The CIDOC CRM, an Ontological Approach to Schema Heterogeneity

Martin Doerr

Institute of Computer Science,
Foundation for Research and Technology – Hellas,
Science and Technology Park of Crete,
Vassilika Vouton, P.O. Box 1385, GR 711 10, Heraklion, Crete, Greece
martin@ics.forth.gr

The creation of the World Wide Web has had a profound impact on the ease with which information can be distributed and presented. Now with more and more information becoming available, there is an increasing demand for targeted global search, comparative studies, data transfer and data migration between heterogeneous sources of cultural and scholarly contents. This requires interoperability not only at the encoding level - a task solved well by XML for instance - but also at the more complex semantics level, where lie the characteristics of the domain. In the meanwhile, the reality of semantic interoperability is getting frustrating. In the cultural area alone, dozens of "standard" and hundreds of proprietary metadata and data structures exist, as well as hundreds of terminology systems. Core systems like the Dublin Core represent a common denominator by far too small to fulfil advanced requirements. Overstretching its already limited semantics in order to capture complex contents leads to further loss of meaning[1].

The CIDOC Conceptual Reference Model (CRM) [2], [3] is a core ontology that aims at enabling information exchange and integration between heterogeneous sources of cultural heritage information, archives and libraries. It provides semantic definitions and clarifications needed to transform disparate, heterogeneous information sources into a coherent global resource, be it within a larger institution, in intranets or on the Internet. Its use comprises: intellectual guidance for conceptual modelling; common upper level for application ontology development; reference ontology for data translation and global schema in federated systems [4], [5], [6].

The CIDOC CRM was developed over the past 8 years by an interdisciplinary working group of the International Committee for Documentation of the International Council of Museums (CIDOC/ICOM) under the scientific lead of ICS-FORTH. It is on vote as Draft ISO standard (ISO/DIS 21127) until February 2005 [7]. The latest version 4.0, which was released on March 12th, 2004, consists of 80 classes and 132 properties. It is developed in the knowledge representation language TELOS, and available in RDFS and other formats.

It is commonly accepted, that the employment of formal ontologies and the respective enabling technology for their use in information systems is currently the only way to reach the precision of human-mediated knowledge. It is equally widely assumed, that ontologies are highly application and domain specific, and necessarily huge, so that a

Dagstuhl Seminar Proceedings 04391
Semantic Interoperability and Integration
http://drops.dagstuhl.de/opus/volltexte/2005/35

generalization from an information technology point of view is not possible. In contrast to most current ontologies, the CRM is a core ontology for capturing the common semantics of heterogeneous data structures in order to support their semantic integration, and not a formal account of expert terminology. It is argued that such an ontology is property-centric, compact and highly generic, in contrast to terminological systems. Furthermore it seems that a core set of relationships is more fundamental to knowledge integration than the mapping of terminology.

The CRM is result of a strategic, careful, long-term knowledge engineering process from existing data structures and experts of various museum disciplines, libraries and archives. Historical knowledge, be it political, cultural, scientific, medical etc., is incomplete and alternative opinions are in general undecidable. Its elements have different statistical stability against knowledge revision. So is the existence of things as elements of our discourse only, such as London, Caesar, King Arthur, Aphrodite and the Sinking of Atlantis, more stable than knowledge if they have existed in reality. I.e. we can agree on what we mean by Atlantis without it having ever existed. If the work of Shakespeare was written by one or more "Shakespeares", is less sure than that there existed at least one writer. If someone is a criminal or hero may be more debatable than that he is a human being. Whereas no one doubts the existence of El Greco and the village Fodele, the relation of El Greco to this village as his birthplace is debated. If someone committed an action may be more debated than his/her participation etc.

Rather than elaborating the differences of specialized terms, the CRM is a system concepts and relationships that are systematically found to be most robust against change of context and perspective, such as *events* and *actors*, *participation*, *classification*, *location*, *identification*, and hence allow for relating seemingly incompatible information. Distinct features are: Generalization of (spatio)temporal entities; explicit modelling of events that are normally hidden in relationships such as "has creator", "birthplace", etc.; consistent modelling of the material and geometric notion of place; systematic connection to terminological systems; explicit modelling of the relationship of identifiers and the identified, see Fig.1.

Applications have already shown that such a well-crafted core ontology can help to achieve a very high precision of schema integration at reasonable cost in a huge, diverse domain [8],[9],[10], [11], [12],[13],[14], [15].

It is argued that three levels can be distinguished in the knowledge life-cycle, where a such a core ontology plays a central role: (1) data acquisition, (2) information integration, (3) interpretation and story-telling.

Data acquisition is mostly characterized by a work-flow elaborating series of analogous items, such as library catalogues, collection management systems of museums, epidemiological studies, biodiversity registers etc. It requires highly ergonomic documentation units, completeness of information, a very case-specific language and data quality control the. The interoperability needs are restricted to the capability to map the documentation records to other systems. A core ontology can be

used to derive compatible , application-specific document structures, and as language to mediate data structure semantics in mapping processes of legacy data to other systems.

