

Electronic esperanto – The Role of the oo CIDOC Reference Model

*Martin Doerr,
Foundation for Research and Technology - Hellas (FORTH), Greece
Nicholas Crofts,
Direction des Systèmes d'Information (DSI), Geneva, Switzerland*

Abstract

Over the past few years, the ICOM/CIDOC document standards group has been developing an object oriented Conceptual Reference Model. The model represents an 'ontology' for cultural heritage information, i.e., it describes in a formal language the explicit and implicit concepts relations relevant to the documentation of cultural heritage. The primary role of the CRM is to serve as a basis for mediation of cultural information and thereby provides the semantic 'glue' needed to today's disparate, localized information sources into a coherent valuable global resource. The model provides mechanisms for dealing a number of complex issues in a coherent manner: varying levels of and precision, transfer of information between 'richer' and 'poorer' systems and extensions to incorporate domain specific information. This explains how the model can be used as a reference in the cultural sector. is intended both to represent good practice in the representation information and to be used as a practical aid in the design implementation of mediation servers, search engines, databases, DTD's Z39.50 access profiles, Metadata, documentation guidelines, and products. using the CRM ensures semantic compatibility between systems and services and removes the need for one to one conversions different native formats. The paper should be of interest to curators other domain specialists, as well as system designers and implementors working in the cultural domain.

1 Introduction

The creation of the World Wide Web has had a profound impact on the ease with which information can be distributed and presented. Museums have been relatively quick to take advantage of the new technology and many now manage their own web sites. However, many of these sites are little more than electronic versions of tourist brochures and offer only a tantalising glimpse of the resources available. At present, few museums make the effort to tap into their information systems and still fewer to integrate their information with that from other institutions. Today's web sites are still predominantly hand-coded productions. The results can be very attractive, but the effort involved in producing and managing a hand-made web site imposes severe restrictions on the level of complexity that can be sustained.

Many writers have evoked the vision of the web as a global resource for cultural heritage information. In order to achieve this vision, museums will have to establish solid and reliable means for integrating and distributing the rich and detailed documentation contained in their information systems.

A major barrier to such integration is the semantic and structural incompatibility of existing systems. Different institutions organise and present the data they use in different ways. Differences may be limited to the naming and arrangement of entities and fields, but they may also affect the level of depth and detail of analysis, or even the entire focus and orientation of the data. Even if the structures are compatible, terminology is often incompatible. To date, most attempts to bridge the gaps between incompatible information systems have been based on hermetic, ad hoc transformation rules, or have resorted to massive simplification, concentrating on a limited subset of 'core' data.

The CIDOC Reference Model (in the following "CRM") aims to overcome this limitation by providing a common semantic reference point, a formal expression of the basic concepts behind the structure of the various data we wish to communicate. It will enable museums to render their information resources mutually compatible without sacrificing detail and precision. To this end the model is presented as an object-oriented semantic model, a "domain ontology", which allows for a great deal of flexibility both in the level of detail which is required and in terms of extensibility.

Ultimately, we hope that the CIDOC model will serve as a basis for the mediation of cultural heritage information and thereby provide the 'glue' needed to transform today's disparate, localised information sources into a coherent and valuable global resource.

After a discussion of museums' communication needs, the present paper gives an overview of the state of heterogeneous data access in the cultural field and other domains before presenting the principles features of the CIDOC reference model along with an introduction to its basic entities. Some illustrations are given as examples of the model's application. The paper concludes with some ideas for future development.

2 Background to the model

Work on the CRM began in March 1996 following a meeting in Crete, hosted by ICS-FORTH. Up to that time CIDOC had maintained a E-R data schema, inspired largely by work done at the Smithsonian, which was intended to fulfil much the same role as the current CRM. However, the need to encompass a sufficiently broad scope of information and domains had resulted in a highly complex and unwieldy model which was proving difficult to maintain. Furthermore, the model suffered from a significant bias towards the fine arts and historical collections – support for the natural sciences, archaeology and ethnography was inadequate. Extension of the already complex model to incorporate further information categories was becoming increasingly difficult. At the Crete meeting, the CIDOC Documentation Standards Working Group decided to adopt an oo approach and develop a new data schema, derived initially from the information categories contained in the Relational Model and from a separate document – the “International Guidelines for Museum Object Information: The CIDOC Information Categories” [CIDOC95] (Hereafter "IC"). The first version of this new model was presented at the triennial ICOM conference in Melbourne in 1998 and is currently being evaluated by ISO (International Standards Organisation) as a potential standard. The model, and associated documentation is available via the web [CIDOC98].

The decision to adopt oo modelling techniques was motivated by a number of factors.

- The oo data model is semantically richer than the E-R model. All E-R modelling constructs find equivalents in the oo model, but the reverse is not the case. Although a straight forward mechanical translation proved to be inadequate, this enabled the working group to translate the primary aspects of the existing E-R schema into an oo schema, and to simplify many redundant constructs in the process.
- Through the mechanism of specialisation, the oo data model is more readily extensible than an E-R model and therefore easier to maintain.
- The specialisation and aggregation of classes provides a means for presentation of variable levels of granularity. This both helps to conceal complexity and unnecessary detail and makes the model more flexible and adaptable.
- Finally, both theory and practice have shown that adopting an object oriented reference model does not necessarily *require* the use of an object-oriented database for implementation. Although they present some drawbacks, mainstream relational database engines can be used for implementation of object oriented schema [Crofts99].

