Electronic Communication on Diverse Data –
The Role of the oo CIDOC Reference Model

Dr Martin Doerr ,
ICS FORTH, Crete, Greece
martin@csi.forth.gr

Nicholas Crofts,
DSI, Geneva, Switzerland
nicholas.crofts@ville-ge.ch

Abstract

The oo CIDOC Reference Model represents an ontology; i.e. it describes in a formal language the overt and implicit concepts and relations relevant to cultural heritage information. The authors argue that the semantic and structural incompatibility of existing information systems constitutes a major barrier to integration. An ontology such as the CIDOC Reference Model can serve as a basis for mediation of cultural heritage information and thereby provide the semantic '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 authors position the role of a reference model with respect to the major projects in the field before presenting the principles features of the CIDOC reference model along with examples of its application. They conclude with some ideas for future development.

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, terminology is often incompatible, and the level of detail varies. 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 sub set of 'core' data.

The CIDOC reference model aims to overcome these limitation by providing a semantic reference point which 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 data model, 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 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 positions the role of a reference model with respect to the major projects in the field before presenting the principles features of the CIDOC reference model along with examples of its application. We conclude with some ideas for future development.

2. Communication needs

Access to museum documentation, presented in an appropriate manner, has the potential to interest a wide audience: researchers, educational institutions, professionals in the field of cultural heritage and the general public. In each case it is important that information presented should be *integrated* both between institutions and across disciplines.

Firstly, the value of information is 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 value. Additional references to other objects, and to an object's historical, geographical, and cultural origins help to place each item in a context and give it meaning. At present, contextual information tends to be distributed across many institutions. Without some form of interaction between the different information systems much of the potential interest of the collections is lost.

Secondly, professionals and researchers working in the field of cultural heritage often require access to detailed and accurate *global* information and statistical analysis. Preservation and conservation of cultural heritage, for example, can benefit from statistical information as an aid to decision making for the distribution of resources. Global analysis only becomes possible when wide scale and integrated sources of information are available.

To illustrate the value of cross collection integration 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 Breugel and his sons since between them, 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. Several versions have been reunited by an enterprising student of art history[1]. 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 subject. The precise date of execution and attribution of each work becomes highly significant since we instinctively want to arrange the images in chronological order.

It is significant 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 events[2].

Different forms of presentation can be imagined to meet different requirements. Statistical analysis and in depth research require systematic and precise query facilities which can generate exhaustive lists of items. This kind of approach may be 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

[1] http://www.cwd.co.uk/babel/bruegel.htm

[2] 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.

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 is not just a question of homogeneous data formatting. Information from different sources embodies different viewpoints: both as different disciplines and as different types of collections. Natural history, fine arts, ethnography, etc. but also archives, libraries, and other types of collections.

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 'creator' is totally irrelevant. Archaeologists 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 incertitude 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 failure of librarians to store information about the *attribution* of books is not simply an oversight - it would be counterproductive to do so. Each discipline and domain embodies a series of implicit assumptions and presuppositions about the semantic value of the data it handles which 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 is precisely what the CIDOC reference model aims to provide.

3. About Mediation

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

3.1. RAMA, CHIO and AQUARELLE

Between 1992 and 1995, the RAMA project successfully demonstrated that large heterogeneous databases of museum objects in different countries could be accessed using a uniform user interface. The project provided a solution to technical question of **interoperability**, i.e. how to issue client requests to different platforms over the net and to present responses. However, it did **not** attempt to resolve the semantic differences between the various information sources, and consequently allowed for accessing databases only one at a time. Nevertheless, it demonstrated the value of a well-designed, **uniform** user interface and the interest of presenting even heterogeneous data. It is further noteworthy that this was undertaken on database records and not on texts.