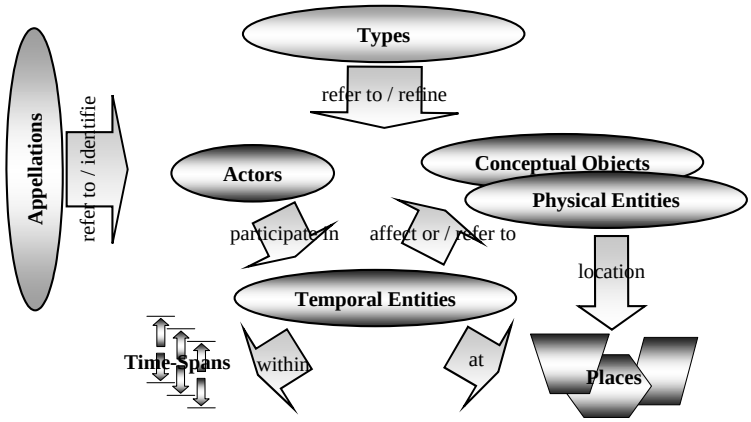

Fig. 1. CRM-metaschema

Information integration is the focus of the CRM. Information integration must break up document boundaries and relate contained information to a wider context, match shared identifiers of items, and compile alternatives. There is no preference direction of access by particular subjects, relationships, or classes of items. Semantic interoperability requires a global schema. A core ontology can be used for that purpose with suitable addition of administrative elements. It should provide the relevant relationships to enable exploration of contexts by data paths across various source document units, see Fig. 2.. Suitable application specific *access profiles* can be defined in terms of the core ontology.

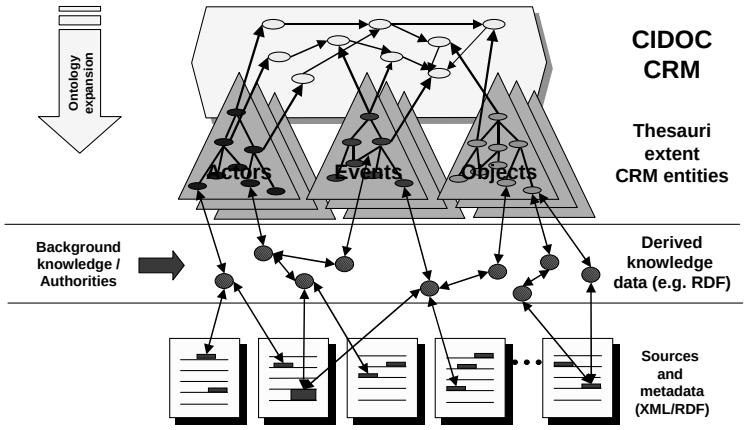

Fig. 2. Information Integration under a core ontology.

Interpretation can only take place on top of integrated knowledge. It will explore contexts, arbitrate between alternatives, make hypotheses and collect evidence by collecting all relevant facts or doing various statistics. A core ontology appears as enabling factor, but may be enriched by various interpretative elements, such as causation, states free of change, etc.

It is further argued that such ontologies are widely reusable and adaptable to other domains. There are good indications that the determinants of such a core ontology are the kind of discourse, such as retrospective, historical analysis, which appears generically across domains, and generic functions, such as knowledge integration, rather than any domain-specific concepts. This opens a way to highly effective and yet economical methods of information integration.

References

[1] Baker, T.. A Grammar of Dublin Core. D-Lib Magazine, Vol. 6, No. 10 (2000)

[2] Doerr, M.. The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, Vol. 24, No. 3 (2003)

[3] Doerr, M., Crofts N.. Electronic Esperanto: The Role of the Object Oriented CIDOC Reference Model, 1999. In Proceedings of the ICHIM '99, Washington DC, September 22-26 (1999)

[4] Guarino, N.. Formal Ontology and Information Systems. In: Guarino, N. (ed.): Formal Ontology in Information Systems. In Proceedings of the 1st International Conference, Trento, Italy, June 6-8 (1998). IOS Press

[5] Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., Rosati R.. Description Logic Framework for Information Integration. In Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98) (1998) 2-13

[6] Wiederhold, G.. Mediators in the Architecture of Future Information Systems. IEEE Computer, Vol. 25, No. 3 (1992)

[7] http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=34424&scopelist=PROGRAMME

[8] Crofts, N.. Implementing the CIDOC CRM with a relational database. MCN Spectra. Vol. 24, No. 1 Spring (1999)

[9] Borbinha, J.. Authority Control in the World of Metadata. In Proceedings of the International Conference Authority Control: Definition and International Experiences, February 10-12 (2003)

[10] Vassilev, V., Stoev, I., Gaydarska, B., Alexandrov, S., Nehrizov G., Vaklinov, M.. Museum Information Systems: CIDOC data model implementation in the ArchTerra project. BOLLETINO CILEA, No. 69, Settembre 1999

[11] Hatala, M., Kalantari, L., Wakkary, R., Newby, K.. Ontology and Rule based Retrieval of Sound Objects in Augmented Audio Reality System for Museum Visitors. In Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus. ACM Press New York, NY, USA (2004) 1045 – 1050

[12] de Haan, G.. The Design of I-Mass as a Tool for Interacting with Cultural Heritage. In Proceedings of the 1st International Symposium on Information and Communication Technologies, Dublin, Ireland. ACM Press New York, NY, USA (2003) 415 - 420

[13] Mulholland, P., Collins, T., Zdrahal, Z.. Story Fountain: Intelligent Support for Story Research and Exploration. In Proceedings of the 9th International Conference on Intelligent User Interface., Funchal, Madeira, Portugal. ACM Press New York, NY, USA (2004) 62 - 69

[14] Doerr, M., Schaller, K., Theodoridou, M.. Integration of complementary archaeological sources, 2004, To Appear in Proceedings of the Conference on Computer Applications and Quantitative Methods in Archaeology, CAA2004, Prato, Italy April 13-17 (2004)

[15] Hunter, J., Combining the CIDOC CRM and the MPEG-7 to Describe Multimedia in Museums. In Proceedings of the International Conference about Museums and the Web Boston, Massachusetts (2002)