3 What the model is for

While the CRM can be used as the basis for implementation of cultural information systems, we see the *primary* role of the reference model as being to define a semantic framework which will enable compatible systems to exchange and share information¹. For CIDOC, this represents a significant paradigm shift away from the assumption that integration of information requires homogeneous data sources.

Many formats are currently available which allow relatively simple, unambiguous exchange of data; however, the meaning of these data, their scope and application, is often far from obvious. Oversimplification of structure results in a need to “stretch” the meaning of structural elements, and thereby introduces a level of ambiguity which renders the contents incompatible. The OO reference model provides a means for defining the semantic values of data structures with the precision needed to ensure reliable communication and mediation of cultural information.

¹ Information exchange includes issuing queries over the net and receiving answers from heterogeneous sources.

3.1 Communication needs

Access to museum documentation, presented in an appropriate manner, has the potential to interest a wide audience: researchers, educational institutions, and the general public. In each case it is important that the information presented should be *integrated* with other sources. The value of information is generally enhanced when it is put in relation with other pieces of information. This is particularly evident with respect to cultural heritage. Descriptions of individual objects are, in themselves, of only limited interest. Additional references to other objects, and to an object's historical, geographical, and cultural origins help to place it in a context and give it meaning. Typically, the contextual information, which can help bring collections to life, is distributed across several institutions. Without some form of interaction between the different information systems, much of the potential interest of the collections is lost.

To illustrate the value of cross collection links it is worth looking at a simple example of juxtaposition of works from a number of collections. The tower of Babel was a theme which clearly fascinated the Breugels since they executed a number of versions of the subject, the best known of which are the Tower of Babel in the Kunsthistorisches Museum, Vienna and the "Little" Tower of Babel (1563) in the Boymans-van Beuningen Museum, Rotterdam. However, other versions exist and they have been reunited on the web by an enterprising student of art history². This web page has no ambition other than to bring together a number of illustrations for the purpose of comparison, and only minimal textual commentary is provided. However, the pedagogical value of even this rudimentary approach is obvious. Differences of detail are thrown into relief and it becomes possible to detect a thread in the evolution of the Breugels' treatment of the subject. The precise date of execution of each work becomes highly significant since we instinctively want to arrange the images in chronological order.

It is significant, though, that this page was *not* created by a museum - each illustration comes from a different institution, none of which has direct access to information from the others. The information systems of the world's museums are a potential gold mine if they can be made to work together. At present, however, the technical problems involved in producing web pages such as this automatically are practically insurmountable.

Presentation of information is another area where current efforts are generally inadequate. Many institutions present only a small selection of their collections with no little or no indication of the extent and nature of the rest. Others adopt an 'inventory' approach based on exhaustive and often cryptic lists of objects. Few sites attempt to integrate information about objects with contextual information about people, places and events³.

Different forms of presentation can be imagined to meet different requirements. Statistical analysis and in depth research obviously require systematic and precise query facilities which can generate exhaustive lists of items. But this kind of approach is inappropriate for general interest browsing and education which would most likely prefer a far less 'technical' presentation with more textual commentary, and some form of guidance to help find a pathway through the available material. There is little use in offering novice users the possibility of typing in search criteria if they are unfamiliar with the subject matter and the content of the collections. These different requirements imply different interface designs, which presuppose different levels of knowledge in the subject matter. Both, however, depend on mechanisms capable of integrating information from different sources.

The challenge of integrating information from different sources and providing well adapted access goes far beyond the question of homogeneous data formatting. The European Community has declared the integration of museum, archive and library information as a current strategic research and development goal. Different disciplines, such as natural history, fine arts, and ethnography, as well as different types of collections - museum information systems, archives, and libraries - provide complementary information and viewpoints. Their combination, rather than their compilation, has the potential to provide new insights into our cultural heritage.

Combining and integrating data in a meaningful way, so that subject matter can be readily identified, requires more advanced mechanisms than are needed for straight forward compilation. It is worth considering a few examples of the divergent information needs of different domains. Ethnography, for example, is typically less interested in the identity of the individual creator of an object than the fine arts, whilst for natural history, the notion of 'author' applies only to the classification system and not to the objects being collected. Archaeologists

² <http://www.cwd.co.uk/babel/bruegel.htm>

³ Some of the major exceptions to this rule are not in fact museums, but sites run by individuals e.g. the WebMuseum <http://www.fhi-berlin.mpg.de/wm/> and CGFA <http://sunsite.unc.edu/cjackson/fineart.htm>.

and palaeontologists habitually deal with fragmented objects, which are then combined, with luck, into a single whole - a process that is highly unusual in other domains. Multiple fragments need to be identified and tracked during the entire process. For historical disciplines, much information is of a hypothetical nature and therefore needs to be 'signed' as an opinion by the author whereas uncertainty about, say, the author of a book is rare, and multiple attributions do not need to be dealt with. We could go on. The point is that information and levels of detail that are essential to one discipline may be unnecessary or even incomprehensible to another.

