Using an ontology-driven system to integrate museum information and library information

Paper presented on the occasion of the Symposium on
Digital Semantic Content across Cultures,
Paris, the Louvre, 4-5 May 2006

Patrick Le Bœuf
Bibliothèque nationale de France

Abstract

"Infodiversity" (i.e., different ways of structuring information and of defining the content of information) exists among museums, and between museums and libraries. It is neither desirable nor feasible to standardise the practices of very different kinds of cultural heritage institutions. Rather than stretching them on a "Procrustean bed," it is now possible to rely, among other possibilities, on Semantic Web techniques, which enable such different institutions to preserve their specific needs and commitments, while ensuring the interoperability of the databases they produce. This is one of the goals of the European Project SCULPTEUR (2002-2005), which resulted in the development of a prototype for a query interface on heterogeneous databases, called "Concept Browser." This tool is driven by an ontology adapted from the CIDOC CRM conceptual model that was developed by ICOM CIDOC (to be published as ISO standard 21127 in the near future). The data structure from each of the museum databases involved in the project was mapped to the CIDOC CRM. In addition, it should be possible to integrate bibliographic databases to the Concept Browser, as the library format called UNIMARC was mapped to the CIDOC CRM as well. The Concept Browser, which is based on TouchGraph technology, allows one to visualise the ontology itself.

Résumé

L'« infodiversité » (c'est-à-dire, l'existence de diverses manières de structurer l'information et de définir le contenu de l'information) existe au sein des musées, ainsi qu'entre musées et bibliothèques. Il n'est ni souhaitable ni faisable d'obliger des types très divers d'institutions de mémoire à adopter les mêmes outils de traitement de l'information. Plutôt que de les allonger sur un « lit de Procruste », il est possible aujourd'hui de faire appel, entre autres possibilités, aux techniques du Web sémantique, qui permettent à ces différentes institutions de préserver leurs besoins et leur « philosophie » spécifiques, tout en rendant interopérables les bases de données qu'elles produisent. C'est l'un des objectifs du projet européen SCULPTEUR (2002-2005), qui a débouché sur la réalisation du prototype d'interface de recherche simultanée sur des bases de données hétérogènes appelé « Concept Browser ». Cet outil fonctionne autour d'une ontologie adaptée du modèle conceptuel CIDOC CRM élaboré par l'ICOM CIDOC (et bientôt norme ISO 21127). La structure des données de chacune des bases muséographiques inclues dans le projet a fait l'objet d'un « mapping » vers le CIDOC CRM. Il devrait être possible d'intégrer en outre au Concept Browser des bases de données bibliographiques, grâce à un mapping du format UNIMARC, en usage dans les bibliothèques, vers le CIDOC CRM. Le Concept Browser permet de visualiser l'ontologie elle-même, grâce à la technologie TouchGraph.

1. Infodiversity

It is a truism to say that information can be structured and expressed in an infinity of ways. As the classical joke puts it, "the nice thing about standards is that there are so many to choose from."

However, when it comes to cultural heritage information, "infodiversity is good,"[1] as such information is produced by different institutions with different needs and is about different types of resources.

The problem is that those differences may seem overrated to end-users. The challenge we have to face is that we would like *at the same time* to preserve a given institution's commitment such as expressed in the information produced by that institution about its collections,[2] and to meet the information needs of end-users who would rather not have to query each database individually and would like to benefit from the potential of Semantic Web technology, such as, for instance, the possibility for machines to make logical inferences from information elements that may be stored in distinct databases and in distinct formats.[3]

Infodiversity exists among museums, and between museums and libraries. But it would be wonderful if all information about cultural heritage could be *integrated*. We could then navigate seamlessly from museum information to bibliographic information and back: we could find information about objects that were created in a given cultural context, and about that cultural context; about objects that were made using a given technique, and about that technique; we could view digitisations of a given object, and find bibliographic references pointing to books or audio-visual materials about that object, etc.

It was the aim of the SCULPTEUR Project to integrate cultural information in such a way.

2. The SCULPTEUR Project

The SCULPTEUR Project was a European-funded project that began in 2002 and ended in 2005. It involved a dozen of partners, including the Centre de recherche et de restauration des Musées de France, the National Gallery and the Victoria and Albert Museum in London, the Galleria degli Uffizi in Florence, etc. The University of Southampton was in charge of technical infrastructure and software development.

The SCULPTEUR Project resulted in the implementation of the "Concept Browser," a graphic interface that makes it possible to query heterogeneous museum databases and to visualise digital reproductions of the works of art described in those databases.

The Concept Browser integrates the information produced by each of the Project's partners by mapping it to a common ontology that formalises the underlying semantics of museum information, no matter in which specific format it is expressed and stored. That common ontology is the CIDOC CRM.

3. The CIDOC CRM

The CIDOC Conceptual Reference Model (CIDOC CRM) was developed specifically for museum information by the ICOM CIDOC (International Council of Museums, International

---

[1] Gill, Tony. "When the rubber hits the road: using the CIDOC CRM in the real world." In: *Sharing the knowledge: International CIDOC CRM Symposium, March 26-27, 2003* [on line]. Heraklion, Greece: FORTH, 2003 [cited 5 August 2005]. Available at: <http://cidoc.ics.forth.gr/docs/symposium_presentations/gill_2003-when-rubber_hits_road.ppt>.