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, that structuring using SGML makes texts far more accessible to more precise questions, and that a standard retrieval protocol allows for accessing a vast range of data sources, could have a considerable impact on the technology used by the museum community. CHIO resolves the problem of divergent data formats by **standardisation** of common mark-up tags and common Z39.50 access points. The freedom of interpretation allowed by Z39.50 access points to the target systems allows for the resolution of some questions of semantic heterogeneity. (This point is discussed in more detail in chapter 4.) 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 gave a focus to 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:

- **Dynamic handling of DTDs**[3]. The variety of applications and the precision needed in a professional environment could not be covered by a single DTD. So-called "folder servers" constitute repositories of heterogeneous document collections. Standardisation was limited to a document header of metadata. Provisions were made to enable querying of other document structures in the near future.
- **Search aids.** The user interface provides support for query formulation: selection of a set of target sources to which the request is broadcast and the possibility of translating terminology as appropriate for each target. Experimentation revealed the extraordinary importance of being able to select terms in accordance with the terminology used in the fields of a specific target system. The project provides a complete environment for multilingual thesaurus development and access through central terminology servers [Doer98]. However, complete coverage of the terminology in the participating data sources exceeds the frame of the project.
- **User control.** Work in a professional environment requires handling of sensitive data. AQUARELLE controls access permissions and makes provisions for further accounting services in a central application server, called "Access server". Other central services are the directory management of attached sources and the **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 problems of structural heterogeneity to be resolved in a relatively straightforward way. The project has highlighted the problem of incompatible terminology used in analogous data fields. Consequently, the project has had to invest considerably efforts 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. "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 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.

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.

[3] A DTD is a document type definition. This is the standard SGML mechanism for defining the semantic structure of a document and the corresponding tags.

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. 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**") to the ontology by simple declarations, (*not* to notions specific to 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.

Obviously, the richness of the ontology ultimately determines the mediation capabilities. Some semantic differences cannot easily be described by declarative statements, e.g. the relation between conditions of preservation and events of deterioration. These cases may need specific custom functions. 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 an intellectual guide for "hard-wired" services and for determining 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. Not the superficial identity of structure, but the semantic compatibility is needed. This will enable far richer services to be created than standardisation 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

Since 1996, members of the CIDOC Documentation Standards Group, including the authors, have elaborated a proposal for a "CIDOC Reference Model" (in the following "CRM"). It represents an ***ontology*** in the sense of computer science [Guar98], i.e. an approximation of a conceptualisation of a domain by a formal language and a vocabulary. 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 differences to arise in the presentation of identical semantic contents due to the different purposes and points of view of the 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 (in the following "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 on 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 in the model, we regard both instances as identical, because the object we refer to is identical. This notion of object identity, which is independent from temporary changes in description, is a key concept of object orientation [Atki89], [Kim90] in contrast to the Relational model. 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.

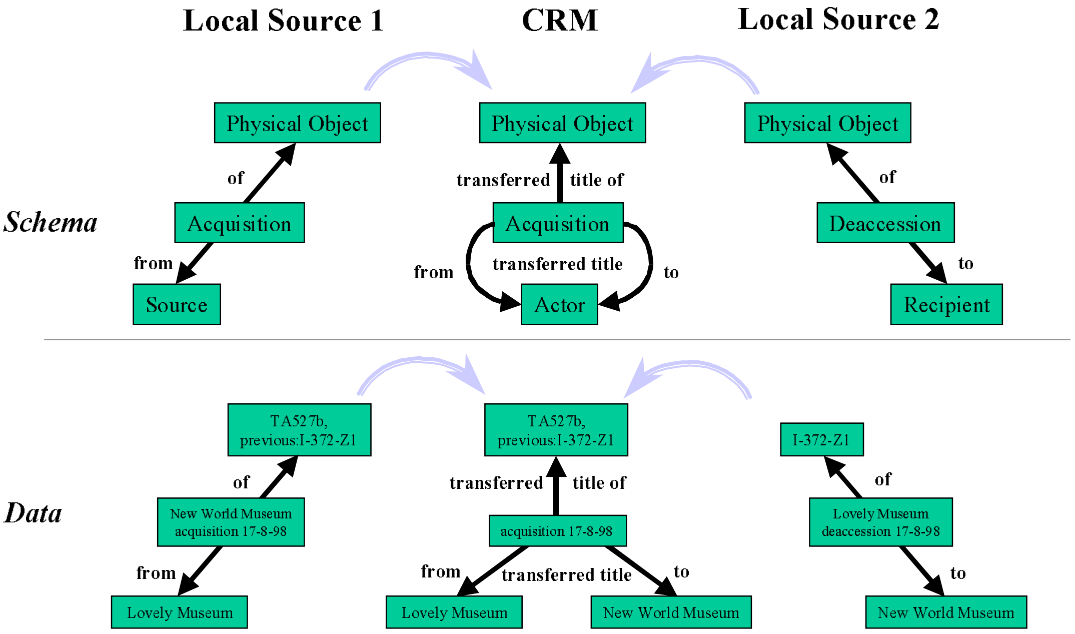

Fig 1: Creating a symmetric data representation

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 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, from a grammatical subject, where the link originates, to a

grammatical object, where it points. All historical information uses past tense, whereas states use 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 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.

