Was It Worth It? Experiences with a CIDOC CRM-based Database

Ellen Jordal
University of Oslo, Norway. *ellen.jordal@usit.uio.no*

Espen Uleberg
University of Oslo, Norway. *espen.uleberg@khm.uio.no*

Brit Hauge
University of Oslo, Norway. *brit.hauge@usit.uio.no*

Abstract:
The paper reflects on an implementation of an event-based database for the Ethnographic Collection at the Museum of Cultural History in Oslo, Norway. CIDOC CRM was used as a guideline for mapping three different sources into the database. The paper considers benefits of having data structured in an event-based database considering the extra work of analysing data and tags and mapping them into explicit events and relations. It gives an overview of challenges and problems encountered, and benefits from having our data structured in this model. The model allows import of rich data from different sources and keeps track of events and relations between object, people, time and place. Will this enrichment of the data enhance the quality of the browse and search functionality?

Key Words: *Event-based Modelling, CIDOC CRM, Ethnography, Rich Datasets*

Background

The database for the ethnographic collection at the Museum of Cultural History at the University of Oslo is the newest result of a grand vision from 1990. With the motto "From Drawer to Screen" the Documentation Project started to build an extensive information system on Norwegian language and culture (Hodne 1988, Ore 1988). The Norwegian university museums have had two consecutive national projects since 1991 with the aim of constructing a full database system for the collections and make them available for researchers, students and the general public. The projects are now continued in the permanent organization MUSIT (museum IT). The database system has been developed to be used for research purposes and not only for the artefact curation at the museums (Uleberg 2008).

The Documentation Project was a cooperation between the faculties of humanities at the Norwegian universities. Archaeology was one of the first sub-projects and it started with digitizing the acquisition catalogues (Holmen and Uleberg 1996a) and also reports and grey literature documents (Engevik et al. 2004). The paper based catalogue texts were first converted to computer readable formats. After transcription, the texts were SGML-tagged according to a grammar developed for these texts and expressed in the museums' tagging schemata (Holmen and Uleberg 1996b, Holmen et al. 2004). The tagged texts were then converted to databases.

In modelling the database for the museums, the projects worked towards an event-based model inspired by work done at the National Museum of Denmark (Rold 1993). Event-based

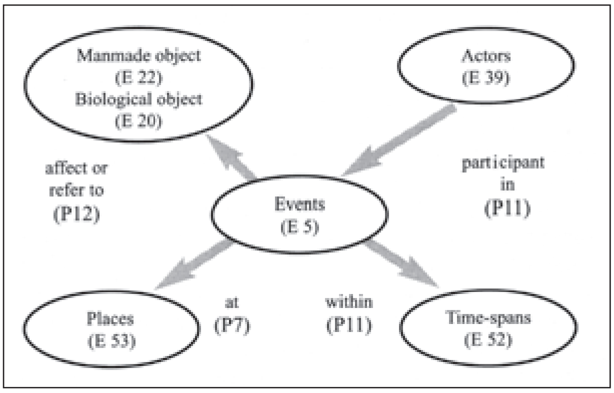


Figure 1. An overview of the data model with CIDOC CRM equivalents (after Jordal et al. 2010).

modelling was at its starting point at the time and it is safe to say that the event-based way of thinking was mainly a back-drop against which the databases were modelled. Today the database for the archaeological collections is becoming even more in line with the CIDOC Conceptual Reference Model (CIDOC CRM: www.cidoc-crm.org).

The Museum of Ethnography at the University of Oslo was included in the project at a later stage, because it belonged to the faculty of Social Sciences until 1998. This delay made the modelling of the database for the Ethnographic collection a good opportunity to implement CIDOC CRM for a much larger dataset than had hitherto been done. By 2004, an event-oriented model (Fig. 1) had first been developed based on the data in the catalogue texts and subsequently the model had been expressed in CIDOC CRM. The model was designed to preserve all objects, relations, activities and events concerning all museum objects. It was shown that it is possible to map into the CIDOC CRM all the information in the museum records in a meaningful way (Jordal et al. 2010). Further maintenance and development in the following years have given valuable experience in applying CIDOC CRM when modelling large museum databases.

Sources

The museum had three overlapping sources for the ethnographic collection that were converted into the database. The first was the (partly handwritten) acquisition catalogue which had been converted and SGML-tagged. This catalogue was maintained until 1987, when it was replaced by a FileMaker application. This database had also enhancements and corrections to information about older items in the collection. In this way some objects could have more than one description. The third source was a MS Access base for the Classical Antiquities Collection. Some of the items here have been transferred from the Ethnographic Museum. In total, the database has more than 50000 artefact entries and we made around 2 million events and 3,6 million relations between the events and the artefacts, places, actors and time. The event-based model made it possible to convert all information and to have a timeline including all versions. The transfer of objects from the Ethnographic Museum to the collection of Classical Antiquities can also be shown as one of the events in the history of the artefact.

