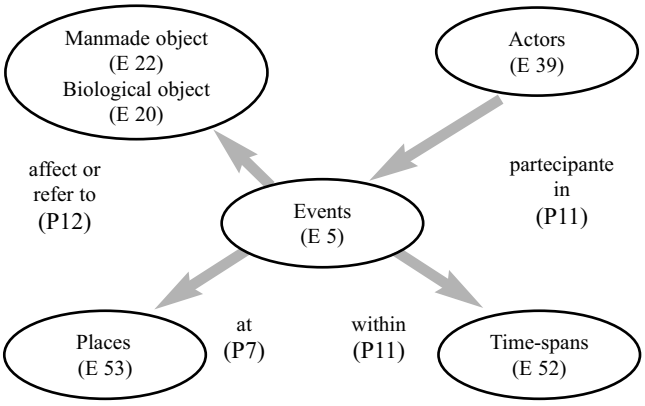

Fig. 1. An overview of our data model with CRM equivalents.

3. Using XML as Mark-up Language

Our solution is to use XML as a tool for marking up the various bits of information and then use parsing techniques to instantiate objects and events in the new model. We have developed a DTD containing elements that combined with some parsing rules maps contents of old catalogues running text into our CRM like database. It does not solve the problem of analyzing the text, but some of the mark up, all extraction and also references back to the original text can be done by programs.

Our case study is the acquisition catalogues at the Ethnographical Museum in Oslo. These catalogues have been

produced for more than 150 years and include more than 45000 catalogue numbers, and an even greater number of objects. All these objects are classified as “man made” and consequently they are the same kind of objects as the archaeological objects. The only difference is that whereas the archaeological objects were dug out of the ground and used to study the past, these objects are collected within existing societies in order to serve as samples related to contemporary human activities.

EM8.
Malayan dagger, taken from pirates of the Indian Oceans.
Beautiful handle, graven as a human figure above waistline.
Snake winded blade. VII, IX, p, 2. Daa,O., 99.
Donated April 11 1856 from Captain Teiste.

Fig. 3. Original catalogue entry dating from 1856.

<NRPAR> <CATNR NRID="EM8">EM8 </CATNR>.
<ARTIFDAT>Malayan
<ARTIFACT>dagger</ARTIFACT>,taken from pirates of the Indian Oceans.
Beautiful handle, graven as a human figure above waistline.
Snake winded blade. VII, IX, p, 2. Daa,O., 99. Donated April 11 1856 from Captain Teiste.</ARTIFDAT></NRPAR>

Fig. 4. The catalogue entry with basic mark-up.

Figure 3 shows a typical entry from the early parts of the catalogue (19th c.). As shown in figure 4 the catalogue entry is given a basic markup, that is, the start and the end of the object information, the catalogue number and the artifact type “dagger". Parsing these elements we create the following records in the database: A man made object with the catalogue number as assigned identifier and an event of type classification that assign the type “dagger” to the object with the author of the catalogue as actor of type person.

<NRPAR><CATNR NRID="EM8">EM 8</CATNR>.
<ARTIFDATA> <PROD> <USE> <PEOPLE>
<PLACE>Malayan </PLACE> </PEOPLE> </USE>
</PROD> <ARTIFACT> dagger</ARTIFACT>, taken from pirates of the Indian Oceans.
Beautiful handle, graven as a human figure above waistline.
Snake winded blade. VII, IX, p, 2. Daa,O., 99.
Donated April 11 1856 from Captain Teiste.
</ARTIFDATA> </NRPAR>.

Fig. 5. The catalogue entry with pre-acquisition information marked.

From the classification of the object as man-made, it is given that the object was brought into existence by an event of creation or production involving human activity.
The only information given in the text about this fact is the not very precise term “Malayan”. Considered out of context and stripped of other information this might mean that the dagger was owned by, used by or produced by somebody with a connection to what was considered to be the Malayan culture in the middle of the 19th century. To know this for sure is not possible. It will always be a question of interpretation. In this case the scholars at the museum interpreted the term “Malayan” to mean both the place of production an of the use of the dagger. Thus the term is enclosed in a nested sequence of tags. The ‘PROD'- and ‘USE’-elements are interpreted as two different events, a production event and a use event and give rise to two event records in the database. The actor connected to the events is of type (Malayan) people. The adjective ‘Malayan’ is also interpreted as a place or location of both events, see figure 5.