In the past, attempts to apply a single, homogeneous data structure to multiple disciplines have foundered on the lack of a discipline neutral viewpoint. The fact that librarians do not store information about the *attribution* of books is not due to an oversight - it would be counterproductive and confusing to do so since, unlike art history, authorship is seldom a contentious issue. Domain specific assumptions and presuppositions about the semantic value of data need to be respected. Applying data structures from one discipline to another leads to unhappy consequences: saying that the 'author' of a fossil specimen is 'unknown', for example, is not simply unclear, it is actually misleading.

In our view, combining information from different sources requires a high level of abstraction and a *discipline neutral* viewpoint, which has the flexibility for different viewpoints to be respected and expressed. This generic level of abstraction is precisely what the CRM aims to provide.

3.2 About Mediation

The recent past has seen several interesting and advanced projects for heterogeneous information access in the cultural area, which gradually provide more and more complex functionality. Other domains, which enjoy more robust economic circumstances, have already implemented solutions, based on "mediation" techniques, which demonstrate the feasibility of effective and rich communication without homogeneous data sources. It is worth passing in review some of the more prominent cultural information access projects which are based on this line of technological development.

3.2.1 RAMA, CHIO and AQUARELLE

Between 1992 and 1995, the RAMA project successfully demonstrated that large heterogeneous databases of museum objects in different countries can be accessed one after another using a uniform user interface. The project solved the problem of the physical connection protocol to multiple databases and the transfer of images to local workspaces, but the conceptual structure of the individual sources is presented unaltered to the user, which prevents further automatic processing. In 1994, the CIMI Consortium initiated the CHIO project, with a strong focus on structured text marked-up in SGML, retrieval using the Z39.50 protocol derived from the library community, and on open standards in general. The basic idea underlying the project was that SGML tagging makes texts accessible to far more precise questions and that a standard retrieval protocol allows access to a vast range of data sources. CHIO resolves the problem of divergent data formats by adopting a 'profile' - a **standardised** set of mark-up tags and Z39.50 access points. The freedom allowed by Z39.50 for the identification of access points to entities in target systems resolves some of the problems of semantic heterogeneity. A great deal of effort has gone into identifying core information and typical user questions although, of necessity, this approach has tended to focus on one viewpoint - that of the museum visitor.

In 1996, the **AQUARELLE** project, funded by the European Commission, took these ideas a stage further and focussed on the interests of professionals in the cultural field: museum curators, urban planners, commercial publishers and researchers, as well as allowing for greater semantic flexibility. Like CHIO, Aquarelle relies on CIMI standards, SGML, HTML, Z39.50, and HTTP. Its major innovations are the **dynamic handling of multiple DTDs**, the use of multilingual thesauri as search aids, [Doer98]. and the central **link manager**, which guarantees referential integrity for hyperlinks over the net.

Many AQUARELLE users work for public bodies concerned with the administration of material cultural, immobile sites in particular. Their need for precise information had a strong impact on the project and taught important lessons for future developments. Their evaluation of the services offered confirmed the importance and feasibility of handling heterogeneous data. It further demonstrated that the success of more advanced systems is only partially dependant on *technical* issues, the major problems are *semantic* in nature - formalising the structures, vocabularies and access points needed for queries. Well-informed and open-minded interdisciplinary teams are needed to deal with these questions [Guar98]. The project has proved an excellent forum for such discussion.

Another interesting project is GRASP. Its focus on the problem of identification of stolen objects allows access by transformation of heterogeneous structure to one fixed format. The project has highlighted the problem of incompatible terminology used in analogous data fields. Consequently, the project has had to invest considerably effort in dealing with questions of terminology. It is a striking demonstration of the fact that precise information retrieval from heterogeneous sources is only possible once semantic issues have been resolved. (Incidentally, the notion "ontology" used in GRASP for terminology resources should not be confused with our use in this paper.)

3.2.2 "Intelligent" services

All the systems so far mentioned use a "3-tier architecture", where a central application server acts as an interface between databases and remote clients. The translation of queries and data is done either locally, by each database, (as for Z39.50 gateways) or by the central service, or by both. Currently, these systems suffer from two severe restrictions:

- 1) The translations are disparate, idiosyncratic and "hard-wired". Consequently, with the exception of the terminology services used by AQUARELLE and GRASP, they cannot be maintained by a domain expert.
- 2) All information is presented in an entirely "object centric" fashion. Information about persons, places, events etc., can only be obtained indirectly. This is due in part to a shortcoming of Z39.50, which does not allow the kind of target object to be specified, although the use of multiple virtual gateways for different types of target could bypass this restriction. But it is also due to the inability of the application servers to analyse the data objects in the information sources.

