The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing

ER2007 Tutorial

Martin Doerr[1], Christian-Emil Ore[2], Stephen Stead[3]

[1] Institute of Computer Science, FORTH, Heraklion,Crete Greece
[2] University of Oslo, Oslo, Norway
[3] Paveprime Ltd, UK

martin@ics.forth.gr, c.e.s.ore@edd.uio.no steads@paveprime.org

Abstract

The tutorial first addresses requirements and semantic problems to integrate digital information into large scale, meaningful networks of knowledge that support not only access to source documents but also use and reuse of integrated information. The pros and cons of developing global ontologies are discussed. It is argued that core ontologies of relationships are fundamental to schema integration and play a completely different role to that of specialist terminologies in practical knowledge management. The CIDOC Conceptual Reference Model (CRM) is presented as an example of such a global model. It is a core ontology and new ISO standard (ISO 21127, accepted September 2006), originally designed for the semantic integration of information from museums, libraries, and archives. It is a product of re-engineering the dominant underlying common concepts from representative data structures. It is not prescriptive, but provides a controlled language to describe common high-level semantics that allow for information integration at the schema level. The tutorial addresses part of the technology needed for information aggregation and integration in the global information environment, namely the question to which extent and in which form global schema integration is feasible. The ability of the CRM to support integration has been demonstrated in a large range of different domains including cultural heritage, e-science and biodiversity. Conceptual modeling by specializing such a well-tested core ontology not only reduces drastically development time and improves system quality, but provides basic semantic interoperability more or less for free. The tutorial will present characteristic applications.

Keywords: Information Integration Semantic Interoperability, Ontology Engineering, Schema Heterogeneity, Event Model.

1 Information Integration, Ontologies and Knowledge Networks

Data-driven science has emerged as a new model which enables researchers to move from experimental, theoretical and computational distributed networks to a new paradigm for scientific discovery based on large scale distributed GRID networks. Hundreds of thousands of new digital objects and immense numbers of encoded facts are placed on the Web, in digital repositories and other information systems everyday, supporting and enabling research processes not only in science, but in medicine, education, culture and government. It is therefore important to build infrastructure and web-services that will allow for exploration, data-mining, semantic integration and experimentation across all of these rich resources.

A prime example of a scientific discovery that emerged from the re-use of existing resources is Mendeleev's Law of Periodicity. As Peter Murray-Rust (2007) points out: "The law of periodicity was thus a direct outcome of the stock of generalizations and established facts which had accumulated by the end of the decade 1860-1870; it is an embodiment of those data in a more or less systematic expression."

Mendeleev's law emerged from a concatenation of facts extracted from the current published chemical literature which appeared in many languages and symbolic formulations; the analysis of relations in the data and metadata – the experimental conditions – were critical for establishing his conclusion.

The ultimate goal of users of a scientific information system is not to retrieve an object or numbers but to *understand* a topic. Understanding is built on associations. Associations are found in digital objects or data structures. Data structures provide explicit associations in the form of relationships and data paths. Tools may extract associations from digital objects, either by interpretation of data structures or by statistical means such as evaluation of co-occurrence patterns, and save them again as data structures. If the semantics of represented relationships are explicit, such as part-whole, membership, creation and participation, then patterns in the network of factual relations (or *material facts* as called by Degen et al., 2001), can reveal new, indirect associations, or can be used for inductive reasoning.

Current data warehouse technology focuses on detecting statistical patterns for data retrieval. Base technology for the more fundamental operation, the concatenation of facts, is widely missing. Factual relations however can be concatenated to form huge meaningful semantic networks – the driving vision behind W3C's promotion of RDF and OWL. Cardoso and Sheth (2006) have stressed the extraordinary importance of access by factual relationships for the Semantic Web, in particular with respect to business applications.