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 heterogeneity between semantically overlapping descriptions. These chains are potentially infinite. While 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 differences in the granularity of information are justified for the one or the other purpose of documentation.

In these cases, we model both cases, the richer and the poorer, and characterise the "poorer reference" as a ***short cut*** of the entity it bypasses. The resulting CRM model may thus appear 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 or how the richer system can be queried from a poorer model.

Furthermore, although one cannot expect to recover the missing data, it is nevertheless possible to transfer data from the poorer to the richer model by filling in the "gaps" with default values, e.g. by assuming that a "condition assessment" event took place at the same date as the state of condition. This condition assessment can be assigned a type "assumed" in order to avoid confusion with real data. Note that a mediation system must be able to handle consistently unknown and assumed values in data fields.

This mechanism allows for extending the model and respective implementations to any level detail and to any number of indirection without loss of compatibility. 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.

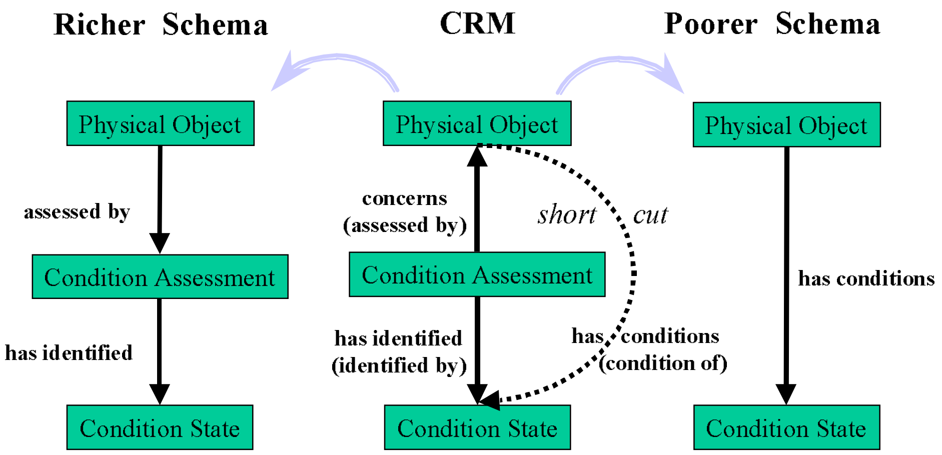

Fig 2: Short cuts of indirect references

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 details 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 reduces the applicability 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, 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 mentioned above 1,2,3.

On this basis, one can extend the 'standard' database, i.e. one built following the CRM, to any more specific use, without loosing 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 man-made 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 provides the respective broader terms, we do not lose information about the subclass of this instance at the "standard" level. Problem 4 can now also be solved, for items for which a specific table was designed. Second, the more generic entities may provide general attributes (links) as "containers" for the additional attributes of subclasses, analogous to entity specialisation.

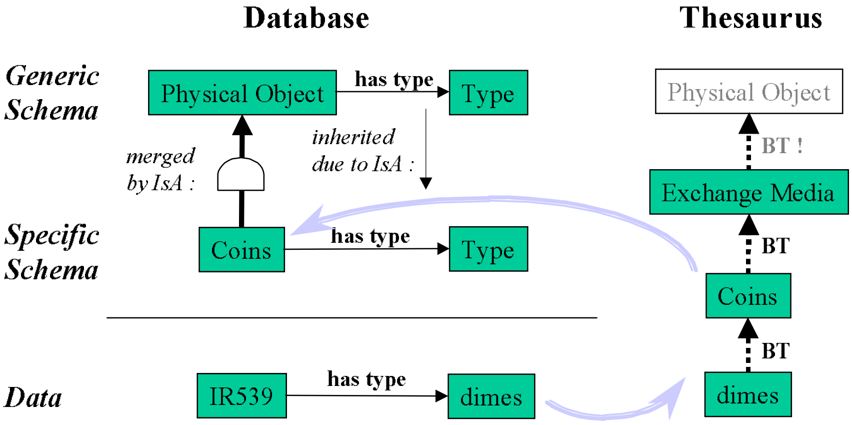

Fig 3: Merging generic and specific tables and the role of thesauri

Of course the flexibility of a standard depends not only on its ability to grow to encompass richer levels of detail, but also its capacity to interpret or to communicate with **poorer** systems which implement