<NRPAR><CATNR NRID="EM8"> 8</CATNR>.
<ARTIFDATA><PROD><USE><PEOPLE><PLACE>
Malayan </PLACE></PEOPLE></USE></PROD>
<ARTIFACT> dagger </ARTIFACT>,
<ACQUISITION>taken from <ACQUFROM> pirates
</ACQUFROM> of the Indian Oceans. </ACQUISITION>
Beautiful handle, graven as a human figure above waistline.
Snake winded blade. VII, IX, p, 2. Daa,O., 99.
<ACQUISITION> Donated <ACQUTIME> April 11 1856
</AQUTIME> from <ACQUFROM> Captain Teiste
</ACQUFROM>.
</ACQUISITION></ARTIFDATA></NRPAR>

Fig. 6. The catalogue entry with acquisition tags.

Another aspect of this catalogue entry concerns the provenience of the dagger. There are two explicitly mentioned events. The dagger is transferred from the alleged pirates to some unknown recipient (could be Captain Teiste). The second is the donation of the dagger to the museum by Captain Teiste. It is unlikely that this is the complete history of the dagger. There must be at least one more implicit acquisition event: from the Malayan producer to some unknown recipient who of cause could be one of the (Malayan?) pirates. We don’t know the dates or time span of these events. However we know the sequence in which they have occurred and we know the date of the last event, the donation. This makes it possible for us to set an upper time boundary for all the events and then to use time arithmetic, such as Allen Operators (see Allan and Hays 1989), on the ordered sequence of the events.
The text is written in the 1850s. Today we may not have used the term ‘pirate’. Thus it is important to add a fact to the database that it is the author of the catalogue who stated the fact that the unknown sailors were pirates.

3.1 Using Attributes as Glue

The catalogues and reports are written to inform the the specialists and an interested public about new acquisitions of the museum. Even in this rather specialized genre it is not desirable or even possible to write a text in accordance with a strictly hierarchical mark-up schema. The information about an artifact or an event is often spread discontinuously in the running text. Thus one needs a mechanism for gluing the information pieces together. This is a well known problem in text encoding and is generally solved by the use attributes (see also Sperberg-McQueen and Burnard 2002).

The catalogue entry in figure 7 displays a simple example of the problem of discontinuous information. In this example the term bush indicates the place of use for the special kind of shirt. At the end of the entry the author locate the bush to be in Mainé-Soroa. In our simple mark-up model we link the two terms by the use of attributes, see figure 8. By doing so we don't necessarily postulate that the entire Mainé-Soroa consists of bush. The purpose is to make the mark-up schema less complicated.

EM 43895.
"Djampa". Blue Manga sleeveless shirt.
Used daily by most Manga peasants in the bush. Shoulder width 57 cm., length 103 cm. People Manga, Place Mainé-Soroa.

Fig. 7. A catalogue entry from 1971.

<NRPAR><CATNR NRID="EM43895"> 43895 </CATNR>.
<ARTIFDATA>"<ARTIFACT><ETERM> Djampa
</ETERM> </ARTIFACT>". <TECHNIQUE> Blue
<PROD><USE><PEOPLE> Manga
</PEOPLE></USE></PROD> sleeveless <ARTIFACT>shirt
</ARTIFACT></TECHNIQUE>.
<APPAREA> used daily by most Manga- <USE><SAGE>
peasants </SAGE> </USE> in the <USE><PLACE
plref="43895.1"> bush</PLACE></USE></APPAREA>.
<MEASURE> Shoulder width 57 cm., length 103
cm.</MEASURE> People <PEOPLE> Manga </PEOPLE>,
Place <PLACE plid="43895.1"> <UNINTERP> Mainé-
Soroa </UNINTERP> </PLACE>.</ARTIFDATA>
</NRPAR>

Fig. 8. The catalogue entry fully encoded.

There are several ways the mark-up shown in figure 8, can be interpreted when the tagged text is used to populate a database. There may be two places with an inclusion relation in either direction or one with two attributes, the name and the vegetation type. The preferred solution depends on the use of the database. However, the text is linked to the database and a user can always check the original source for the facts in the database.