To overcome such restrictions, Wiederhold [Wied92] introduced the notion of "mediation services". This approach has since been successfully implemented in a number of different systems in other domains (e.g. [Chaw94], [Subr94], [Baya96]). In his terms, "... **mediation** covers a wide variety of functions that enhance stored data prior to their use in an application. Mediation makes an interface intelligent by dealing with representation and abstraction problems ... Mediators have an active role. They contain knowledge structures to drive transformations". They have to be maintained by domain specialists. Major functions are:

- Transformation of databases using view definitions.
- Methods to access and merge data from multiple databases
- Abstraction and generalisation of underlying data
- Handling of information that is incomplete or at different levels of detail or abstraction
- Methods to integrate information from structured texts
- Maintenance of derived data

A mediator is a software module that exploits encoded knowledge about certain sets or subsets of data to create information for a higher layer of applications. This knowledge is stored in a knowledge base, referred to in recent literature as an "**ontology**" ([Kash97], [Guar98]). It describes in some formal language the entities of a domain of discourse and their relations, and their correspondence with expected data items and notions used for retrieval, in a way which can be understood by a domain specialist and can be accessed by interpretation software. To date, and without exception, all ontologies are formulated in some object-oriented paradigm, with a preference for semantic models. (We do not use the term "ontology" for thesauri, as sometimes found in literature.) Real systems still vary widely in the ease of integration of new sources, semantic capabilities and quality of service.

In order to integrate a new source into an information access environment, the schema or structure of the source is related - "**mapped**" - by simple declarations, to the ontology (rather than to notions of the various applications). The mediator "knows" by itself how to reshuffle data between fields and entities, rename fields, call translation functions for values, follow paths over multiple sources to find values, and reformulate queries etc., in order to execute a request such as a query or data transfer. Furthermore, the mediator contains "metadata" about the capabilities of each attached source, in order to determine which source can answer a question, by which mechanism and in what way: precise, approximate, incomplete or probabilistic.

A particular added value lies in the possibility of assembling *new* information objects from complementary data in different sources: the goal underlying the European Commission's declaration in Vth Framework, to focus on the connection of museum, archive and library data. This can only be achieved on a wide scale by the use of mediation techniques.

Obviously, the richness of the ontology ultimately determines the mediation capabilities. In some cases, only approximations to wider or narrower concepts can be made, or one must derive or "guess" missing values. In particular, in the cultural domain, terminology used in data fields is tightly related with structure. This must be

reflected in the ontology (see below). However, the value of a formal ontology goes beyond its use in mediation systems as it can also serve as a blue-print for information system implementation and as an intellectual guide for good practice in the development of information systems.

To summarise, we are on the brink of a technological revolution, which will render obsolete the need for homogeneous data formats for communication. Rather, we must engage in providing formal definitions of the underlying semantics in our data. The need for semantic compatibility goes beyond the superficial identity of structure. This will enable far richer services to be created than standardisation of structure could ever provide. The effort of CIDOC to define an object-oriented Conceptual Reference Model is both timely and appropriate since the currently adopted formalism conforms with that used in the emerging field of semantic integration systems.

4 A Conceptual Reference Model

The CRM represents an *ontology* in the sense of computer science [Guar98], i.e. an approximation of a conceptualisation of a domain in a formal language and a vocabulary⁴. In other words, we try to capture, in a consistent logical framework, the overt or implicit concepts which the museum community typically works with and agrees upon. (For more information on ontological principals see. e.g. [Guar98b].) This framework is designed to promote the creation of high quality information systems for the museum and cultural community which are either developed according to an ontology or actively "ontology driven", and in particular, to enable communication between heterogeneous but semantically overlapping systems, as outlined in chapter 3. In the following, we justify the major organisation principles of this model by simple examples and discuss development strategies and examples of use. The examples may be debatable. Our intention here is to demonstrate the principles involved rather than the contents.

4.1 Principles

We anticipate that differences will arise in the presentation of identical semantic contents due to the different purposes and points of view of individual systems. A reference model must adopt a well-defined "neutral" position, which implies a number of structural principles described below. This leads quite naturally to an object-oriented paradigm. A set of naming conventions is also adopted in order to assist the reader and to facilitate the unambiguous identification of parts of the model.

4.1.1 Symmetry

Let us assume that an object is sold from one museum to another. In accordance with the CIDOC Information Categories (IC) both institutions document this event. Even though they describe the same action, the obvious identity of "deaccession" and "sale" on the one hand and "acquisition" and "purchase" on the other, is unintelligible to a computer and cannot be automatically combined into one. We therefore "normalise" this information as an "Acquisition" action, which refers to two "Actors", one who surrenders the legal title, and one who acquires the title (see fig. 1). Acquisition is thus defined as the "transfer of the legal title to an object". This view is "institution neutral", a necessary precaution when querying some hundreds of databases over the net, which would result in retrieving identical information from a number of organisations. Incidentally, this approach is not incompatible with the IC; it is just another view of the same information.

Note that information about the documenting organisation has been made *explicit* in order to achieve symmetry. Note further that the object acquires a new inventory number, hence the description is different. Nevertheless, the model regards both instances as identical because the object referred to is identical. In contrast with the Relational model, this notion of object identity, independent of temporary changes in description, is a key concept of object orientation [Atki89], [Kim90]. Obviously a mediation system must contain specific operators in order to establish which incoming data possibly refer to the same item, which is not always possible. In our example it is based on the registration of the previous inventory number.