In order to support any advanced services, relationships (i.e. classes of relations) should conform to a schema or ontology. Even though it is widely believed that there is no agreement on a global ontology, the wide acceptance of Dublin Core demonstrates the opposite. If there is one or a few core ontologies, does not make any difference in their ability to give rise to global networks of knowledge. Empirical studies show (Maganaraki, et all 2002) that the number of relationships in ontologies is orders of magnitudes smaller than that of classes and hence quite manageable. Doerr (2003), Doerr, Hunter, and Lagoze (2003), and Sinclair (2006) have shown that a core ontology of ten to a hundred relationships can capture semantics of data structures across many domains. This simplification was possible under the assumptions and restrictions as described in the next sections.

2 Scope, History and Form of the CIDOC CRM

The CIDOC CRM is a formal ontology (Crofts, et all 2005) intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It started bottom up, by reengineering and integrating the semantic contents of more and more database schemata and documentation structures from all kinds of museum disciplines, archives and recently libraries.

The development team applied strict principles to admit only concepts that serve the functionality of global information integration, and other, more philosophical restrictions about the kind of discourse to be supported, as described below (for more details see Doerr, 2003). The application of these principles was successful in two ways. On the one side, the model became very compact without compromising adequacy. The very first schema analyzed in 1996, the CIDOC Relational Data Model with more than 400 tables (described by Reed, 1995), could be reduced in 1996 to a model of about 50 classes and 60 properties, with far wider applicability than the original schema.

On the other side, the more schemata were analyzed, the fewer changes were needed in the model (see version history, http://cidoc.ics.forth.gr/2007). The present model contains 80 classes and 132 properties, representing the semantics of may be hundreds of schemata. As a result of the successful reformulation of the original relational model CIDOC started the standardization process in collaboration with ISO in 2000. The model was accepted as ISO21127:2006 in Sept. 2006.

Deliberately, the CIDOC CRM ontology is presented in a textual form to demonstrate independence from particular knowledge representation formats. There exists however a formal definition in TELOS (Mylopoulos, et all 1990, Analyti, Spyratos and Constantopoulos 1998). It distinguishes individual classes from properties (binary relationships). Properties are directed and bidirectional, with distinct labels for each direction. It employs strict multiple inheritance (without exceptions) for both classes and properties. It foresees multiple instantiation, i.e. one particular item can accidentally be instance of more than one class. Domain and range of properties are associated with quantifiers zero, one or many. There exist valid equivalents in KIF, RDFS and OWL, to the degree the respective constructs are supported.

3 Intended Use and Development Methodology

The focus of the development of the conceptual model was on the ontological commitment, i.e. the empirical confirmation from experts and documentation examples that the concepts in the ontology are the relevant ones that experts actually share. Further, the functional adequacy of the ontology was validated and is continued to be validated carefully against characteristic *sets of questions or queries from domain experts*. Since most of the concepts in the model are primitive ones, definitions in the form of logical expressions beyond subsumption, domain and range, are not yet given. Rather, concepts are defined textually and supplied with examples from the scholarly discourse to make sure that the domain experts will have the same understanding of their meaning. Several frequent deductions are contained implicitly in the ontology. For instance, a person ("E39 Actor") acquiring a title to an object will also be the owner of this object. Logical definition of such deductions is due to future work.

The aim of the CIDOC CRM was to develop an ontology for data interchange in the very wide sector of cultural heritage. Without a set of strict design principles the domain experts would readily provide an immense number of concepts and relationships as relevant for their specialized domains, and no generic concept may be found. The most important principles are

1. For each concept there must be evidence from actual data structures widely used. The concept must be found to be underlying elements of a data structure. The data structures are the representatives of their domain of use. The frequency of use of a data structure is a measure for its relevance for the core ontology. Rather than expert opinions, this is the warrant that people actually analyze their data in such detail. As this is associated with labour, it is assumed that this analysis is useful for their work.
2. The development is bottom up. A relationship is only declared to the degree of genericity well

understood from the empirical base. If it turns out to be more generic, the respective change to the model is backwards compatible. This is a major difference to attempts such as Dublin Core.

