Abstract

Ontologies play an important part in the development of the future 'semantic web'; the CIDOC conceptual reference model (CRM) is an ontology aimed at the cultural heritage domain. This paper describes a Concept Browser, developed for the EU/IST-funded SCULPTEUR project (semantic and content-based multimedia exploitation for European benefit environment (programme IST-2001-no. 35372); May 2002 to May 2005), which is able to access different museum information systems through a common ontology, the CRM. The development of this Concept Browser has required mappings from the legacy museum database systems to the CRM. The crucial process of creating the mappings is described, using the C2RMF catalogue (EROS) and library databases as a case study.

Keywords

ontology, CIDOC CRM, Concept Browser, SCULPTEUR, UNIMARC, EROS

Using an ontology for interoperability and browsing of museum, library and archive information

Patrick Le Bœuf*
Bibliothèque nationale de France
DSR/ABN/SCO
Bureau de normalisation documentaire
Quai François Mauriac
75706 Paris cedex 13
France
E-mail: patrick.le-boeuf@bnf.fr

Patrick Sinclair, Kirk Martinez and Paul Lewis
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ
UK
E-mail: km@ecs.soton.ac.uk; phl@ecs.soton.ac.uk; pass99r@ecs.soton.ac.uk

Geneviève Aitken and Christian Lahanier
Centre de recherche et de restauration des musées de France
Palais du Louvre
Porte des Lions
14, quai François Mitterrand
75001 Paris cedex 01
France
E-mail: genevieve.aitken@culture.fr; lahanier.christian@culture.fr

*Author to whom correspondence should be addressed

Introduction

An ontology is a shared conceptualization of a particular domain. It serves to structure the concepts, relations and instances of the concepts associated with the domain.Ontologies play an important part in the development of the future 'semantic web' (World Wide Web Consortium, Semantic Web Activity, http://www.w3.org/2001/sw/). Their use as a way of mapping from legacy database fields to a more commonly accepted naming convention provides a useful tool for interoperability between diverse collections as well as browsing them. This paper describes a Concept Browser which has been developed specifically for accessing museum collections. The mapping process is one of the areas that requires considerable effort and this is also described.

The CIDOC CRM

The CIDOC CRM (conceptual reference model) (Heraklion, Greece: Institute of Computer Science, Foundation for Research and Technology, http://cidoc.ics.forth.gr/) is an ontology based on the CIDOC Information Categories (CIDOC 1995). It has been under development since 1996 and is currently being agreed as an ISO standard. It has been described by its originators either as an 'electronic Esperanto' (Doerr and Crofts 1999) or as a 'semantic glue' (CIDOC CRM website: http://zeus.ics.forth.gr/cidoc/docs/dc_to_crm_mapping.pdf; also available in rtf format: http://zeus.ics.forth.gr/cidoc/docs/dc_to_crm_mapping.rtf): both phrases indicate its role as a mediator between systems that are supposed (or that may happen) to be incompatible. The CIDOC CRM can be used as the basis for data exchange between systems, as a reference guide for the design of new cultural heritage information systems, and as the basis for integrated query tools and mediation systems' data schemas (Crofts et al. 2003).

The Concept Browser

One of the aims of the European Union funded SCULPTEUR project is to devise an ontology-based system to query several museum databases built with different architectures, containing different metadata and developed with several

computer operating systems (Goodall et al. 2004).

The Concept Browser, developed by the University of Southampton, UK, is able to display the CIDOC CRM ontology in a graphical way, using a graph-based approach for the visualization (Figure 1). Owing to the complexity of the ontology, this view is generally hidden from the user. A simplification of the ontology is displayed instead, which will only display the concepts and relations that are present in the museum metadata structure. These concepts and relations are further refined and simplified, in some cases using terms from the original metadata schema to increase familiarity of the users with the interface.