Let us now suppose that someone is interested in the actors involved, rather than the transactions. In this case, he or she would like to see the transaction as an attribute of the actor, rather than vice versa, or even as an attribute of the object, as in the IC. Therefore we model these references as symmetric, directed links, in the manner of semantic networks or conceptual graphs, between entities without internal information. Links carry two labels, one for each direction in which they can be read (e.g. "transferred title from" inverts as "surrendered title of", as

⁴ This sense is derived from, but not to be confused with that used in philosophy.

shown in fig 2.). Implementers of database schemata can choose the reading which is most appropriate for their viewpoint, and transform links into attributes, fields or references. We decided to avoid the cryptic naming practice of many computer programmers and name links in verb form, originating from a grammatical subject and pointing to a grammatical object. For all historical information we use the past tense, whereas for states we use the present tense.

Summarising, the symmetry principle allows us:

- to establish if apparently different information is in fact identical, but has been documented from the point of view one of the different entities involved;
- to transform the view from any entity involved into a view from another one;
- to derive view-specific, compatible information systems.

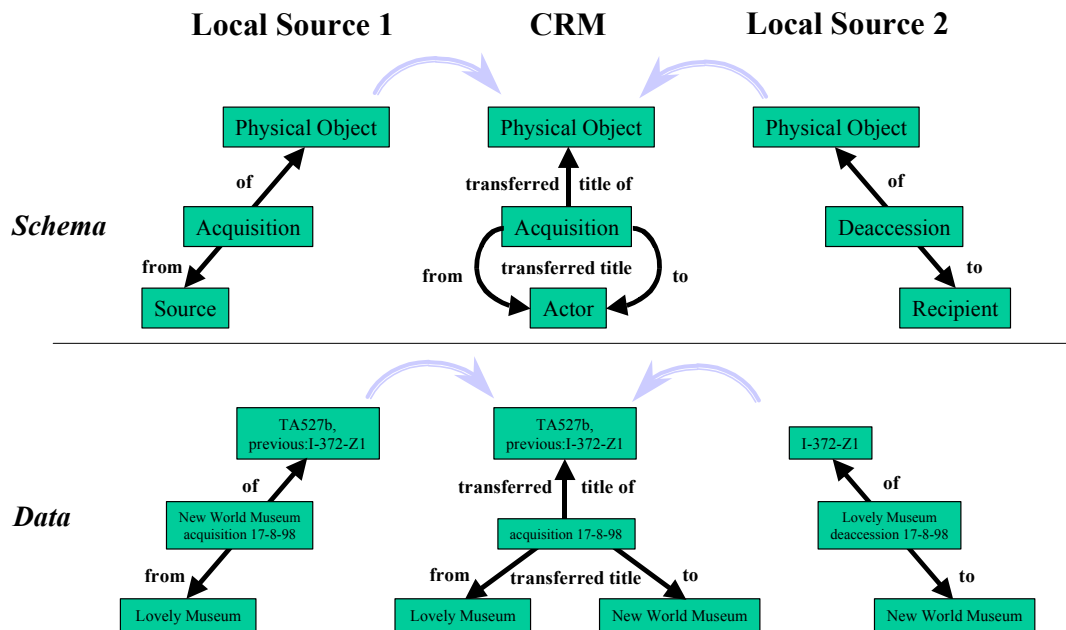


Fig 1: Creating a symmetric data representation

4.1.2 Extensible Granularity of Reference

Let us assume that one collection management system documents the condition of an object in accordance with the IC as a composite entity with a classification term, a date and a text (called "Condition State" in the CRM). Another database, used in a laboratory, may register the same information as an **act of condition assessment** with reference to persons, methods, documents created as well as the Condition State already described (see fig 2). Consequently, the table for objects will have no link to the Condition State, but to Condition Assessment, which in turn links to Condition State. This variable indirection or granularity of reference is another major source of incompatibility between semantically overlapping descriptions. These chains are potentially infinite. One system may refer to the condition of an object as an assessment of the outcome of a number of measurements carried out by a number of people over a period of time. A 'poorer' system may not even refer to the date and text, but simply register a term such as 'good' 'bad, or 'indifferent'. Such differences may be entirely justified by the intended use of the information in a given context. We have encountered numerous cases where radical differences in the granularity of information are justified by the intended purpose of the documentation - there is no one *right* way to do things and richer systems are not necessarily *better*⁵.

In such cases, we model two paths, direct and indirect, and characterise the "poorer", direct reference as a **short cut** of the entity it bypasses (a simple kind of deduction in database technology terms). The resulting CRM model thus appears to be redundant (fig 2). The idea is however, that any given implementation would use only one of the two alternatives. The Reference Model thus defines how data from the richer to the poorer system are transformed and how the richer system can be queried from a poorer model.

⁵ The use of the words 'richer' and 'poorer' is not intended to imply a value judgement concerning the applicability or appropriateness of any given information system, but is restricted to a comparison of the level of granularity which a system supports.

Although one cannot expect to recover the missing data, it is nevertheless possible to transfer data from the poorer to the richer model. The "gaps" can be filled with default values and conservative "guessing", for example, by making the assumption that a "condition assessment" event took place on or before the date associated with the condition state. This condition assessment event can be assigned a type "assumed" in order to avoid confusion with real data. Other assumptions can be derived from general knowledge about the database, like *termini postquem* and *antequem*, names of actors etc. Note that a mediation system must be able to handle, consistently, unknown and assumed values in data fields.