3.2 Using the CRM as a Guideline for Data Capture

There are two main aspects of the use of the CRM as a guide for capturing data in the catalogues and reports: an information technical and a museum-scholarly point of view. A main feature of the CRM is the possibility to choose between different levels of generalisation/specialisation of the data. An entry in a catalogue may simply state: “Vase. Borneo” while an entry in another catalogue may give a detailed description of the production and provenience of another vase. In the first case one would model this as a single event connecting the place, Borneo, with the object, the vase. The event could be given a type say “produced_aquired”, reflecting its undefinedness. The information in the second thought entry will give rise to several well defined events, actors etc. Thus the use of the CRM as a guideline enable the mapping of several mark-up schemata into on common structure and store the information in a common fact database. From a scholarly point of view the CRM could be a guideline to good cataloguing practice. In many cases the information in the catalogues and reports is not very elaborated. A lot of the entries lacks recordings of possible geographic referable information, e.g. the bush of the Manga in Mainé-Soroa. In the texts one does normally not differ between actual use of an object and intended or typical use, while our model does. A literal interpretation of the text in the example in Fig. 7, states that this particular djampa has been used by most Mangas, which is obviously wrong. In fact it is not clear whether this particular shirt has been used at all.
The CRM is the result of more than 10 years of international interdisciplinary work. Our experience is that it is possible to create a meaningful mapping into the CRM of the information in all the museum records, catalogues and reports we have been working with in the last 14 years.

4. Conclusions

We have created a CRM compatible, event-based data model for our museum database and we are able to transfer data from old text based catalogues that are tagged with XML into objects and events in this database. The encoding process is time-consuming but can be done by non-specialists. The alternative and more widespread method where a specialist reads and key the core data into a registration form, may be faster measured by the number of pages processed per day. However, the job requires a scholar/specialist and it is very hard to do the proof reading and 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.
The two methods both rely on a human extracting from or marking up a text. In our work we have also written and used simple mark up programs. This is efficient and time saving. However, it is hard to write program that is capable to do nontrivial tagging correctly. This is perhaps more difficult than writing program for automatic topic indexing of large document collections. This is however an interesting field of research (see also Makkonen and Ahonen-Myka 2003).
By mapping the catalogue texts to our CRM-based model we make information that is hidden in the text visible and explicit. By using the model as a “template” for types of information to search for, we also detect possible improvements in practice. All in all we create a skeleton for the story we can tell about these objects.

References

Allen, F. J. and Hays, P. J., 1989. Moments and points in an interval-based temporal logic. *Computational Intelligence* (5), 225–338. Oxford UK, Blackwell.

Crescioli, M., D'Andrea, A. and Niccolucci, F. 2002. XML Encoding of Archaeological Unstructured Data. In

Archaeological Informatics: Pushing the Envelope. Proceedings of CAA 2001. Oxford. 267–275.

CIDOC-CRM: CIDOC Conceptual Reference Model. Proposed ISO 21127. URL: http://cidoc.ics.forth.gr/

D'Andrea, A., Holmen, J., Niccolucci, F., Ore, C.-E. and Ryan, N. An XML-compliant, CIDOC-CRM compatible methodology for documenting archaeological sources, Presentation on CAA2004. Prato Italy.

Holmen, J. and Ore, C.-E., 1996. New life for old reports – The Archaeological Part of the National Documentation Project of Norway. *ALLC-ACH '96*, *Book of Abstracts: Conference abstracts, posters and demonstration, No. 70.* Report Series of the Norwegian Computing Centre, Bergen.

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

Holmen, J., Ore, C.-E. and Eide, Ø., 2004. Documenting two histories at once, Magistrat der Stadt Wien, Referat Kulturelles Eerb, Stadtarchäologie Wien (ed.) *The E-way into the Four Dimensions of Cultural Heritage, Procs. CAA2003*. Bar International Series 1227, 2004, Oxford.

Makkonen, J. and Ahonen-Myka, H., 2003. *Utilizing Temporal Information in Topic Detection and Tracking, ECDL'2003*. Berlin. Springer LNCS No.

Meckseper, C., 2001. *XML and the publication of archaeological field reports.* Master's dissertation, University of Sheffield.

Rold, L., 1992. Syntheses in object oriented analyses. In Andresen, J., Madsen, T. and Scollar, I. (eds), *Computing the Past CAA92*. Aarhus, Denmark, Aarhus Univeristy Press.

Schloen, D., 2001. Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities* 35. 123–152.

Sperberg-McQueen, C. M. and Burnard, L. (eds), 2002. *TEI P4, Guidelines for Electronic Text Encoding and Interchange*, Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford, Oxford.