The graph-based visualization is based on TouchGraph (TouchGraph LLC,

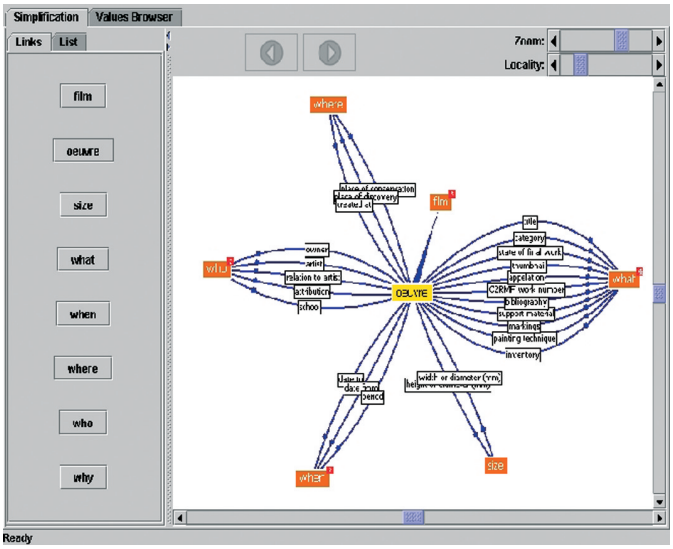

Figure 1. Concept Browser

http://www.touchgraph.com), an open source package of Java classes which provides a way of drawing a graph and interacting with it.

An important aspect of ontological visualization tools is querying for instances of concepts. Although the visualization of instance information within a graph-based interface has been investigated, complex visualization challenges must be overcome to cope with the scale of the datasets present in the SCULPTEUR project. For example, the C2RMF EROS database (Lahanier et al. 2002) contains information on many tens of thousands of objects; trying to display even a subset of this data in a graph-based visualization will typically result in a confusing and messy display for the user.

Instead, the Concept Browser has based instance visualization and query on mSpace interfaces (http://mspace.ecs.soton.ac.uk, Schraefel et al. 2003). mSpace is an interaction model designed to allow a user to navigate in a meaningful manner the multi-dimensional space that an ontology can provide.

mSpace interfaces are based on slices through an ontological space, with each slice represented as a list of values; slices are presented as columns arranged from left to right (Figure 2). Selection in a slice will update the display so that the values displayed in the next slice (that is, to the right of the current slice) are related to that value. For example, if there is a slice of artists and the next slice is painting titles, selecting an artist will display only that artist's paintings in the titles slice. Values in each slice are filtered, so that there are always results to view in the next column when a selection is made. When an item is chosen in a slice, details about that item are displayed in a detail panel; if no details are available for that item, examples of related objects are shown. Slices can be freely interchanged, removed and new slices can be added to the mSpace.

The museum metadata being dealt with in Sculpteur is large and varied, so there are many possible slices as well as combinations of slices that users may be interested in. The ontology simplification interface, based on TouchGraph, allows users to browse and add the slices in which they are interested into the mSpace browser, where they can be arranged to suit the user's preference. A preview panel

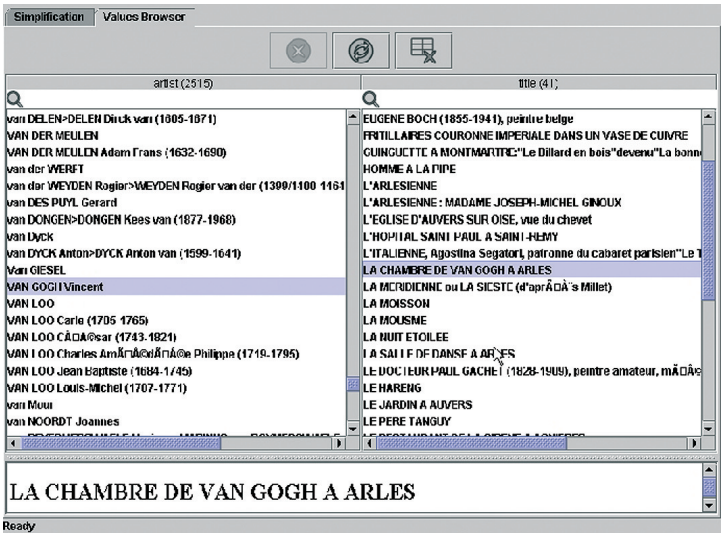

Figure 2. The mSpace interface component of the Concept Browser

displays the current slice arrangement, so that users can view the mSpace slices as they are put together in this interface. Predefined groups of slices can be selected, and users are able to save and load their own arrangements. As well as the TouchGraph display, the Concept Browser provides a more traditional interface involving a simple list of available slices that can be added and removed from the mSpace browser.