coarser grained information[4]. 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. We can now invert the role of the ontology, and use the CRM as a ***reference ontology*** [Guar98], which serves to formalise the 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 serve correct mediation. This has consequences for both sides, the ontology creators and the 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, the compatibility of thesauri with an ontology may assist in the creation of multilingual thesauri. Ontology and terminology can, of course, be seen as two aspects of the same thing: The ontology applies to details of attributes and links, and the terminology to hierarchies of entities.

Summarising, the "genericity principle" allows for querying or transferring data with well-defined restrictions or losses between levels of specialisation. Together with the extensible granularity, one can cover any foreseeable extension of modern data structures. The more detailed the "standard", the better the communication. A compatible system of (multilingual) thesauri of topical subjects provides a substantial added value.

4.1.4. Multiple and Ambiguous Nature

The last principle has to do with the uniqueness of aspects. As well known from thesauri, 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.
Multiple aspects of real things are explicitly represented in the model. In fact 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***"). The CRM plays an explanatory role rather than a format. Decisions on formats are implementation details. Therefore we have separated certain aspects into different entities according to their causality, even though they may co-occur. E.g., a "Destruction" of an object is always an event, but not necessarily caused wilfully by people. We regarded it even as problematic to draw the borderline between "wilful" and "unwilling" destruction. Therefore the entity "Destruction" has no actor, and we see any activity of people resulting in destruction as an event with double nature: an "Activity" and a "Destruction".
We have not however engaged in a formalisation of which entities can co-occur on an instance and which cannot. This is not necessary for a posteriori taxonomy, but 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 therefore should capture all potentially relevant aspects, i.e. *it is better say something wrong than to leave something out*. This is quite the opposite approach to that of a scientific taxonomy, which would *better say nothing than something wrong*. The pure scientific aspect has to be captured by the data itself, in texts and any other appropriate form.

4.2. Development Strategy

4.2.1. About Form and Standards

Specific mediation systems may select relative "light-heartedly" a powerful knowledge representation model of their choice, but for a community such as ICOM, open standards, ease of use and availability of tools is mandatory. Obviously, the use of the object-oriented paradigm is necessary. Beyond that, the

[4] The use of the word '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.

model must be intelligible for non-experts and domain specialists. This is particularly true for graphical and semantic models.
We have so far formulated the model in two forms, as MERISE graphics for illustration and a rigid definition in a "light-weighted" object-oriented semantic model, which is readable by non-experts and converts easily to other paradigms. It actually follows the TELOS knowledge representation language ([Mylo89]), an open research result implemented in several systems, that has been successfully employed by one of the authors to built a cultural documentation system [Dion94]. It is our intention to reformat the model in the EXPRESS-G language, ISO10303-11, part of ISO10303 "Industrial automation systems and integration – Product data representation and exchange", called "STEP" for the following reasons:

- It is the only ISO standard of an object-oriented modelling language.
- It is the only standard for a graphical object-oriented language, which can be formally processed and incrementally interchanged.
- It is appropriate. Product definition has some deep similarities with museum objects.
- It has logical extensions to formulate missing semantics of pure oo languages.
- The majority of database providers interface to it.
- There are multiple suppliers of tools

4.2.2. About Proceeding

Many directions can be taken to develop a conceptual model. Virtually any entity can be refined and extended. Without a specific program and hard discipline, working groups tend to get bogged down in details and often focus on the special fields of interest of some participants. On the other hand, with even a small set of examples, the "core" notions, the meta entities, 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 more 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.

The Documentation Standards Group therefore decided to apply a program of restrictions in several conceptual dimensions, which allowed for defining clear work packages and a completeness of the model with respect to each work package. These restrictions cover:

1. The **conceptual framework** (viewpoints) of the intended users (scholars, museum professionals and museum visitors, etc.)
2. Common museum **activities** (collections management and conservation, research and analysis, promotion and communication)
3. The **objects** collected by museums
4. The level of **detail** and **precision** required for providing an adequate level of quality of service.
5. Considerations of **technical complexity**.