3. The CIDOC CRM concentrates on the definition of relationships, rather than terminology, in order to support mediation, transformation and integration between heterogeneous database schemata and metadata structures, as well as good practice of conceptual modeling for documentation systems. The core ontology contains only the fundamental classes necessary as basic constraints of range and domain for the relationships ("properties") in the ontology.
4. We deal with information about the present or past, in contrast to enforcing plans. Besides others, this means there is no need not enforce cardinality constraints in the integrated resource. Data are assumed to come in valid from the acquisition phase. Any violation of cardinality constraints in the integrated information space is interpreted as an aggregation of alternatives. Since data are not used to control systems, no harm is done. (For instance, this does not hold for mediation of command structures for Web Services).
5. We deal only with discrete roles and entities, as they appear in a mesoscopic, human-centric environment, but not numerical values from continuous mathematical spaces. We assume that in experimental science and business applications, an integration process of discrete entities is always prior to integration of numerical spaces, and that the respective non-discrete information units can be passed through the discrete integration process unanalyzed to subsequent processing. For instance, experimental scientific data can only be integrated if the experiments are comparable on the ontological level (are about the same kinds of phenomena).

Point 3 needs some additional explanation: Whereas one can imagine a schema with only one class and many relationships, a schema with many classes and only one relationship makes hardly any sense. Schema semantics lives from relationships. Therefore we introduce into the ontology only classes necessary for the definition of relationships. If the ontology is to be used as a schema, then the classes that are not needed as domain or range for some relationship ('property'), eg. terminology, are treated as "data" and connected to the ontology by the general property "P2 has type". This implies that the actual information integration process needs a separate mechanism to deal with the mapping of all the more specialized classes and terms. The Semantic Web literature is full of such methods. It is argued that the process of integration of relationships, corresponding to schema integration (mediation, transformation and merging), can be separated from integration of classes.

We have *empirical confirmation*, from mapping many database schemata from different domains and actual harmonization efforts with competitive proposals, that under the above restrictions and conditions the CIDOC CRM describes data structures of a wide range of domains without loss of meaning in the aggregation process. Even though it was initially engineered from data structures in museum applications, most of the classes and relationships are surprisingly generic. They are characteristic for the logic of retrospective documentation as it occurs generically in science, culture studies and news. Furthermore, we have found that applications from quite different domains can be dealt with as straight-forward specializations of this single ontology. This makes us believe that other application classes may also be described by similar, highly generic core ontologies. If incompatible ontologies are found, research should reveal the functional requirements leading to the incompatibilities. This analysis may provide the key arguments to harmonize or merge incompatible ontologies (see also Doerr, Hunter, and Lagoze 2003).

The benefits of a common core ontology are immense. Research should focus more on the engineering, harmonization and *empirical validation* of common core ontologies of relationships. The separation of terminology from schema semantics has a great practical advantage. The stability and hence possible agreement on common semantics for schema-level semantics is much higher than for terminology. The scale is much smaller and language does not pose a major obstacle.

4 Central Concepts of the CRM

The tutorial will present the major constructs of the CRM in the form of graphic representations pertaining basic information functions, such as identifying, classifying, locating things, part decomposition, participation, reference and influence. The driving principle is the explicit modeling of events. It allows for the representation of metadata, such as creation, publication, and use, as well as content summarization. The representation of events allows for connecting facts into coherent representations of history. The use of CRM concepts is not prescriptive, but provides a controlled language to describe common high-level semantics that allow for information integration at the schema level.

Four ideas are central to the CRM (see Figure 1):

1. The possible ambiguity of the relationship between entities and the identifiers ("Appellations") that are used to refer to the entities are a part of the historical reality to be described by the ontology rather than a problem to be resolved in advance. Therefore, we distinguish the nodes representing the real item from the nodes representing the names of the item.
2. Types and classification systems are not only a means to structure information about reality from an external point of view, but also part of the historical reality in their nature as human inventions. Similarly, all documentation is seen as part of the reality, and may be described together with the documented content itself.

