Modelling Intellectual Processes: The FRBR - CRM Harmonization

Martin Doerr, ICS-FORTH
Patrick LeBoeuf, Bibliothèque National de France

Even though the Dublin Core Metadata Element Set is well accepted as a general solution, it fails to describe more complex information assets and their cross-correlation. These include data from political history, history of arts and sciences, archaeology or observational data from natural history or geosciences. Therefore IFLA and ICOM are merging their core ontologies, an important step towards semantic interoperability of metadata schemata across all archives, libraries and museums. It opens new prospects for advanced global information integration services. The first draft of the combined model is published in June 2006.

1 Introduction

Semantic interoperability of Digital Libraries, Library- and Collection Management Systems requires compatibility of both the employed Knowledge Organization Systems (KOS; eg classification systems, terminologies, authority files, gazetteers) and of the employed data and metadata schemata. Currently, the notion and scope of Digital Libraries covers not only traditional publications, but also scientific and cultural heritage data. The difference between traditional publication in text form and structured data in form of databases is more and more blurring, with databases containing texts in XML form, texts and multimedia data being described by structured metadata records, and Natural Language Processing techniques extracting structured information from free texts. The grand vision is to see all these data integrated so that users are effectively supported in searching for and analyzing data across all domains. Even though the Dublin Core Metadata Element Set is well accepted as a general solution, it fails to describe more complex information assets and their cross-correlation. These include data from political history, history of arts and sciences, archaeology or observational data from natural history or geosciences etc.

Core ontologies describing the semantics of metadata schemata are the most effective tool to drive global schema and information integration [1], and provide a more robust, scalable solution than tailored ‘cross-walks’ between individual schemata. Information and queries are mapped to and from the core ontology, which serves as a virtual global schema and has the capability to integrate complementary information from more restricted schemata. Many scientists question the feasibility of such a global ontology across domains. On the other side, schemata like Dublin Core reveal the existence of overarching concepts. Ideally, the envisaged European Digital Library would be based on one sufficiently expressive core ontology, not by selection, but by harmonization and integration of the relevant alternatives. The challenge is to explore practically the limits of harmonizing conceptualizations from relevant domains.

2 The Harmonization Project

The CIDOC Conceptual Reference Model (CRM) has been developed since 1996 under the auspices of the International Committee on Documentation (CIDOC) of the International Council for Museums (ICOM) Documentation Standards Working Group. This is occurring with the initiative and support of ICS-FORTH, Heraklion, and the CRM is about to be accepted as ISO standard (currently ISO/DIS 21127) in 2006. It is a core ontology aiming to integrate cultural heritage information [2,6]. It already generalizes over most data structures used by highly diverse museum disciplines, archives, and site and monument records. Even the common library format MARC (‘MAchine Readable Cataloguing’) can

be adequately mapped to it. Its innovation is to centre descriptions not around the things, but around the events that connect people, material and immaterial things in space-time. Further, it explicitly describes the discourse on relations between identifiers and the identified, a powerful feature for the integration of information assets. Finally it bridges the role of typologies as classification systems with their nature as objects of the cultural-historical discourse.

Quite independently, the FRBR model ('Functional Requirements for Bibliographic Records') was designed as an entity-relationship model by a study group appointed by the International Federation of Library Associations and Institutions (IFLA) during the period 1991-1997. It was published in 1998. Its innovation is to cluster publications and other items around the notion of a common conceptual origin – the 'Work' in order to support information retrieval. It distinguishes four levels of abstraction from conception to the book in my hands: The Work, Expression, Manifestation, Item. Its focus is domain-independent and can be regarded as the most advanced formulation of library conceptualization [3,8].

Initial contacts in 2000 between the two communities eventually led to the formation in 2003 of the International Working Group on FRBR/CIDOC CRM Harmonisation. It is headed by Martin Doerr from ICS-FORTH and Patrick LeBoeuf from BNF Paris, and brings together representatives from both communities. The common goals are to express the IFLA FRBR model with the concepts, ontological methodology and notation conventions provided by the CIDOC CRM, and to merge the two object-oriented models thus obtained. This Working Group has published the first complete draft of FRBRoo, ie the object-oriented version of FRBR, harmonized with CIDOC CRM, in June 2006. This formal ontology is intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation and interchange of bibliographic and museum information.

3 Selected Results

The combined model on one side enriches the CIDOC CRM with notions of the stages of intellection creation and refines its model of identifiers and the associated discourse. On the other side, it makes available to FRBR the general model of historical events of the CRM. FRBR is not event-aware. As a consequence, many attributes are attached to entities they do not causally belong to, and the precise semantics remains unclear. E.g., it was a surprise that the date and place of publication is in reality not necessarily related to the event of printing of a book. The process of developing this model turned out to be very demanding. The intellectual rigour of the methodology of the CIDOC CRM demanded clarification and explication of many notions more vaguely specified in FRBR. After that, FRBRoo could completely be formulated as a specialization of the CRM, some smaller, upwards-compatible modifications of the CRM not withstanding.