Whereas the first two points are obvious, let us comment the other three. If we do not restrict the instances, the differences between most well-defined notions begin to blur, as e.g. places, landscape features and physical objects; mobile and immobile items; crystals, viruses and living beings, etc. Consequently, the well-intentioned "global model" looses its precision, power and utility. This problem is reminiscent of those referred to by Paul Feyerabend with respect to the global logical foundations of science.
Formal definitions of many data fields with specific semantics are attractive, because they read like guidelines for contents of documentation. This is however the purpose of guidelines and not of an ontology for information mediation. We have therefore foreseen free-text fields for all information, on which we did not expect a specific global query or which could be queried through other obvious associated information, as e.g. "Mark description" or "Physical description" of the IC. Some formal definitions that require composite rules were also regarded as being too complex to be worth investigating at present; e.g. the relation between state and state transition information, such as ownership and acquisitions.
By virtue of these principles, we have succeeded in 'finishing' our model of the IC. Naturally, the scope and depth of the model can be widened incrementally, as appropriate to meet future needs, purposes and cost-benefit considerations.

4.3. Immediate Use

What is the immediate use of such a model? Obviously, one can implement information systems, which conform to common notions from the outset and hence are easy to integrate, or which simply represent good practice of the field. Furthermore, one can expect that the existence of an well-accepted reference model will foster activities to create active mediation services for the domain. However, we see a most prominent immediate use as being the definition and processing of Z39.50 access points, Metadata definitions, SGML DTD and guideline creation. The use of the CRM will allow these different standards to be integrated, even though they were made for different but overlapping purposes, into interoperable forms.

4.3.1. AQUARELLE, CIMI, Z39.50 and SGML

Because of its simplicity, Z39.50 is a very attractive access protocol for wide data access based on minimal assumptions. Z39.50 effectively sees any database as a single table of objects with a flat attribute set. How these attributes are created is the mystery of the wrapper, the Z39.50 gateway. Seen from the CRM, these attributes are equivalent to "short cuts" from the "Physical Object" entity to respective entities, combined with a few direct attributes. The AQUARELLE project has invested considerable effort in defining an attribute set which satisfies professional access on items as divers as museum objects, architecture and secondary information. The outcome [AQUA98] is a successful but "minimalist" approach, with very few high-level concepts such as "where", "when", "who", "what", and a few others, which cover maximal heterogeneity, as pointed out in [LeVa98]. It was merged for reasons of interoperability with the complementary CIMI profile, which is "maximal", and requires more homogeneous sources. With this merge, both profiles cover a large world-wide consensus for the cultural area.

Two observations where important: First, the attributes in the two profiles have a great deal of semantic overlap with each other which has not been formalised. Hence users and implementers have no precise idea, which "what"s include which other attributes. Second, the individual implemented Z39.50 gateways exhibited large differences in interpretation of the attributes with respect to the fields of the local database, which rendered search results inconsistent and difficult to compare. For this purpose, AQUARELLE introduced a central quality control for such mappings. Obviously, an object-oriented CRM would allow both aspects to be formalised and hence to decentralise quality control. Moreover, it would facilitate derivation of consistent attribute sets for different purposes without the need for interminable discussions between expert needed simply to rediscover basic concepts. Hence a CRM has immediate relevance for the quality of Z39.50 interoperability in the cultural domain.

In accordance with the needs of the professional cultural community for a variety of well-structured documents, AQUARELLE has the capacity to handle DTDs dynamically. The philosophy behind this was that the DTD could be used to express the semantic structure of a document, rather like a database schema, and not just formatting information such as with HTML. Consequently, specific tags are directly or indirectly related to attributes of the Z39.50 profile [Chris97], [Chris97b]. A mapping mechanism between tagged elements and Z39.50 attributes was introduced in the document repository (folder server) similar to the mapping of attributes to database fields. The original CIMI DTD followed another philosophy, where Z39.50 attributes are marked as link anchors in the document. This approach does not enforce consistency of structure and access points and requires additional work. CHIN’s new information system (http://www.chin.gc.ca) also relies on the mapping of tagged elements to solve heterogeneity problems.