In a museum context, acquisition, cataloguing and the handling of artefacts in connection with photography, conservation, exhibition, storing and so on are all events. In line with CIDOC CRM, each event is connected to a place and a person. Acquisition is the last of the provenance events. The provenance is divided into *produced at*, *typically used at*, *actually used at* and *acquired*. In this way it is possible to register more detailed structured information about the history of each object. The Museum of Ethnography was founded in 1857. Due to the long timespan throughout which the museum catalogues have been written and also the many people involved in collecting and registering information about the artefacts, the actual information content in the original catalogue texts varies a lot. Information that anthropologists find useful today was not always considered to be important by a diverse group of collectors that provided artefacts to the collection. Therefore it is not always possible to connect an event with a specific place and person.

Extracting information from full text into a database involves interpretation. It is therefore essential to give access to the original source, in this case, the catalogues that go back more than 150 years. The original source gives the user access to other events that are not published on the web through the database, such as the acquisition, as all artefacts in the same acquisition event are presented together in the original catalogue text.

Some of the catalogue texts are brief, while other contains more extensive information about the object. The following text will illustrate this (our translation): "these artefacts are the collection brought back from the Netchilli-eskimos by captain Roald Amundsens Gjøa expedition. King Williams Land". King Williams land is a place in Canada, but what relation is there between these artefacts, the Inuit and King Williams land? Are the artefacts produced there or used there - the text does not say.

This kind of data was first stored as events of unknown type with an unknown relation to the artefact but with the relation "located at" to the specific place. In this way the provenance could be converted into the database. This information could have been shown in the application in an extra view for the event of unknown type, but the museum decided instead to state explicitly an interpretation of such information.

The question was what is meant by stating that an object is from a certain place. The statement is not complicated in everyday language and the database user will also often wish to query the base without a more specified relation between place and object. However, the views in the application are more specific, and the relation between an object and a place should be stated as a "produced at","typically used at" or "actually used at" event. The precise meaning of the relation "being from" is not intuitively evident. Back in time, it was more likely that the place of collecting, producing and using was the same. This has changed due to globalization. Asian products can be purchased in Europe; perhaps the products are even made in Europe by Asians who grew up with the local Asian tradition of making the items. The Department of Ethnography concluded that the provenance "being from" in daily speech generally means "typically used at" the location. The database has been updated according to this and the event-based model made it possible to store the information that this was an interpretation and by whom and when it had been added.

The application is made to contain richer information, more detailed events than what you can find in many of the original catalogue texts. That is because the model shows how social anthropologists would like to describe the artefacts today and not the minimum of what should be registered for each artefact. The result is that several fields remain empty for the earlier finds in the collection. Revisions of the collection will however benefit from the possibility of storing rich information in the system. The database will also structure the cataloguing of new acquisitions as it shows what kind of structured information the museum expects to receive with new artefacts.

Challenges

Data model

The semantic standards like RDF were in their early stages (and premature) when the project started in the beginning of 2000. We decided therefore to implement our model in a relational database. One of the main challenges was how to implement this model in a way that it would be easy to access data, to maintain the content and to be able to extend the model. We decided to limit ourselves to implementing only the entities and properties from CIDOC CRM that we had data for and we kept the number of tables at a minimum; essentially one table with three columns. All data were converted into this model, but we had to conclude that the result was not optimal. It was easy to append data and to extend the model but it turned out to be difficult to access data. The system responded too slowly with this large amount of data. The relational database lost its capability to maintain constraints and therefore made it difficult to verify the content. It also became difficult to browse data. The structure could not be presented and several joins had to be made even for the easiest relations. The rich structure of the data model had been simplified into these three columns in one table, and it was necessary to rebuild the structure every time data were accessed or browsed. The lesson learnt from this is never to store triples in a relational database – unless you have an adequate browser.

In the implementation of the next version we followed the traditional way of modelling a relational database. We made more tables and a structure closer to the end user's concepts, but it should be emphasized that the event model was maintained. With this solution, it was much easier to maintain data and to optimize search queries and data browsing was fast. Our experience tells us not to be tempted to make too many tables of a complex conceptual model like CIDOC CRM. The lesson learnt here is the importance of making the right amount of tables; neither too few, as in our first attempt, nor so many that you lose track of the content.

