Knowledge Representation seminar Meeting #2

Michele Pasin
Kings College, London
June 2010

Some key concepts in Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Magazine archive (2003) vol. 24 (3) pp. 75-92

The context..

The creation of the **World Wide Web** has had a profound impact on the ease with which information can be distributed and presented. In the sequence, there has been an increasing interest from professionals, public and politicians to **make publicly available** the tremendous wealth of information kept in **museums**, **archives** and **libraries**, the so-called "**memory organisations**". Quite naturally, their development has focussed on **individual presentations** – **websites** and **interfaces** to their **local databases**.

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 contents. This requires **interoperability not only on the encoding level** - a task solved well by XML for instance - **but also of the complex semantics**, which are characteristic for this domain.

Data capturing VS interpreting data

We make the hypothesis that much of the diversity of data and metadata structures is due to the fact, that they are **designed for data capturing** – as guide for good practice of what should be documented, and to optimise coding and storage costs for a specific application, **by far more than for interpreting data**.

Necessarily, these data structures are relatively flat (in order to suggest a workflow of entering data to the user) and full of application specific hidden constants and simplifications.

Read-only integration

Instead of seeking any more a common schema as prescription for data capturing, which would be supposed to ensure semantic compatibility of the produced data, we have followed with the CIDOC CRM what Bergamaschi et.al [8] call in the meanwhile a **semantic approach to integrated access**.

Being convinced that collecting information is already done well by the existing data structures – possible improvements not withstanding, **we aim only at read-only integration**, as e.g. in the DWQ project [9][10]. This comprises **data migration**, **data merging** (materialized data integration as in data warehouse applications), and **virtual integration via query mediation** [11].

Recently more and more projects and theoreticians support the use of formal ontologies as common conceptual schema for information integration

Property-centric ontology

Only classes were introduced in the CRM that are **domains or ranges for the relevant relationships**, such that any other ontological refinement of the classes can be done as additional "terminological distinction" without interfering with the system of relationships (see also [2]).

Such a conceptual model seems to cover the ontological top-level automatically, and provides an integrating framework for the often isolated hierarchies found in terminological systems (Fig. 2). We call such a model a "**property centric ontology**" to stress the specific character and functionality.

Property-centric ontology (II)

[...] we require **semantic uniqueness of properties**: no two properties with the same semantics should be declared for two different domains, e.g. a property *has age* from people to number of years and a property *life-span* from animals to number of years should be merged and assigned to one appropriate domain (living beings), often introducing multiple inheritance.
In practice, this principle turned out to be quite powerful to detect concepts of wide validity and to make the model stable under further extensions.

After a short time of development, we had evidence in 1996 for a set of base classes like: **Temporal Entities, Actors, Physical Objects, Conceptual Objects, Place, Time**, similar to Ranganathan's Fundamental Categories [32], **not so much from philosophical insight, but as appropriate domains for properties found in the source models**.

Property-centric ontology (III)