The use of DTDs for semantic structuring and the respective support of structured queries with semantic mapping of tags is just another potential field of use for the reference model. If attribute sets and SGML tags are formally related to the ontology, the mapping is a direct consequence, which can even be automated, as shown in chapter 3. On the other side, various DTDs can be derived from an ontology according to the objects and aspects to be documented, thereby maintaining interoperability from the outset. The same holds for documentation guidelines. The IC, for example, can be easily formulated as a DTD. Since the CRM represents the semantic contents of the IC, this DTD would be a projection of the CRM from the point of view of a museum object. Another projection could be made for artists. Consequently, an access system incorporating the CRM can automatically combine information from both sources.

4.3.2. Metadata

Metadata has become a buzzword recently. In the proper sense, it means any data or information about data. A good analysis of the term can be found in [Kash97]. The museum community seems to have adopted the far narrower sense used by the library community: "a description of objects, documents or services which may contain data about their form and content" [Haka96]. The most prominent representative is the Dublin Core. As part of the document itself or not, it is just an extension of the document structure for the purpose of querying [LeVa98], and RDBMS-based collection management systems will typically not make a distinction between schema and such metadata. Recently, the notion seems to have widened, particularly with the Resource Description Framework (RDF) [RDF98], and at the Helsinki Metadata Workshop DC-5 [Weib98] the need of a common formal data model for the exact definition of RDF and Dublin Core was recognised.

This formal model is going to be formulated as a conceptual graph in manner comparable with the CRM we propose. It effectively becomes part of a domain ontology for the library world. Some harsh criticism by archaeologists [Mill97] on the applicability of certain attributes of the Dublin Core, shows that library metadata are not transferable without thought to museum objects. One of the main reasons for this is that metadata for documents describe documents, which may **in turn** describe objects, whereas similar records on museum objects are **themselves** documents. The differences can only be resolved with reference to a clear ontology. To this end the CRM contains a more detailed analysis of the "subject" information for museum objects. In AQUARELLE, too, this difference also became apparent. The project defined a so-called "Folder Profile", a document header of metadata derived from the Dublin Core and Warwick Framework.

Precisely as in RDF, the museum community could give a formal account of its metadata on the basis of a CRM. A library and a museum ontology, and hence the resulting metadata, could be consistently merged. It should be underlined that a major purpose of ontologies is the ease of merging they provide, which in turn facilitates the merge of all derived products (see e.g. [Kash97]), however, the ontology does not **replace** metadata, Z39.50 attributes, DTDs etc.

5. Future work

An important aspect of future work involves integration with other standards. As mentioned earlier, the Documentation Standards Group intentionally imposed a restricted scope on the CRM in order to render the workload more manageable. Consequently, information dealt with by the CRM is essentially limited to that already present in the IC. The guidelines produced by other CIDOC groups, such as archaeological sites and ethnography, are obvious candidates for incorporation in the CRM. Natural history is another area, which is not fully developed in the CRM. This is an inevitable reflection of the lower level of representation of the natural sciences within CIDOC as a whole. However, we consider it to be quite as important as other disciplines.

Equally important is integration with existing *terminological* resources. Authority files and thesauri effectively embody ontologies or fragments thereof. Optimal use of the CRM requires that terminology hierarchies be mapped onto the class hierarchy in a consistent manner.

Finally, we hope that the CRM will be taken up by other projects in its role as a neutral semantic reference and used in the conception of information systems and mediation services. Inevitably this will generate comments and requests for extension and modification of the model. Maintenance of the model will most likely be an ongoing task in the foreseeable future.

6. Conclusions

Along with many others we share a vision of museum information systems integrated into a working resource covering all disciplines and aspects of human culture. At present, museum information systems are isolated, incompatible and underexploited. The CRM provides a blue print for integration and the construction of links between individual systems and, we believe, constitutes an important step towards the realisation of a global network for cultural heritage.

7. References

[AQUA98] Systems Simulation Ltd., «Aquarelle Z39.50 Profile», Revision 1.22, 15th May 1998
http://www.cimi.org/documents/aqua_profile_0598.html

[Atki89] M. Atkinson, F. Bancilhon, D. DeWitt, K. Dittrich, D. Maier, and S. Zdonik, «The Object-Oriented Database System Manifesto» In: Proceedings of the First International Conference on Deductive and Object-Oriented Databases, pages 223-40, Kyoto, Japan, December 1989. Also appears in «Building an Object-Oriented Database System: The Story of O2.» F. Bancilhon, C. Delobel, and P. Kanellakis. (eds.)Morgan Kaufmann, 1992.
http://www.cs.cmu.edu/People/clamen/OODBMS/Manifesto/index.html

[Kim90] Won Kim, «Object-Oriented Databases: Definition and Research Directions», invited paper in: IEEE Transactions on Knowledge and Data Engineering, Vol.2, No.3, Sept. 1990

[Baya96] R.J.Bayardo et al.: «InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments», MCC Technical Report, MCC-INSL-088-96, October 1996.