Database interface

It is a challenge to present a rich dataset with many layers of information in a way that the end users can easily browse and edit the data. An interface to view the data is in itself a challenge and when the users are able to edit the number of fields in the interface it increases. This is due to the fact that editing information about an artefact involves editing an event or an attribute to an event. Automatic registering of actor, time and place for the data entry is not enough. The user may need to edit these fields as the event may be connected to another time and place than the actual registration in the database.

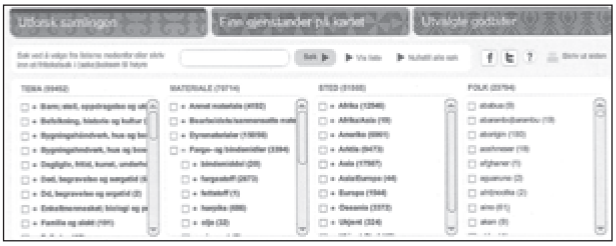
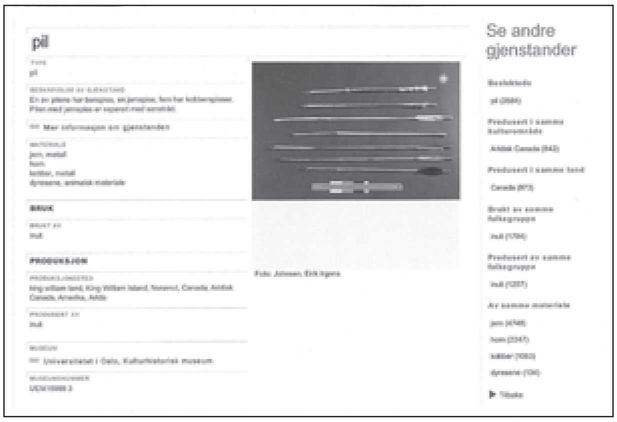


Figure 2. The ethnographic collection's webpage with facetted search.

The interface must also handle the distinction between actual activities and opinions about earlier events. The difference between these two incidents is not always clear. When a user adds information about an event in the past, it can be said that this is always an opinion about this earlier event. It is an opinion in the same way that every artefact description is an interpretation. When the system registers which user has filled in new information, it is also specified which group of fields is changed. Registering fields like artefact type, material, measures etc. are all separate events and each part of the description of the artefact has an actor connected to it.

Web interface

The Norwegian university museums through MUSIT have presented vast amounts of objects through several web pages on www.unimus.no. This newest addition (Fig. 2) presents the ethnographic collection of the Museum of Cultural History to a wide range of users. Facetted queries give users without prior knowledge of the material several inroads to the collection. The facets are Theme, Material, Place and People. Where applicable, the database has been normalized according to Outline of Cultural Materials, Outline of



pil

TYPE
pil

BESKRIVELSE AV GJENSTAND
En av pilene har benspiss, en jernspiss, fem har kobberspisser. Pilen med jernspiss er separert med senetråd.

Mer informasjon om gjenstanden

MATERIALE
jern, metall
horn
kobber, metall
dyresene, animalsk materiale

BRUK

BRUKT AV
Inuit

PRODUKSJON

PRODUKSJONSSTED
king william land, King William Island, Nunavut, Canada, Arktisk Canada, Amerika, Arktis

PRODUSERT AV
Inuit

MUSEUM
Universitetet i Oslo, Kulturhistorisk museum

MUSEUMSNUMMER
UEM15999:3

Foto: Johnsen, Eirik Irgens

Se andre gjenstander

Beslektede
pil (2684)

Produsert i samme kulturområde
Arktisk Canada (842)

Produsert i samme land
Canada (873)

Brukt av samme folkegruppe
Inuit (1794)

Produsert av samme folkegruppe
Inuit (1257)

Av samme materiale
jern (4748)
horn (2247)
kobber (1063)
dyresene (106)

▶ Tilbake

Figure 3. Page with links to other objects through the events produced at/by and used at/by.

World Cultures and authoritative lists for place names. The type assignment, the artefact term, is published along with two provenance events: the production and the typical use events.

Despite the fact that the database model is event-oriented, we decided to let the search criteria be artefact-oriented. The main reason was that we wanted to re-use the lay-out from the Archaeology Web Portal. In this way the Ethnography Web Portal will be recognized by the users as one belonging to the set of pages presented by the university museums on the unimus-website. Another reason is that the way of describing the artefacts has changed from the first items in 1857 till today. The consequence is, as mentioned above, that the rich information that the data base allows for is not recorded in many of the old entries.