But at the heart of the work, the major innovation is a realistic, explicit model of the intellectual creation process (see Figure 1), which should still be developed further in the future for the benefit of librarians and scholars from the various museum disciplines. FRBRoo makes the following distinctions:

- The substance of Work is the concepts or internal representations of our mind. The unity of a Work is given by the intellectual coherence of its concepts. Work can be created by multiple people together, and be understood and continued by other people, such as by translation, derivation, completion. A stage or part of a Work is regarded as an Individual Work, if it is complete from its elaboration and logical coherence of its content, or regarded as a complete unit by its author.
- The substance of Expression is signs or symbols. It is only representation. It has no direct intellectual qualities, but humans can interpret the signs and recognize the Work behind.

Consequently, an Expression cannot be translated, but only be used to translate the Work it represents. Expressions can be complete in the sense, that they represent an Individual Work. Then they are regarded as "Self-Contained". Else they are fragments.

- "Manifestation" can be two completely different things: Either it is an industrial product, i.e., a Type, like a particular car model, or it is a Physical Man-Made Thing that was produced as a unique carrier of an Expression. Industrially printed books belong to the first category, and are indirectly related to the main author's original creations.

The idea is that products of our mind, as long as they stay in one person's mind only, are relatively volatile and not evident. Even though a person may claim having conceived a Work at a certain date, it is not before the Work is **"externalized"** for the first time that its creation becomes evident. Further, we all have experienced how thought takes shape during communicating it to others. Therefore we basically tie the intellectual creation with the event of "first externalization", the Expression Creation. In practical terms, externalization means that the expression must be transferred to another physical carrier. This can be just another person's memory, as in the case of oral tradition (the CRM regards persons also as physical objects), or more usually a paper manuscript or, in these days, a computer disc. A good example are jokes, which have a recognizable identity and may go around the world once uttered to another person.

The transfer to another carrier is a physical process, which leaves more or less material traces. In terms of documentation, we would normally regard that a manuscript is produced as a new object. However, if we do not use raw writing material, but scribble a text on a wall, the object is rather modified than produced. One may argue that a new Physical Feature is produced. In a sense, creating a file on a computer disc or even a memory in a human mind might be seen as a physical modification, but this may not be of practical use from a documentation point of view. For FRBR, we regard a production of a manuscript or visible Physical Feature as the relevant case (see figure). The CRM allows for combining Creation (E65) of immaterial items (such as Expressions) with Production (E12) to model the FRBRoo concept Expression Creation. However, E12 currently does not apply to Physical Features, probably a good reason to extend the CRM here. To our knowledge, this is the first time that the material and immaterial aspects of intellectual creation are modelled explicitly. Also, explicitly modelling oral tradition may be worthwhile doing.

Another important part of the discussion had to do with work containing other work, such as collections of poems. In the course of discussion however it was recognized, that virtually any book is composed of multiple, distinct works: the text, the illustrations, the editors work on lay-out, type phase etc. The latter was widely ignored in FRBR, and discussions tend to confuse the question of which contribution and work is most relevant with how to make the necessary distinctions in a model. This situation demanded for a general model explicating both the individual contribution and the unity of the integrated product. The solution provided regards that the containment happens at the Expression level, the signs. I.e., in a collection of poems, the final Expression is both, a representation of the collector's work, and of the collected works. This does not make the collection work contain other work, nor is the Expression necessarily separable into the different contributions: If all poems are cut out, the collection is not properly expressed by the – potentially empty – rest.

Finally, library practice has a lot to do with complex identifiers with meaningful parts. The CRM will benefit from an explicit model of parts of an identifier, so far ignored by the CRM.

4 Conclusions and Future Work

The work covered so far the FRBR Entities and Attributes. Whereas FRBR promoters claim that the model is applicable to any intellectual production process, the argumentation was deliberately restricted to written material in order to avoid over-generalizing from the very beginning notions well understood in the normal library context. However, a good equivalence of material publishing to electronic publishing could already be established, the basic idea being that the individual copy of a file on a particular machine corresponds to the creation of an Item, such as a book in my hands. The equivalences to performing arts seem to be so debatable, that a particular discussion round will be devoted to this topic. Work will further continue with modelling the FRBR Relationships, and authority records (the so-called FRAR, "Functional Requirements for Authority Records").