Interpreting a reference as short cut of newly introduced entity allows reference chains to be extended indefinitely, without loss of compatibility, to the level of detail required by any implementation. The model can also define appropriate simplifications as "compatible alternatives". Obviously, the notion of compatibility used here is dynamic: a level rather than a fixed number of concepts. This is just one aspect of extensibility and reduction. The next paragraph deals with another dimension: extension to more specific concepts.

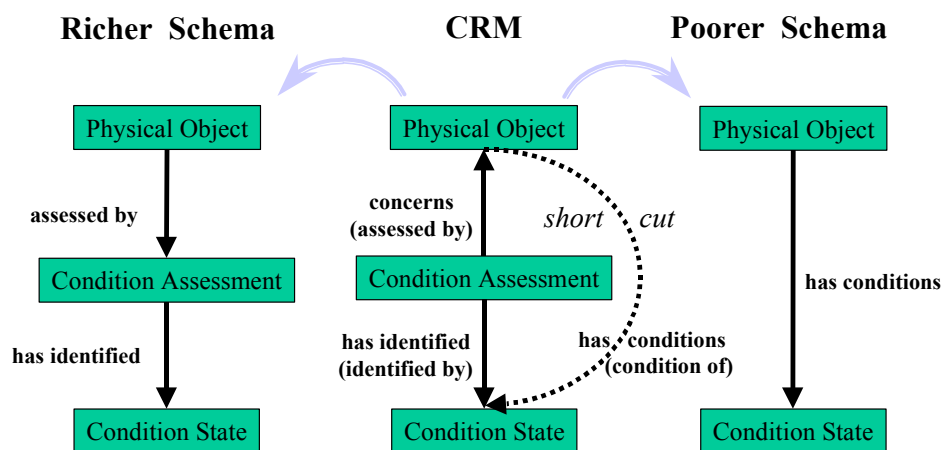


Fig 2: Short cuts of indirect references

4.1.3 Extensibility and Genericity

Let us imagine two collections management systems, one designed for coins and one for paintings. Both use specific tables. A third system follows the IC and uses a single table for any kind of physical object. Obviously, coins and paintings are physical objects, and the "standard" system is more generic than the other two. This relation is called "isA" in knowledge representation, often its inverse is called "subsumption", and its mapping into entities of an object-oriented database schema is called "generalisation/ specialisation" or "superclass / subclass", etc. (See fig. 3 for coins). For more detail, see the rich literature on this topic. Many theories provide many terms, each with a slightly different flavour. But all describe the same basic notion, the second key concept of object-orientation. Specialisation increases the number of known features of an entity and restricts the application of the entity to fewer instances. Four problems arise in a heterogeneous environment:

1. One may wish to query all three databases for, say, painting and coins, without the need to be aware of the respective differences in implementation.
2. Even though coins and paintings do not overlap, related places, persons, periods, times etc., may overlap. Hence one may wish to formulate queries on any common abstraction of coins and paintings.
3. One may wish to load data from the specific to the generic database.
4. One may wish to load appropriate data from the generic to the specific database, e.g. all coins.

From the point of view of database implementation, a subclass may be seen as table, which has all the fields of its superclass(es) ("*inheritance*"), plus some additional fields. When we query the superclass, the database will regard all instances of the subclasses as instances of the superclass. Therefore the "isA" construct allows us to "merge" the two databases with the standard one, physically and/or logically in a mediation system. This deals with the principle problems 1,2 and 3 mentioned above.

On this basis, one can extend the 'standard' database, i.e. one built following the CRM, to any more specific use, without losing compatibility. Following our example in the common object-oriented paradigm, the "Physical Object" entity can be queried and will return coins and paintings simply as physical objects, without however telling us about their specific nature. Furthermore, no specific attribute of "Coins" or "Paintings" can be queried using the "Physical Object" entity. In other words, with generalisation we lose information about the type of the

subclass and its specific features. In this view, the CRM plays the role of a coarse "*shareable ontology*", the maximal common contents of all possible extensions.

Two simple tricks help to reducing this loss of information. First, all entities in the CRM carry a "type" field, which either encodes directly the subclass a data object belongs to, or encodes a "narrower term" of the type of the subclass (e.g. "coin, NT: dime"). Given that all data are appropriately classified, and a thesaurus is used to provide the respective broader terms, we do not lose information about the kind subclass of this instance at the "standard" level. Problem 4 can be solved in the opposite way. Second, we may attach general attributes (links) to more generic entities as "containers" for the additional attributes of subclasses, analogous to entity specialisation.