The instance information is obtained from the search and retrieval web service (SRW). This SRW is able to dynamically map queries expressed in terms of the CIDOC CRM mappings to the relevant database field, and then publish the results as XML structured according to the CRM. These results can then be displayed as slices in the 'values browser', the mSpace component of the Concept Browser interface. This approach is able to cope with the scale of the museum metadata that we are dealing with, and avoids issues such as data redundancy and maintenance of the semantic web storage version of the data (for example RDF).

The mapping of the EROS database with the CIDOC CRM

C2RMF decided to map the data format used in its EROS database to the CIDOC CRM. The aim is to develop an integrated query tool for both the EROS database and the C2RMF library catalogue—and perhaps in the future for other information resources as well.

Mappings to the CRM are important in that they allow us to explain implicit meanings that we take for granted or which are self-evident; although they are so only in the eyes of a given group of people in a given culture at a given time. Several mappings have already been made available: for the AMICO data model (Doerr 2000a), for the Dublin Core metadata element set (Doerr 2000b), for the Encoded Archival Description (EAD) DTD element set (Theodoridou and Martin 2001), and for the FRBR model (Le Boeuf 2002). Our complete mapping is available below as an appendix, and here we present the general principles that have guided us in the mapping process.

Our mapping of the EROS database format to CIDOC CRM had of course to take the specific structure of the EROS database itself into account. The EROS database contains three kinds of entries:

- entries describing the works themselves; such entries are characterized by field names beginning with the prefix 'oeuvre_'
- entries relating to various kinds of photographs of the works (X-ray photographs, pictures of details, and so on); field names begin with the prefix 'film_'
- and entries relating to restoration reports, all field names of which begin with the prefix 'report_'.

What binds together all the entries related to the same work is a code, called 'work

number', which appears respectively as 'oeuvre_worknbr', 'film_worknbr', and 'report_worknbr' in all three kinds of entry. These three data fields are mapped to the same CIDOC CRM class, E42 Object Identifier. As a rule, the instances of E42 resulting from 'oeuvre_worknbr', 'film_worknbr', and 'report_worknbr', should always be identical every time all three entries are related to the same instance of E84 Information Carrier. This condition, which is mandatory for the overall consistency of the database, could not be expressed as such in our mapping.

All fields in the EROS format that are only related to such 'housekeeping' activities as, for instance, the loan of photographs (such as film_empr for 'borrower name' or film_dtemp for 'lending date'), were excluded from our mapping efforts, as such information is explicitly declared out of scope of the CIDOC CRM model, requiring therefore specific extensions.

As the EROS database is primarily devoted to the *physical* aspects of works of art (their *physical* restoration), it was decided to take the class E84 Information Carrier as the focus for our mapping, even for those data fields that deal with conceptual/artistic aspects of the works rather than with their physicality, and even for such data fields that may seem to have only a loose relationship to the works themselves, as film_caimfilm (the technical legend of a picture). This allows for the overall consistency and unity of our mapping: one work in the original database = one instance of E84 Information Carrier throughout all fields of all kinds of entry.

It was also decided to regard photographs of works as instances of the class E31 Document, and restoration reports as instances of the class E73 Information Object. Given the structure of CIDOC CRM, this allows us to declare that these photographs 'document' (property P70) the works and that the restoration reports 'are about' (property P129) the works. All three kinds of entries are, in turn, modelled as instances of E73 Information Object (or of its sub-class E33 Linguistic Object), that are 'about' (P129), respectively, the work (instance of E84 Information Carrier), the photographs (instances of E31 Document), and the restoration reports (instances of E73 Information Object).