Even though the initial query is artefact-based, the user can navigate on the web from one catalogue number to other items produced or used at the same place or by the same actors; it follows from the database model that people, cultures or ethnic groups, are actors in events as they are registered as users or producers of the objects. In this way the user will query the database according to events carried out at a specific place or by certain actors (Fig.3). An event-based query that would return all persons and places connected to a specific event is however not possible to do through the web pages.

Benefits

The most important benefit from the event-based model was that we could convert data from several sources with a time depth. When two sources gave diverging information both could be stored in the database and the user would get access to all the information. Implicit information was made explicit through interpretation done in the process of converting from SGML-tagged text to the defined views in the database. This rich dataset gives a better basis for telling a story for each artefact. Our database can be seen as consisting of two parts – the event-based model and the more domain based model. The domain based model describes the tables and columns that are specific for the ethnographic collection. The event-based model is generic and has already shown itself to be reusable for other domains, such as archaeology.

Conclusions

Was it worth it, and what is the impact of the event-based model on the web-pages? Whether it has been worth all the efforts or not, clearly depends on how the system is used. It has the capabilities of an object management system with information of the object and its provenance, but it is also a lot more. This is an information system that provides a possibility to do research on the objects with as complete, accurate and rich data as possible. The ontology, mapped data from the old catalogues and databases are available for exploration by the museum conservators at any time. In addition, it will guide and structure the registration of new items in the collection.

The underlying model is fundamental to how the material can be presented on our own internet pages and to how it can interconnect with other web pages. The possible impact of the event-based model on the web-publication will be further explored and developed in the future. As of now, events are shown, but it is not possible to query an event specifically. We have worked with mapping of both the non-event-based archaeology database and the event-based ethnography database. Our experience from mapping into Museumdat and Dublin Core, indicate that the event-based model is an asset when it comes to mapping to other formats, such as for example RDF.

So, was it worth it? The answer is clearly YES.

Bibliography

Engevik, A., Holmen, J., Innselset, S., and Stabell, J. 2004. "Digital Archaeological resources at the University of Bergen: An efficient Tool in Research and Heritage Management." In *Making the connection to the Past CAA 99.Computer Applications and Quantitative Methods in Archaeology: Proceedings of the 27th Conference. Dublin, April 1999*, edited by K. Fennema and H. Kamermans, 43-48. Haveka BV.

Hodne, B. 1998. "Dokumentasjonsprosjektet: historikk, målsetting og utfordringer." In *Fra skuff til skjerm: Om universitetenes databaser for språk og kultur*, edited by K. Aukrust, and B. Hodne, 13-18. Universitetsforlaget.

Holmen, J., Ore, C. E., and Eide, Ø. 2004. "Documenting two histories at once: Digging into archaeology." In *Enter the Past, the E-way into Four Dimensions of Cultural Heritage. Computer Applications and Quantitative Methods in Archaeology 2003*, edited by K. F. Ausserer, W. Börner, M. Goriany, and L . Karlhuber-Vöckl, 221–224. Oxford: Archaeopress.

Holmen, J., and Uleberg, E. 1996a. "The National Documentation Project of Norway - the Archaeological sub project." In *Proceedings of the 23th CAA conference held at Leiden University, Leiden, 1995*, edited by H. Kamermans, and K. Fennema. Leiden: Leiden University Press.

Holmen, J., and Uleberg E. 1996b. "Getting the most out of it: SGML encoding of archaeological texts." Paper presented at *the IAAC'96, Iasi, Romania.* http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html

Jordal, E., Holmen, J., Olsen, S. A., and Ore, C. E. 2010. "From XML-tagged Acquisition Catalogues to an Event-based Relational Database." In *Beyond the Artifact. Digital Interpretation of the Past. Proceedings of CAA2004 Prato 13-17 April 2004*, edited by F. Niccolucci, and S. Hermon. 81-85.

Ore, C. E. 1998. "Hvordan lage databaser for språk og kulturfag." In *Fra skuff til skjerm: Om universitetenes databaser for språk og kultur*, edited by Knut Aukrust, Bjarne Hodne. 29-57. Universitetsforlaget.

Rold, L. 1993. "Syntheses in object oriented analysis." In *Proceedings of the 20th CAA conference held at Aarhus University, Aarhus, Denmark, 27–29 March 1992*, edited by J. Andresen, T. Madsen, and I. Scollar, 213-220.

Uleberg, E. 2008. "Scale Dependent Patterns in Large Museum Datasets." In *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007*, edited by A. Posluschny, K. Lambers, and I. Herzog, 266-271.