[2] "Collections usually come with a point of view/commitment which is lost during the aggregation of descriptions." Isaac, Antoine. "Accessing cultural heritage collections using Semantic Web techniques." In: *[Proceedings of] DE [Digitaal Erfgoed] Conferentie, 9-10 november 2005* [on line]. [Den Haag]: [Digitaal Erfgoed Nederland (DEN)], 2005 [cited 21 April 2006]. Available from World Wide Web: <http://www.den.nl/bestanden/Conferentie_2005/A_Isaac.pdf>.

[3] "For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning." Berners-Lee, Tim; Hendler, James; and Lassila, Ora. "The Semantic Web." In: *Scientific American.com* [on line]. Available from World Wide Web: <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>.

Committee for Documentation[4]). It has been in development since 1996 and it is about to be published as an ISO standard, ISO 21127. 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.[5]

The central notion in CIDOC CRM is the notion of Event: something that happens in space and time and brings about some change in the world.

An event (i.e., an instance of the Event class) can involve:

- instances of the Actor class (persons, groups…), who can play a decisive role in provoking the event or just witness it or undergo it, and who are referred to through instances of Actor Appellation;
- bits of the physical world and/or creations of the mind (i.e., instances of the class named Physical Thing and/or the class named Conceptual Object; e.g., canvass and paint on the one hand, and the image formed by the paint on the canvass, on the other hand), which are referred to through instances of Appellation (names, titles, codes, whatever).

In addition, an event:

- occurs in time, and has therefore a duration, i.e., an instance of the class named Time-Span, which is referred to through instances of Time Appellation (e.g., instances of Date);
- and occurs in space, and can as such be related to an instance of Place, which is referred to through an instance of Place Appellation.

Finally, we use the notion of Type (a further class declared in CIDOC CRM) to categorise all those things and produce a well-organised view of this complex, chaotic real world.

The CIDOC CRM is an impressive semantic model that has 81 classes and 132 properties.

The Concept Browser is able to display the CIDOC CRM ontology in a graphical way, and allows users to navigate to concepts of interest and request to view instances of certain concepts. The graph visualisation interface is based on "TouchGraph," an open source graph layout system that has been extended and adapted to suit the Project's requirements.

Informal user trials involving museum and gallery partners showed that the terminology and complexity of the CIDOC CRM proved too challenging to visualise in an intuitive way. This led to the creation of customisable simplifications based on each museum's legacy metadata structure to increase familiarity for the museum users. These simplifications are more appealing to end-users, as well-known concrete notions such as "material," "technique" and "place," are used instead of the awkward CIDOC CRM concept and property labels.

4. Bibliographic information

Although the CIDOC CRM focuses on museum information, it proved possible to use it in the context of bibliographic and museum information integration. Much of the semantics defined in the CIDOC CRM for museum objects is also valid for the description of bibliographic resources. The main difficulty comes from the fact that museum descriptions relate to physical, individual, "unique" objects, whereas bibliographic descriptions focus on the abstract notion of "publications", which are exemplified by the individual items actually held by libraries. An international working group is currently striving to harmonise those two views by expressing the FRBR model (Functional Requirements for Bibliographic Records), the conceptual model developed by IFLA (International Federation of Library Associations

[4] <http://www.willpowerinfo.myby.co.uk/cidoc/>.

[5] Crofts, Nicholas; Doerr, Martin; & Gill Tony. "The CIDOC Conceptual Reference Model: a standard for communicating cultural contents". In: *Cultivate Interactive* [on line]. Issue 9. February 2003 [cited 21 April 2006]. Available from World Wide Web: <http://www.cultivate-int.org/issue9/chios/>.

and Institutions) for bibliographic information, and "plugging" it into the CIDOC CRM. However, this objective is far from being achieved, and there is no publicly available documentation about this process for the time being. Since bibliographic and museum information integration was one of the objectives of the European-funded SCULPTEUR project, and this project had a rather tight deadline, it was decided not to wait until FRBR and CIDOC CRM models were harmonised, and to start working using solely the CIDOC CRM model.

It took over two years to map the UNIMARC Bibliographic format to the CIDOC CRM, and that work was finished quite recently, in April 2006. As a consequence, it has not been possible yet to check the validity of the mapping by actually integrating bibliographic information and museum information in the Concept Browser. Although it is therefore too early for an evaluation of the Concept Browser as a pathway between libraries and museums, we firmly believe, however, that this tool opens the way to innovative interfaces of the future, and for an implementation of Semantic Web techniques in the field of cultural heritage information.

5. Possible next steps

To map the formal structure of metadata schemas to the concepts declared in CIDOC CRM, though it is a helpful but time-consuming activity, is not sufficient. In order to enrich the potential for information integration and logical inferences, we still have to map thesauri, classification schemes and subject authority files to CIDOC CRM. For instance, if all subject headings for techniques and materials were explicitly declared as representing instances of the corresponding high-level concepts in CIDOC CRM, the Concept Browser would enable end-users to navigate seamlessly from bibliographic records for studies on a given technique or material to descriptions and digitised reproductions of museum artefacts the creation of which actually employed that technique or material, or vice-versa. But such an objective would still require a significant effort in development and is neither planned nor scheduled for the time being.