The semantics of each individual EROS data field is 'translated' into a chain of CIDOC CRM properties expressed from a domain class to a range class. As indicated above, the first domain class that begins that chain for any data field is systematically E84 Information Carrier; one range class in the resulting chain of properties must necessarily be instantiated with the value of the original EROS data field. For instance, the underlying semantics of the EROS data field oeuvre_record ('record author') is that the entry that was made in the EROS database for a given work of art, was made by a person who bears the name (or code or whatever) that is entered in the data field oeuvre_record. The whole chain of resulting properties in our mapping reads therefore as follows:

A given work of art (instance of the CIDOC CRM class E84 Information Carrier) is the subject (CIDOC CRM property P129 'is about // is subject of') of a specific text (instance of E33 Linguistic Object) that was created (property P94 'created // was created') through the event of it being conceived or 'written' (instance of E65 Creation Event), which event was realized (property P14 'carried out by // performed') by a given individual (instance of E21 Person) as (sub-property P14.1 'in the role of') author of the entry (value of an instance of E55 Type), which person, in turn, can be recognized as bearing (property P131 'is identified by // identifies') a specific name or code (instance of E82 Actor Appellation), which name or code is none but the value that was entered in the EROS data field oeuvre_record (value = 'value of 'oeuvre_record'). Once made explicit, the underlying semantics is always lengthy.

The EROS database allows almost every 'name' (of artist, technique, material, and so on) to be expressed in several languages, so that it can be searchable in any language among those that are proposed on the online database, according to the end-user's preferences. This feature is not reflected in our mapping, as it does not appear from the data format itself (the value actually entered, for example for an artist name, being a code, not the name itself), but the CIDOC CRM provides a means of expressing it, if this proved necessary, thanks to the P139 property, 'has alternative form', that can be declared between two instances of the E41 class, Appellation (or any of its sub-classes). For instance, the instance of E82 Actor

Appellation that has value 'Leonardo' (in Italian) 'has alternative form' (property P139) the instance of E82 that has value 'Léonard de Vinci' (in French) — both instances 'identifying' (property P131) the same instance of E21 Person.

UNIMARC mapping of the C2RMF library database

The mapping effort was then extended to the format used by the Louvre library, which collaborates with C2RMF, so that the Concept Browser could launch simultaneous queries on descriptions of works of art that are available in the EROS database, *and* bibliographic descriptions that are available in the library's catalogue, and so that information elements derived from both sources could contextualize each other and get interrelated.

This second undertaking proved much longer than the first one, as bibliographic formats are more detailed than the EROS format, and it is not finished yet. But it is (politically) very important, because bibliographic formats are highly standardized and in use in many different libraries. As a consequence, the resulting mapping will be available for other CRM-related projects involving libraries, well beyond the SCULPTEUR Project, and may be a milestone on the path toward the much debated notion of 'interoperability' between libraries and museums.

Just like many libraries, the Louvre library uses three distinct formats:

- a bibliographic format, for the description of 'publications', without reference to the physical copies actually held by the library: UNIMARC(B)
- an authority format, for the recording of information regarding authors, uniform titles, subjects, and so on: UNIMARC(A)
- and a local format for the recording of information relating to holdings.

At the time of writing (end-December 2004), UNIMARC(B) has been mapped to the CIDOC CRM, and the mapping that was done still has to be reviewed and validated. The two remaining formats will follow.

Although the Louvre library does not use *all* the fields and subfields present in the UNIMARC format, the decision was made to undertake an integral mapping of the format, so that the resulting document could be proposed to other institutions as well and could serve for other purposes. As a matter of fact, however, a few very specialized UNIMARC fields that are in use only for the description of cartographic materials were not included in the mapping.

The main semantic difficulty when mapping UNIMARC(B) to CIDOC CRM is that museum information as represented in CIDOC CRM focuses on *unique*, *physical items*, whereas library information focuses on the abstract notion of 'publication', that is, on *archetypal patterns* that are common to *sets of physical items*, all of which are supposed to be 'identical', or at least that are produced in the course of the same process consisting in the reproduction of the same content with the same layout.

The 'central' CRM class that had to be taken as a 'departure point' for each semantic chain in the mapping could therefore *not* be E84 Information Carrier, as in the EROS mapping, but E73 Information Object (despite the term 'object' in the name, this CRM class covers an abstract notion, the intangible *content* infixed on physical carriers).

Subject headings are especially important in this context, as names of artists and uniform titles for works of art, used as subject headings in catalogue records, will be in many cases the main or even only junction points between the library catalogue and the EROS database. No matter how accurate the mapping can be, no matter how effective the Concept Browser can be, if the strings for the same name or the same title are not perfectly identical between both databases, interoperability will be ruined. This is something that, perhaps, is not sufficiently taken into account by information specialists.