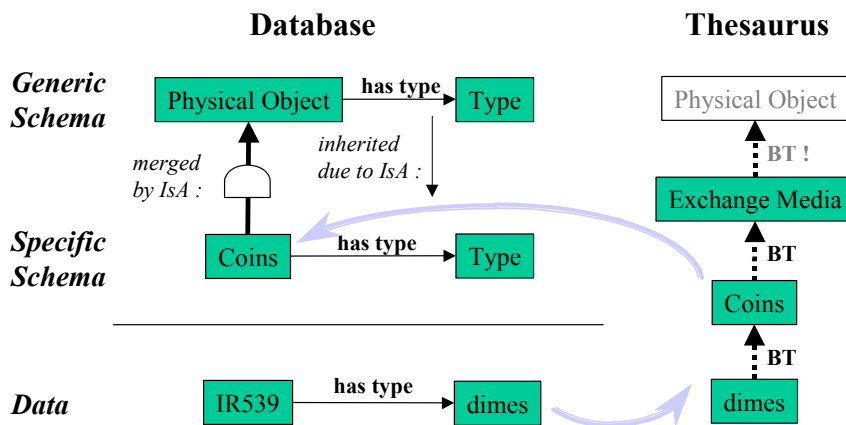


Fig 3: Merging generic and specific tables and the role of thesauri

Of course the flexibility of a standard depends not only on its ability to grow and encompass richer levels of detail, but also its capacity to interpret or to communicate with **poorer** systems which implement coarser grained information. We therefore analyse systematically the entities we need in the reference model for common generalisations or abstractions that may be useful for queries at different abstraction levels or data transfer to poorer systems. The level of specialisation of the "standard" becomes a relative state of development. It can become richer and richer, and one can define a dynamic range of compatibility levels, as outlined above for the extensible granularity. The richer the ontology, the more it can mediate. Viewed in this way it becomes possible to invert the role of the ontology, and use the CRM as a *reference ontology* [Guar98], which serves to formalise poorer systems and their relative semantics.

Simultaneously, we observe that the types in the hierarchies of entities of the CRM tend to cover most of the topical subject hierarchies known from thesauri in the domain. This implies that the terminology hierarchies contained in thesauri have to be closely coupled with the respective ontology hierarchies in the CRM in order to allow correct mediation. This has consequences for both ontology creators and thesaurus providers. As both represent deep knowledge of the field, only a co-operative harmonisation can result in a sound formulation. Since ontologies approximate to a language-independent conceptualisation of a domain, multilingual thesauri may adopt an ontology as conceptual back-bone structure. Ontology and terminology can, of course, be seen as two aspects of the same thing: the ontology gives more detail concerning attributes and links, whereas the terminology focusses on nuances between different entities.

Summarising, the "genericity principle" allows for querying or transferring data with well-defined restrictions or losses between levels of specialisation. When combined with extensible granularity, it becomes possible to encompass any foreseeable extension of the data structure which remains consistent with the underlying conceptualisation. The more detailed the "standard", the better the communication. A compatible system of (multilingual) thesauri of topical subjects provides substantial added value.

4.1.4 Multiple and Ambiguous Nature

The last principle has to do with the uniqueness of points of view. As terminology work on thesauri has shown, particular concepts can have multiple generalisations and real things can be seen under different aspects. Multiple generalisations ("*multiple isA*") can be directly described in the ontology. For example, the CRM handles a "Person" as both an "Actor" **and** as a "Biological Object", an "Inscription" as both a "Mark" **and** a "Linguistic Object" etc.

Many multiple aspects of real things are explicitly represented in the model. However, strictly speaking there is no need to do so, since entities of the model are not a priori mutually exclusive. A framed collection of butterflies can be both, a "Man-Made Object" and a "Biological Object" ("**multiple instantiation**"). It is important to bear in mind that the CRM plays an explanatory role rather than a that of a standard format. Decisions concerning formats are essentially implementation details. We have therefore separated certain aspects into different entities according to their causality, even though they may co-occur. e.g., the "Destruction" of an object is always an event, but not necessarily wilfully caused. However, we regarded it as unhelpful and problematic to draw a sharp distinction between "wilful" and "accidental" destruction. Therefore the entity "Destruction" has no actor. Intentional activities by people which result in destruction are seen as events with a double nature: both "Activity" and "Destruction".

We have not attempted to formalise which entities can co-occur on an instance and which cannot. This is not necessary for an a posteriori taxonomy, though it may be helpful for system design. Obviously, multiple instantiation helps to avoid decision conflicts on things with ambiguous nature. We wish to stress here that the purpose of the ontology is to support communication and retrieval, and that it should therefore capture *all* potentially relevant aspects, i.e. *it is better to say something wrong than to leave something out*. This is quite the opposite approach to that adopted by a scientific taxonomy, which would *rather say nothing than something wrong*. The purely scientific aspect has to be captured by the data itself, in texts and any other appropriate form.

4.2 Overview of the Model

4.2.1 Basic Entities

Many directions can be taken to develop a conceptual model and virtually any entity can be indefinitely refined and extended. Without a specific work program and considerable discipline, working groups tend to get bogged down in details and may often focus on the special fields of interest of some participants. On the other hand, a carefully selected set of examples, which represent the "core" notions, metaentities readily emerge which glue together specialisations such as types of events, objects, actors etc. As the creation of a reference ontology is in principle an endless task, it is important to establish the correct methodology, one which allows different groups to "build" co-operatively over an extended time frame on one common consistent logical construct, rather than to worry about questions of detail.

In its current state, the model is the result of a program of restrictions in several conceptual dimensions, which allowed a clear work package and criteria of completeness to be defined. The current restrictions were:

1. In the **conceptual framework** (viewpoints) of the intended users with an emphasis towards physical history and physical analysis.
2. In the intended use for common museum **activities** (collections management and conservation, research and analysis, promotion and communication)
3. In the kind of features of typical **objects** collected by museums
4. In the level of **detail** and **precision** required for adequate communication between institutions.
5. In **technical complexity** to declarative forms without the use of logical rules or algorithms.