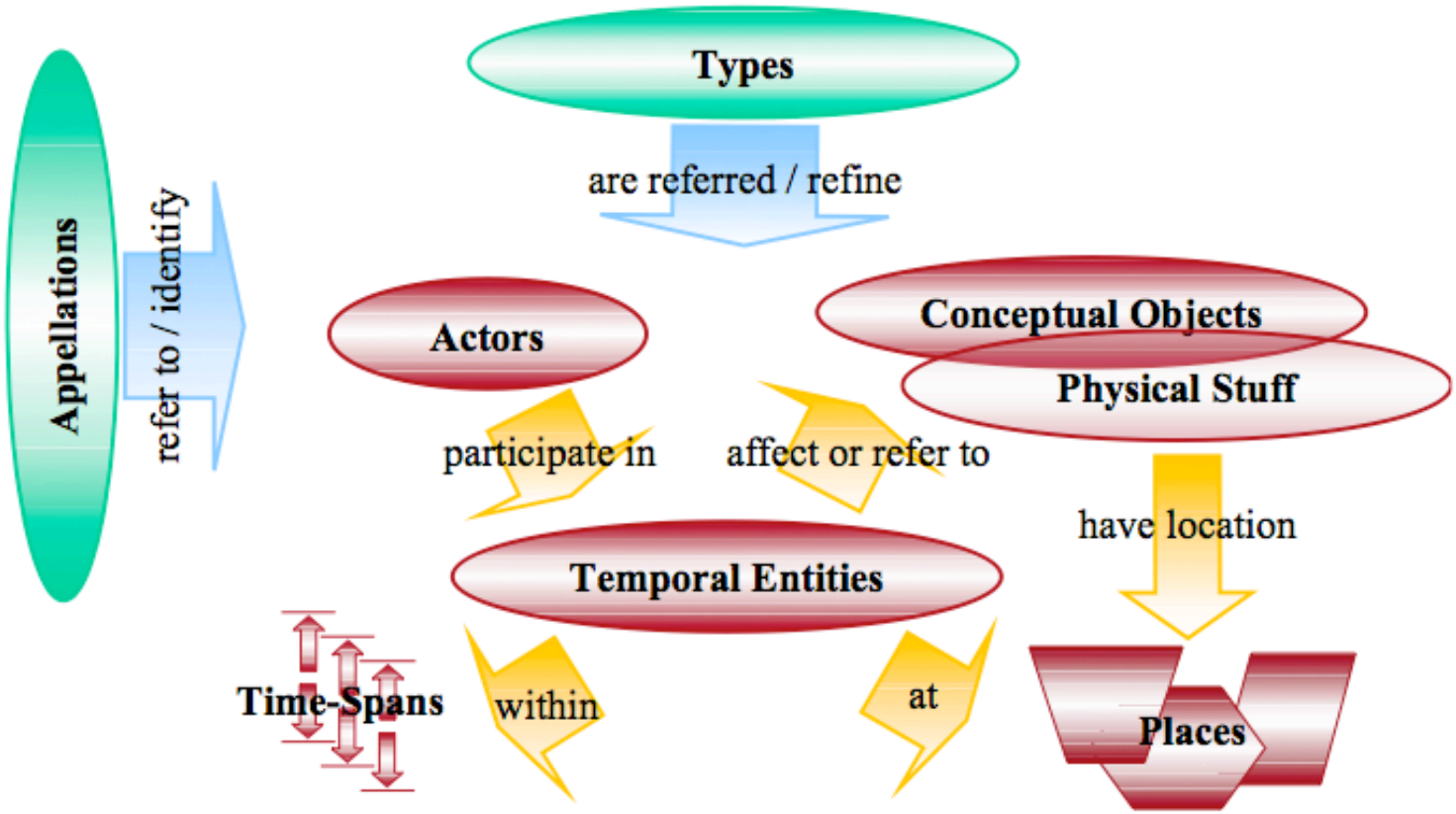

Figure 3: A qualitative metaschema of the CIDOC CRM

Property-centric ontology (IV)

Step 1 : Create the list of properties

Step 2 : Detect new classes from attribute values.

Step 3: Detect entities hidden in attributes.

Step 4 : Property consistency test.

Step 5 : Create the class hierarchy.-

Step 6 : Create property hierarchies.

Step 7 : Closing up the model.

The nature of cultural-historical knowledge

Another contribution of this work are considerations about the specific nature of cultural – historical knowledge and reasoning, which **aims at the reconstruction of possible past worlds from loosely correlated records** *rather than at control and prediction of systems, as in engineering knowledge*.

"Historical" must be understood in the widest sense, be it cultural, political, archaeological, medical records, managerial records of enterprises, records of scientific experiments, criminalistic or jurisdictional data.

The nature of cultural-historical knowledge (II)

Historical data - it means any description done in the past about the past, be it scientific, medical or cultural – **is normally unique and cannot be verified, falsified nor completed in an absolute sense**.

In history, any conflict resolution of contradictory records is nothing more than **yet another opinion**.

So, for an ontology capable to collect and relate knowledge from historical data, **ontological** principles about how we perceive things must related, and **epistemological** principles about how knowledge can be acquired must both be respected. There can be huge differences in the credibility of propositions.

Necessity of data structure diversity

Let us regard here data structures for long-term storing of data, as database schemata, but also tagging schemes like SGML/XML DTD, RDF Schema, and data structures designed as fill-in forms guiding users to a complete and consistent documentation [...]

These structures are always are **compromise** between the ***complexity of the information*** one would like to make accessible by formal queries, the ***complexity the user can handle***, the ***complexity of the system*** the user can afford to implement or to pay for, ***the cost to learn those structures*** and to fill them with contents.

As most applications run in a relatively uniform environment – a library, a museum of modern art, a historical archive of administrational records, a paleontological museum etc., much of the complexity of the one application is negligible for another. **This allows for a variety of simplifications, which are unavoidable** to create efficient applications.

CRM principles and methodology

The problems computer scientists and system implementers have to comprehend the logic of cultural concepts seems to be equally notorious as the inability of the cultural professionals to communicate those to computer scientists.
The CIDOC CRM working group is therefore **interdisciplinary**, **aiming at closing that gap**.

- ***clarity***
- ***extensibility***
- ***coherence***
- ***context-free interpretation (?)***
- ***alternative views***
- ***appropriate granularity***
- ***minimal ontological commitment***

Monotonicity

Gruber [13] writes: “An ontology…should offer a conceptual foundation for a range of anticipated tasks, and **the representation should be crafted so that one can extend and specialize the ontology monotonically**”, i.e. **without revision of the existing definitions**. Only if this holds, one can propagate an extensible ontology as an ISO standard. To our opinion this is both, an *ontological* and an *epistemological principle*:

From the *epistemological* side, monotonicity under the addition of knowledge which is not in contradiction to previous one, is needed on both, the categorical and the factual level; else the integration of facts as they come in over time becomes a non-scalable task. *On the other side*, it is an ontological question, because the notion of what is in contradiction and what is not, is grounded in the domain conceptualisation

Monotonicity can be regarded at least on three levels, ***classification***, ***attribution*** (properties), and ***modelling constructs***