The CIDOC CRM and an
Integrated Approach to Semantic Interoperability

Martin Doerr

Decades of research have been devoted to the goal of creating systems which integrate information into a global knowledge network, yet we still face problems of cross-repository interoperability, lack of public infrastructure, and a coherent research agenda - both theoretical and practical – to face these challenges. Interest in the Semantic Web has revived the dream, but many are sceptical. The presentation will address semantic problems and requirements to integrate digital information into large scale, meaningful networks of knowledge that support not only access to source documents but also information use and reuse. We present a new approach based on (i) interdisciplinary research of scholarly and scientific discourse, (ii) a generic global ontological model based on relations and co-reference rather than objects, (iii) semi-automatic maintenance of co-reference links, and (iv) public engagement in the creation and development of the network.

We regard Digital Libraries or better *Digital Memories* as : information systems preserving and providing access to source material, scientific and scholarly information, such as libraries of publications, experimental data collections, scholarly and scientific encyclopedic or thematic databases or knowledge bases. There is still a widely held and traditional view of the task of libraries as institutions limited to the collection and preservation of documents and to providing assistance in finding specific items of literature or information. In this view, the library's role is completed when the (one, best) document is handed out: 'All you want is in this document.'

This view has not helped much in raising the level of new functionality that semantic interoperability of resources would permit. There is little or no support for the searches to produce new and informed responses from aggregated sources or to retrieve them by contexts (e.g. "Which excavation drawings show the finding of this object?"). There is little or no support to allow integration of complementary information in multiple sources into new insight (e.g., "What is known about the people who participated in this excavation"). Finally there is typically no support for cross-disciplinary search (e.g. to find relevant related information from the many disciplines that contribute to archaeological knowledge, such as ecology, ethnology, biodiversity, etc.). Central to our approach is a suitable knowledge management. We distinguish:

1. Core ontological relationships for "schema semantics", such as: "part-of","located at","used for", "made from" which are localized atomic relationships, but and rich in potential structural information, relating to content.
2. Categorical data": taxonomies used for reference to and agreement on sets of things, rather than as means of reasoning, such as: "basket ball shoe", "whiskey tumbler", "burmese cat", "terramycine". These terms define and order concepts rather than providing structural information. They aggregate categories as opposed to integrating sources. The leaves of the taxonomic structure would be entries in a thesaurus.
3. Factual background knowledge for reference and agreement as objects of discourse, such as particular persons, places, material and immaterial objects, events, periods, names. These would be elements of the taxonomic classes.

Global core ontologies play a fundamentally different role to that played by specialist terminologies in practical knowledge management. They are small and can be manually created. They support schema mediation, data transformation and migration.

The CIDOC Conceptual Reference Model is presented as an example of such a global model. It is a core ontology and new ISO standard (ISO 21127, accepted Sept. 2006) designed for the semantic integration of information from museums, libraries, and archives. It has been developed by CIDOC, the International Committee for Documentation of the International Council of Museums (ICOM), and an international multidisciplinary team of experts. 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. It is a product of re-engineering the dominant common meanings from the most characteristic schema elements in use in these institutions. It is not prescriptive, but provides a controlled language to describe common high-level semantics that allow for information integration at the schema level. This integration has been demonstrated in a large range of different domains including cultural heritage, e-science and biodiversity.

The CRM foresees domain-specific extensions, such as the integration of the conceptual model contained in the Functional Requirements for Bibliographic Records (FRBR), developed by the International Federation of Library Associations (IFLA), with the CIDOC CRM.

Whereas the second level, the “categorical” data, have been extensively treated by information science and the Semantic Web likewise, the third level of factual knowledge, which is orders of magnitude larger, is widely overseen as topic of semantic interoperability. There is a growing awareness of the need for information systems which provide reasoning capability, but before any “reasoning” can be done over integrated knowledge resources, the data must be connected in a “global network of knowledge.” This requires:

- A sufficiently generic global model as presented above
- Methods of knowledge extraction / data transformation to populate the network.
- Massive, distributed, semiautomatic detection of co-reference relations (data cleaning) across contexts.
- Referential integrity of co-referencing needs to be curated in order to create, maintain and improve the consistency of global networks of knowledge as a continuous process.

Further research is needed on co-referencing to make global information integration a reality. Advocating a global model does not mean using a common schema. The global model needs only to be used only on a virtual level in integrated information management. A common schema is counter productive and hinders evolution. A global core ontology is a question on agreeing on a common understanding of basic concepts. We argue that metadata for digital memories are based on a similar enough discourse for most domains. This is demonstrated in a few examples, extensions to the presented ontology not withstanding.