[Chaw94] S.Chawathe, H.Garcia-Molina, J.Hammer, K.Ireland, Y.Papakonstantinou, J.Ullman, J.Widom, «The TSIMMIS Project: Integration of Heterogeneous Information Sources», In: Proceedings of the IPSI Conference, pp 7-18, Tokyo, Japan, October 1994

[Chri97] Christophidis V., Doerr M., Fundulaki I. (1997). "The specialist seeks expert views: Managing digital folders in the AQUARELLE project." In Museums and the Web, 1997, Selected Papers from an International Conference, Los Angeles, California, March 15-19, 1997. Pittsburgh, Pennsylvania: Archives & Museum Informatics.

[Chri97b] V. Christophides, M. Doerr, I. Fundulaki. «A Semantic Network Approach to Semi-Structured Documents Repositories». In: Carol Peters and Constantino Thanos, editors, Research and Advanced Technologies for Digital Libraries, Lecture Notes in Computer Science, pages 305-324, Pisa, Italy, September 1997. First European Conference on Digital Libraries ECDL'97, Springer-Verlag.

[Dion94] I.Dionissiadou, M.Doerr, "Mapping of material culture to a semantic network", in : Automating Museums in the Americas and Beyond, Sourcebook, ICOM-MCN Joint Annual Meeting, August 28-September 3, 1994

[Doer98] Martin Doerr, Irini Fundulaki «The Aquarelle Terminology Service», ERCIM News Number 33, April1998, p14-15

[Guar98] Guarino N. Formal Ontology and Information Systems. In N. Guarino (ed.), Formal Ontology in Information Systems. Proc. of the 1st International Conference, Trento, Italy, 6-8 June 1998. IOS Press

[Guar98b] Guarino N. Some Ontological Principles for Designing Upper Level Lexical Resources. Proc. of the First International Conference on Lexical Resources and Evaluation, Granada, Spain, 28-30 May 1998

[Haka96] Juha Hakala, Ole Husby, Traugott Koch, «Warwick framework and Dublin core set provide a comprehensive infrastructure for network resource description», Report from the Metadata Workshop II, Warwick, UK, April 1-3, 1996 , http://www.ub2.lu.se/tk/dcwsrept.html

[Kash97] Vipul Kashyap, Amit Sheth, «Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies», Academic Press 1997, Cooperative Information Systems: Current Trends and Directions. M. Papazoglou and G. Schlageter (editors), (with A. Sheth)

[LeVa98] Ralph R. LeVan, «Dublin Core and Z39.50», OCLC Research Project Report 1997 http://www.oclc.org/oclc/research/publications/review97/levan/dublincoreandz3950-v1_1.html

[Mill97] Paul Miller, Alicia Wise, «Resource Discovery Workshops: Final report from the Archaeology Data Service», prepared in accordance with guidelines from the Arts & Humanities Data Service (AHDS) and United Kingdom Office for Library & Information Networking (UKOLN) Resource Discovery Workshop series, 4 August 1997, http://ads.ahds.ac.uk/project/metadata/workshop1_final_report.html

[Subr94] V.S.Subramanian, Sibel Adali, Anne Brink, James J.Lu, Adil Rajput, T.J.Rogers, R.Ross, C.Ward, «HERMES: A Heterogeneous Reasoning and Mediator System».

[RDF98] «Resource Description Framework (RDF)» July 21 1998, http://www.w3.org/RDF/

[Weib98] Stuart Weibel, Juha Hakala, «DC-5: The Helsinki Metadata Workshop, A Report on the Workshop and Subsequent Developments», ISSN 1082-9873, D-Lib Magazine, Feb.1998

[Wied92] Gio Wiederhold, «Mediators in the Architecture of Future Information Systems», in : IEEE Computer, March 1992.