3. The normal human way to analyze the past is to split up the evolution of matters into discrete events in space and time. Thus the documented past can be formulated as series of events involving "Persistent Items" (also called endurants, see Doerr, 2003) like Physical Things and persons ("Actors"). The involvement can be of different nature, but it implies at least the presence of the respective items. The linking of items, places and time through events creates a notion of history as "world-lines" meeting in space and time. Events, when seen also as processes, are generalized as "Periods" and further as "Temporal Entities" (also called perdurants Doerr 2003). Only these classes are directly connected to space and time in the ontology. The temporal entities have fuzzy spatiotemporal boundaries which can be approximated by outer and inner bounds.
4. Immaterial objects ("Conceptual Objects") are items that can be created but can reside on more than one physical carrier at the same time, including human brains. Immaterial items can be present in events through the respective physical information carriers. Immaterial items cannot be destroyed, but the last carrier may be lost.

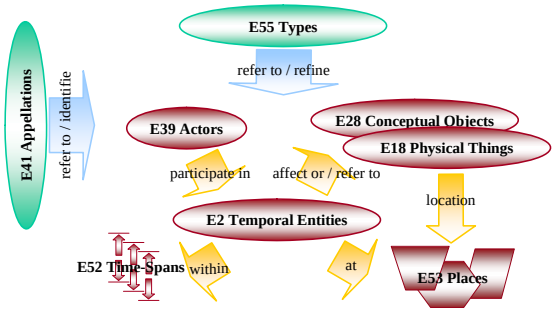

Figure 1: Fundamental concepts of ISO21127

The best way to summarize the contents of the CRM is to look at the functions supported by its relationships. Those are:

- Identification of real world items by real world names.
- Classification of real world items.
- Part-decomposition of immaterial and physical things, temporal entities, groups of people (Actors), Places and Times.
- Participation of persistent items in temporal entities.
- Location of temporal entities in space-time and physical things in space.
- Influence of objects on activities and products and vice-versa.
- Reference of information objects to any real-world item (aboutness).

The content of the CRM can be presented on one side in terms of the IsA hierarchy of classes and relationships. A better understanding is achieved if it is presented as distinct logical units answering questions with respect to the above functions, such as "by which constructs does the model describe where things are located, have been located or how they changed location". All relationships can be associated with such questions. In the tutorial, we show the most important parts of the model.

5 Application

We discuss in this section applications at three stages in the information lifecycle: information acquisition, aggregation and reuse.

Information acquisition usually happens in very specialized environments. People may register information about potsherds of Ancient Greek vases, the letters of a poet, the history of buildings, the water levels of a river or the results from a zoological field trip. The data entry systems and the underlying data structures are usually highly specialised to support the users in entering the data fast, accurately and completely. There is a characteristic discourse associated with the maintenance, correction and preservation of such primary information. The *preservation of the original documentation unit* is useful to maintain knowledge about its circumstances of creation and to update information.

It makes in general no sense to express the respective data structures directly in terms of a global ontology.

However, whether objects are man-made, such as the Ancient Greek vase or buildings, or created naturally, such as the insect on the pin or the river, 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 do. As objects they have been collected by somebody, at a place within a specific period of time using known or unknown methods.The pinning of the insect and the construction of the building, the measurement of the wingspan of the insect and of the water level of a river share fundamental contextual characteristics. In most cases they can be mapped into a rather generic form of activities and their parts, subjects and objects of those parts, kinds of dimensions, the where and when etc. The process can also be inverted. A suitable data structure can be created by *specializing a core ontology*, so that the mapping is explicit from the beginning. We demonstrate this process with a data structure for tracing the workflow provenance of digital images, proposed for complex 3D image processing steps used in archaeology.

The shared fundamental contextual characteristics mentioned above are the key for basic information aggregation and integration of complementary, related information. Technically, it is achieved by the mapping between a specialized structure and the core ontology. In this application, the latest stage of knowledge should appear as *one network of knowledge* that can be read in any direction. Document boundaries appear rather as an obstacle.

This network can best be represented by a core ontology of relationships combined with suitable domain terminologies. Besides a common conceptual model, the

creation of the network requires detecting all duplicate identifiers, a fairly complex process beyond the scope of this tutorial. It would be the best if this network could be created at any time from its sources to avoid the complexity known from data warehouse updating. We demonstrate this case with an example from biodiversity, relating field observation with collection data, an example from archaeology relating person names written on Roman stones with places of finding also described by Doerr, Schaller and Theodoridou (2004), and an example from cancer research relating clinical observation with gene activation measurements.