Further work will widen some or all of these restrictions.

Presenting an overview of the CRM in a succinct and comprehensible form presents a major challenge. The scope and depth of the model, the level of detail, and the intimate relations that exist between all its elements, make it difficult to find an appropriate starting point. The class hierarchy itself, thanks to its pyramidal structure, suggests a natural 'top down' presentation. However, this point of entry also has the inevitable drawback of starting with some extremely high-level abstractions which may be difficult to grasp and which have no obvious practical application. We therefore beg the indulgence of readers impatient to get to the 'nuts and bolts' of the model.

The schema below presents the *main* branches of the class hierarchy, omitting detailed subclasses, links and attributes (fig. 4).

The highest level class in the model, **CIDOC Notion**, serves as an abstract container for all other classes in the model. It has no other significance beyond this and can therefore be ignored for most intents and purposes.

CIDOC Type is the class for the definition of a parallel type hierarchy - the thesaurus-like structure described above which provides a mechanism for enhancing the level of granularity of the model and which facilitates its implementation using relational database engines. It can, in fact, be seen as a metaclass, since its instances characterise classes.

CIDOC Entity is the parent class for all the main classes in the model.

The subclasses of CIDOC Entity are separated here into three groups for presentation purposes. However, all are *direct* descendants of CIDOC Entity.

The first group is composed of the four basic concepts which are fundamental to the model and which constitute the primary focus of cultural heritage documentation.

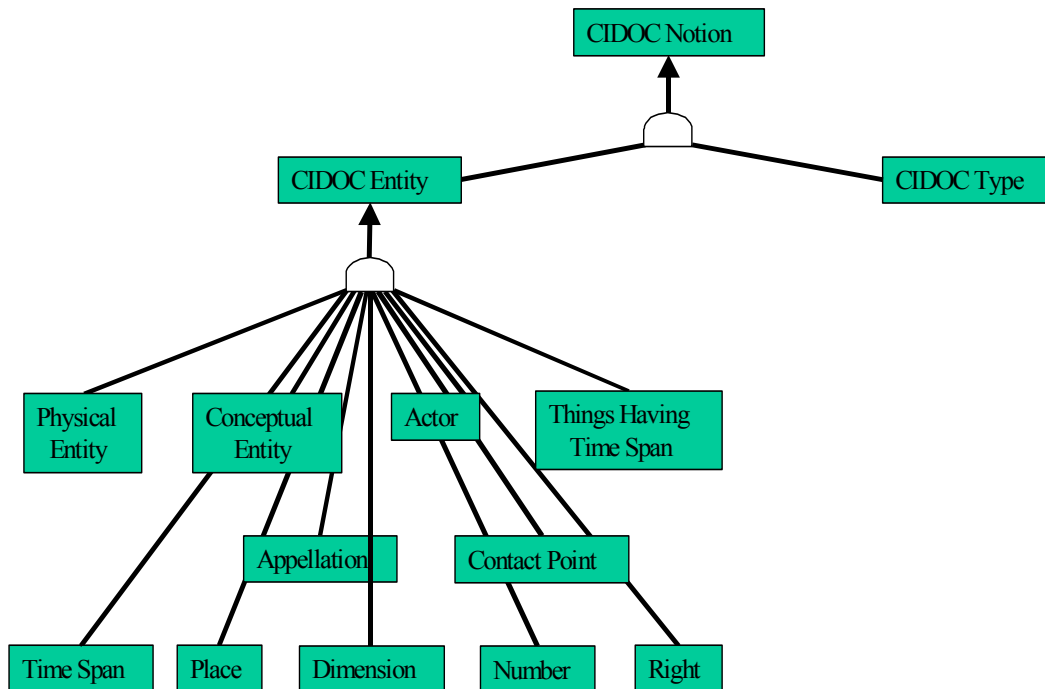


Fig 4: Overview of the CRM

Physical Entity is the parent class of all physical objects, and physical features which includes objects in museum collections, but also things like valleys, rivers, holes etc.

Conceptual Entity is used for intellectual or conceptual objects, independent of their physical manifestation or support. This distinction will be familiar to librarians as that between an edition of a book, the basic unit of bibliographic documentation, and the physical copies which are on the shelves. The CRM extends the class to include other conceptual objects such as *Designs and Procedures*, *Linguistic objects* such as inscriptions and titles, and *Visual items* such as marks, images and symbols. No attempt is made to provide a theoretical definition of the scope of this class because of the obvious philosophical and logical problems involved. It is best considered simply as the union of its subclasses - a dynamic convention.

Actor is the class of all agents - persons, groups and institutions - capable of actions, and therefore potentially responsible for events which result in changes of state.

Things having Time Span is, unfortunately, the best name we could come up with for the class which groups together periods, events, and states, all manifestations which are volatile in time.

These primary entities can be combined in specific ways to create simple propositions - like sentences in natural language - in which “**Things having Time Span**” function as a verb. The attributes of the model formalise the anticipated combinations and their meanings, (as well as the connections with the “ancillary concepts” mentioned below.)

Generally speaking, **Actors**, and **Physical** or **Conceptual Entities**, are connected through periods, events or