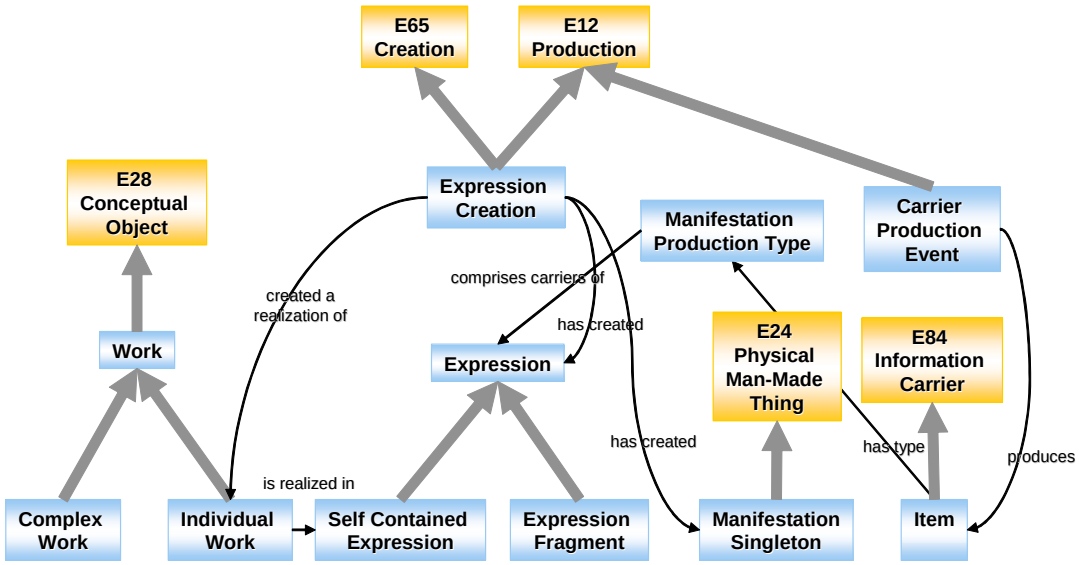

Figure: Partial model of the intellectual creation process.

The potential impact of the combined models can be very high. The domains explicitly covered by the models are already immense. Further, they seem to be applicable to the experimental and observational scientific record for e-science applications. From a methodological perspective, the endeavour of core ontology harmonization experimentally proves the feasibility of finding viable common conceptual grounds even if the initial conceptualizations seem incompatible [3,4]. Even though this process is intellectually demanding and time-consuming, we hope the tremendous benefits of nearly global models will encourage more integration work on the core-ontology level. A recent practical application of these models is the derivation of the CRM Core Metadata schema [5,7], which is compatible and similar in coverage and complexity to Dublin Core, but much more powerful. It allows for a minimal description of complex processes, scientific and archaeological data, and is widely extensible in a consistent way by the CRM-FRBR concepts. CRM Core can be easily used by Digital Libraries.

[1] Manjula Patel, Traugott Koch, Martin Doerr, Chrisa Tsinaraki, Nektarios Gioldasis, Koraljka Golub and Doug Tudhope, "Semantic Interoperability in Digital Library Systems", DELOS Network of Excellence on Digital Libraries – deliverable 5.3.1, June 2005.

[2] Doerr, M., "The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata", AI Magazine, 4(1), 2003.

[3] Patrick LeBoeuf (Editor), "Functional Requirements for Bibliographic Records (Frbr): Hype or Cure-All?", Haworth Press, Inc, January 2005, ISBN: 0789027984

[4] Martin Doerr, J. Hunter , Carl Lagoze, "Towards a Core Ontology for Information Integration", 2003, In *Journal of Digital information*, volume 4 issue 1, April 2003

[5] Patrick Sinclair, Matthew Addis, Freddy Choi, Martin Doerr, Paul Lewis, Kirk Martinez, "The use of CRM Core in Multimedia Annotation", to appear in: Proceedings of First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006), part of the 15th World Wide Web Conference (22-26 May 2006, Edinburgh, Scotland

[6] Definition of the CIDOC CRM: http://cidoc.ics.forth.gr.

[7] Definition of CRM Core: http://cidoc.ics.forth.gr/working_editions_cidoc.html

[8] IFLA Study Group on the functional requirements for bibliographic records. "Functional requirements for bibliographic records : final report", volume 19 of UBCIM Publications : New Series. K. G. Saur, Munich, 1998.

Links:
IFLA: http://www.ifla.org
ICOM: http://icom.museum
Definition of the CIDOC CRM: http://cidoc.ics.forth.gr.
Definition of CRM Core: http://cidoc.ics.forth.gr/working_editions_cidoc.html
Definition of FRBR: http://www.ifla.org/VII/s13/frbr/frbr.htm
DELOS NoE deliverable 5.3.1: http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs

Please contact:
Dr. Martin Doerr
Center for Cultural Informatics, Institute of Computer Science,
Foundation for Research and Technology - Hellas (FORTH)
Tel: +30(2810)391625
E-mail: martin@ics.forth.gr