The virtual or physical network of aggregated knowledge is the basis for information use and reuse. It can support the understanding of contextual factors (see Figure 2), indirect relationships, conflict resolution, statistical evaluation for inductive reasoning, deductive reasoning, testing of hypotheses and alternatives, presentation of views and story telling. In this phase, advanced reasoning methods and highly case specific tools may be employed. The global ontology is useful for a retrieval and selection of relevant related data, but only partially useful for the subsequent kinds of use.

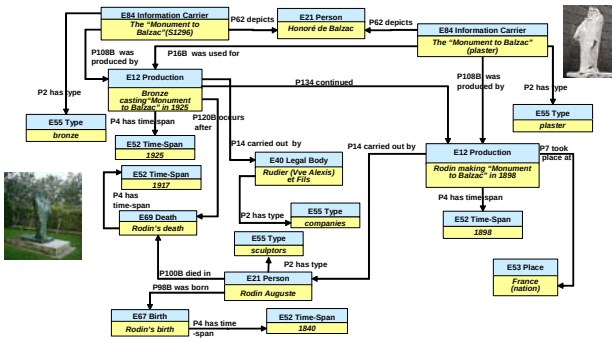

Figure 2: Graphical model of associations about the casting in bronze of Rodin's "Monument to Balzac" after his death, demonstrating the importance of explicit event.

In this light, we see the major applications of the CIDOC CRM in

- good practice of conceptual modeling for data acquisition systems;
- physical or virtual information integration;
- data transformation to an application neutral form for migration and preservation

The ability of the CIDOC CRM to support integration has been demonstrated in a large range of different domains including cultural heritage, e-science and biodiversity.

6 Extension

The CRM foresees domain-specific specialization. With or without suitable specialization and extensions, the CRM can be used as a backbone model for conceptual modeling of documentation systems in a wide range of domains. This is demonstrated by the above mentioned model of clinical observation in cancer and a model tracing workflow provenance of digital images (see Figure 3). We are always surprised how few additional concepts may be needed to cover applications in different domains.

Over the past three years, there is a collaboration of CIDOC with the International Federation of Library Associations on harmonizing the generic library model "Functional Requirements for Bibliographic Records" (FRBR) (IFLA Study Group 1998) with the CIDOC CRM. The work resulted in a model ("FRBRoo") of about 50 classes and 60 relationships that are actually subsumed by the CIDOC CRM (see LeBoeuf 2005). This model details the intellectual creation process, performing arts, recording and publication work, as well as bibliographic practice related to tracing the identity of the related intellectual products. Since it involved reengineering a set of poorly designed entities from an information system point of view, it can be seen as a larger conceptual modelling task in its own. It gave rise to several minor, backwards compatible changes to the CRM itself, i.e. essentially generalization due to the extended evidence from library practice. Both models together have been verified to cover the key conceptualization of documentation done by memory institutions like museums and libraries, facilitating access to an important portion of the global cultural heritage information.

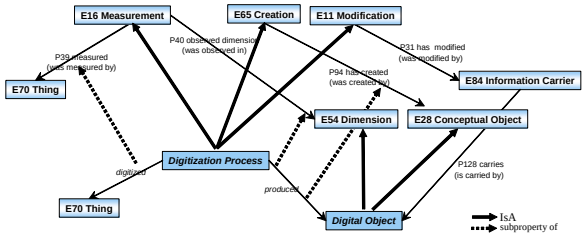

Figure 3: Part of a model of digital provenance: representing a digitatization process as a specialization of the CRM concepts E16 Measurement, E65 Creation and E11 Modification.