Conclusion

Interoperability between different databases related to domains as different as library and museum artefacts is more and more required for information retrieval. Mappings are a crucial element in projects such as SCULPTEUR: by referring to a common semantic model as robust as the CIDOC CRM distinct formats are able to communicate with each other, ensuring the overall consistency of the query system. The ontology reference model developed by the CIDOC is applied to simultaneously query both the C2RMF conservation and library databases. The system will be described and demonstrated to users as a result of the EU/IST funded SCULPTEUR project.

Acknowledgements

We thank the European Commission and Hewlett Packard for their support for the SCULPTEUR project.

References

CIDOC, 1995, *International Guidelines for Museum Object Information: The CIDOC Information Categories*, Paris: International Committee for Documentation of the International Council of Museums (CIDOC), ISBN 92-9012-124-6. Also available from World Wide Web: http://www.willpowerinfo.myby.co.uk/cidoc/guide/guide.htm.

Crofts, N, Doerr, M and Gill, T, 2003, 'The CIDOC Conceptual Reference Model: a standard for communicating cultural contents' in *Cultivate Interactive* [on line], issue 9, February, available from http://www.cultivate-int.org/issue9/chios/.

Doerr M, 2000a, 'Mapping of the AMICO data dictionary to the CIDOC CRM' technical report FORTH-ICS/TR-288, June 2000 (on line), Heraklion (Greece): Institute of Computer Science, Foundation for Research and Technology (cited 4 September 2003). Available from World Wide Web in pdf format: http://www.ics.forth.gr/proj/isst/Publications/paperlink/mappingamicotocrm.pdf; also available in rtf format: <http://zeus.ics.forth.gr/cidoc/docs/mappingamicotocrm.rtf.

Doerr M, 2000b 'Mapping of the Dublin Core Metadata element set to the CIDOC CRM' technical report FORTH-ICS/TR-274, July 2000 (on line), Heraklion (Greece): Institute of Computer Science, Foundation for Research and Technology (cited 4 September 2003). Available from Internet in pdf format:

Doerr, M and Crofts, N, 1999, 'Electronic Esperanto: the role of the oo CIDOC Reference Model' in *Proceedings of the ICHIM '99, Washington DC, 22–26 September 1999*. Also available from Internet: http://cidoc.ics.forth.gr/docs/doerr_crofts_ichim99_new.doc.

Goodall, S, Lewis, P H, Martinez, K, Sinclair, P A S, Giorgini, F, Addis, M J, Boniface, M J, Lahanier, C and Stevenson, J, 2004, 'SCULPTEUR: multimedia retrieval for museums. image and video retrieval' in *Proceedings of the Third International Conference, CIVR 2004, Dublin, Ireland, 21–23 July 2004*, 3115/2004, 638–646.

Lahanier, C, Aitken, G, Shindo, J, Pillay, R, Martinez, K, Lewis, P, 2002, 'EROS: an open source, multilingual research system for image content retrieval dedicated to conservation-restoration exchange between cultural institutions' in *ICOM-CC 13th Triennial Meeting, Rio de Janeiro, 22-27 September 2002, vol 1*, 287–294.

Le Boeuf, P, 2002, '*Mapping FRBR ≠ CRM*' (revised version, on line), Heraklion (Greece): Institute of Computer Science, Foundation for Research and Technology, February 2002 (cited 22 May 2003). Available from Internet in pdf format: http://cidoc.ics.forth.gr/docs/mapping_frbr-crm_revised.pdf; also available in doc format: http://cidoc.ics.forth.gr/docs/mapping_frbr-crm_revised.doc.

Schraefel, M C, Karam, M, and Zhao, S, 2003, 'mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia' in De Bra, P (ed.) *Proceedings of AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems*, *Nottingham, UK*, 217–235.

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, June 2001 (on line), Heraklion (Greece): Institute of Computer Science, Foundation for Research and Technology, June 2001 (cited 4 September 2003). Available from World Wide Web in pdf format: http://www.ics.forth.gr/proj/isst/Publications/paperlink/ead.pdf; also available in rtf format: http://zeus.ics.forth.gr/cidoc/docs/ead.rtf.