Any major extension is expected to shed some light on aspects such a core ontology as the CRM, that can be generalized, and more specific constructs, that may be of general use. Important is to keep these changes backwards compatible, a thing we have widely succeeded in

7 Conclusions

Systematic analysis of data structures under the perspective of information integration can provide new insight into conceptual modeling. Generic patterns and relationships emerge, that are overlooked in a local application context. These can be formulated in core ontologies of very wide applicability. Conceptual modeling by specializing a well-tested core ontology not only reduces drastically development time and improves system quality, but provides basic semantic interoperability more or less for free. Whereas this has been theoretically required since a longer time, we show this in practice, and with a global model which is

significantly smaller than others (such as SUMO) and has passed the status of an ISO standard.

The CIDOC CRM provides interesting perspectives for good practice of conceptual modeling, global interoperability and information integration in a wide area of domains, in particular for recorded knowledge. With this tutorial, we present some of the wealth of experience and empirical evidence collected with this model by an open, international, interdisciplinary team over nearly fifteen years, and encourage wider application. We also encourage our audience to look into areas outside the scope of the CRM to see if common ontologies can be crafted. It is a very time consuming and intellectually demanding process, but we believe its results justify the effort.

8 References

Murray-Rust, P. (2007): Data Driven Science – A Scientist's View, *In NSF/JISC 2007 Digital Repositories Workshop*, http://www.sis.pitt.edu/~repwkshop/papers/murray.html

Degen, W., Heller, B., Herre, H. and Smith, B. (2001): GOL - Towards an Axiomatized Upper-Level Ontology. *Electronics and Computer Science.*

Cardoso, J. and Sheth, A. (eds.) (2006): *Semantic Web Services, Processes and Applications.* New York, NY: Springer.

Magkanaraki, A., Alexaki, S., Christophides, V. and Plexousakis, D. (2002): Benchmarking RDF schemas for the Semantic Web. Proc. *First International Semantic Web Conference on the Semantic Web*, Sardinia, Italy, **12**:132-146, Springer Berlin / Heidelberg.

Doerr, M. (2003): The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, **24**(3).

Doerr M., Hunter, J. and Lagoze C. (2003): Towards a Core Ontology for Information Integration. *Journal of Digital Information*, **4**(1):Article No. 169.

Sinclair, P., Addis, M., Choi, F., Doerr, M., Lewis P. and Martinez, K. (2006): The use of CRM Core in Multimedia Annotation. Proc 1st *First International Workshop on Semantic Web Annotations for Multimedia part of the 15th World Wide Web Conference (SWAMM 2006*), Edinburgh, Scotland.

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff M. (eds.) (2005): *Definition of the CIDOC Conceptual Reference Model. (June 2005)*, http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc

Reed, P. A. (1995): CIDOC Relational Data Model. A Guide. ICOM/CIDOC. http://www.willpowerinfo.myby.co.uk/cidoc/model/relational.model/ . Accessed 19 Sept 2007.

The CIDOC Conceptual Reference Model http://cidoc.ics.forth.gr/. Accessed 15 Aug. 2007.

Mylopoulos, J., Borgida, A., Jarke, M. and Koubarakis, M. (1990): Telos: Representing Knowledge about Information Systems, *ACM Transactions on Information Systems* **8**(4):325-362.

Analyti, A., Spyratos, N. and Constantopoulos, P. (1998): On the Semantics of a Semantic Network. *Fundamenta Informaticae* **36**(2-3):109-144.

Doerr, M., Schaller, K. and Theodoridou, M. (2004): Integration of complementary archaeological sources. Proc. Computer Applications and Quantitative Methods in Archaeology Conference (CAA2004), Prato, Italy, http://www.ics.forth.gr/isl/publications/paperlink/doerr3_caa2004.pdf. Accessed 15 Aug.2007.

IFLA Study Group on the functional requirements for bibliographic records: Functional Requirements for Bibliographic Records: Final Report", vol. 19. Of UBCIM Publications: New Series. K. G. Saur, Munich, 1998, http://www.ifla.org/VII/s13/frbr/frbr.htm.

LeBoeuf P. (eds.) (2005): Functional Requirements for Bibliographic Records (Frbr): Hype or Cure-All? Haworth Press, Inc, ISBN: 0789027984.