Towards a Core Ontology for Information Integration

Martin Doerr[1], Jane Hunter[2], Carl Lagoze[3]

[1]Institute of Computer Science, Foundation for Research and Technology, Heraklion, Greece, martin@ics.forth.gr
[2]DSTC Pty, Ltd., Brisbane, Australia, jane@dstc.edu.au

[3]Computing and Information Science, Cornell University, Ithaca NY, lagoze@cs.cornell.edu

Abstract

In this paper, we argue that a core ontology is one of the key building blocks necessary to enable the scalable assimilation of information from diverse sources. A complete and extensible ontology that expresses the basic concepts that are common across a variety of domains and can provide the basis for specialization into domain-specific concepts and vocabularies, is essential for well-defined mappings between domain-specific knowledge representations (i.e., metadata vocabularies) and the subsequent building of a variety of services such as cross-domain searching, browsing, data mining and knowledge extraction. This paper describes the results of a series of three workshops held in 2001 and 2002 which brought together representatives from the cultural heritage and digital library communities with the goal of harmonizing their knowledge perspectives and producing a core ontology. The knowledge perspectives of these two communities were represented by the CIDOC/CRM [31], an ontology for information exchange in the cultural heritage and museum community, and the ABC ontology [33], a model for the exchange and integration of digital library information. This paper describes the mediation process between these two different knowledge biases and the results of this mediation – the harmonization of the ABC and CIDOC/CRM ontologies, which we believe may provide a useful basis for information integration in the wider scope of the involved communities.

1. Introduction: Information Integration

In a recent presentation on digital library developments, Lorcan Dempsey described interoperability as "recombinant potential" [21]. This phrase epitomizes the potential for interoperability mechanisms and architectures to integrate heterogeneous information across the *internet commons* [45]. As articulated by Paul Ginsparg [27], *information integration* provides the basis for a rich "knowledge space" built on top of the basic web "data layer". This knowledge layer is composed of value-added services that process and offer abstracted *information* and *knowledge*, rather than returning *documents* (in the manner of most current web search engines).

Information integration on the web involves a number of architectural building blocks that are the focus of work of the W3C and the related semantic web community [12]. This work includes mechanisms for information encoding and manipulation (e.g. XML [3], RDF[35], XSLT [17]), and ontology construction and reasoning (e.g., RDFS [15], DAML+OIL, OWL [20]). Information integration also motivates much of the metadata work in the digital library community. Some of this work is focused within specific domains (e.g., FGDC [40] in the geospatial community, IMS LTSC [7] in the educational/instructional community), while other metadata initiatives are looking beyond domain specificity towards providing services across heterogeneous content (e.g., Dublin Core [2] and its goal of cross-domain resource discovery).

This paper describes work on a *core ontology*, arguably another of the building blocks to information integration. The goal of a core ontology is to provide a global and extensible model into which data originating from distinct sources can be mapped and integrated. This canonical form can then provide a single knowledge base for cross-domain tools and services (e.g., resource discovery, browsing, and data mining). A single model avoids the inevitable combinatorial explosion and application complexities that results from pair-wise mappings between individual metadata formats and/or ontologies.

The distinction between a core ontology and *core metadata*, such as Dublin Core, is subtle but important. Both are intended for information integration. They differ, however, in the relative importance of human understandability. Metadata is in general created, edited, and viewed by humans. Therefore, human factors, including limits on complexity, should play a primary role in its design. In contrast, a core ontology is a underlying formal model for tools that integrate source data and perform a variety of extended functions. As such, higher levels of complexity are tolerable and the design should be motivated more by completeness and logical correctness than human comprehension.

The core ontology described in this paper is the result of a series of three workshops held throughout 2001 and 2002 under the sponsorship of the DELOS Network of Excellence on Digital Libraries [1] and the NSF/DSTC/JISC-funded Harmony Project [5]. The nature of the workshops and the participants reflect a specific design philosophy. One approach to the creation of a core ontology is to attempt to "represent all knowledge". A notable example of such an effort is Cyc [37]. The utility and applicability of such a complete approach may be proven over the long-term. However, in the nearer term, we argue that a more useful approach is to bring together representatives of major communities and *harmonize* their knowledge perspectives. We define harmonization as a process of modifying two ontologies, preserving their intended functionality, but integrating them into a coherent wider model. In general, this process may imply changes in the initial concepts. The communities engaged in this harmonization effort were the developers of the CIDOC/CRM [31], a basis for information exchange in the cultural heritage and museum community, and the ABC ontology [33], an effort to create a core ontology for digital library information.

Our goal in this paper is therefore two-fold. Firstly, we will describe the mediation process between the subtly, yet overtly different knowledge biases of these two efforts and how that process may provide a model for mediation amongst other ontological biases. Secondly, we will describe the results of this mediation – the *harmonization* of

the ABC and CIDOC/CRM ontologies. We believe that the harmonized ontology presented here may provide a useful basis for information integration in the wider scope of the involved communities.

The organization of the paper is as follows. Section 2 describes the motivation for a core ontology in an information integration scenario. Section 3 summarizes the two ontologies, the CIDOC/CRM and ABC, which were the inputs to the harmonization process. Section 4 then describes the harmonization process and describes how ontology experts and formal tools assisted in the process. The actual results of the harmonization are presented in Section 5. The paper closes with Section 6 that describes opportunities for future work.

2. Motivation: An architecture for information integration

The work described in this paper to develop a core ontology is motivated by an information integration scenario as described by Diego Calvanese et al [16]. Implementing such a scenario requires that the ontology represent complete data structures, rather than just terms with the data. A brief description of this scenario is as follows.

Central to this scenario is the notion of an *enterprise model*, which, as described in [16], "is a conceptual representation of the global concepts and relationships that are of interest to the application". The enterprise model provides a basis for the integration of an open set of *source data* (i.e., independent communities on the web or any other open information space). The mapping is possible due to:

- The establishment of *source models* that provide conceptual representations for each category of source data. These models are not necessarily complete (representing all knowledge in the source data) but are sufficient for the application demands on the enterprise model.
- The definition of a *domain model* that expresses the interdependencies between and amongst the enterprise and domain models.
- The formalization of these conceptual models and relationships via decidable frameworks such as *description logic* [16]. Both the ABC and CIDOC/CRM ontologies described here are expressed in a form that could be transformed into the description logic framework and are, thus, compatible with this scenario.

One consumer of such an integrated model might be a cross-domain search service. If the information integration is correct then a query to the *enterprise model* should produce the same answer as that yielded by applying the respective query over each of the sources.

We regard a core ontology as the upper level of such an enterprise model. A well-formulated core ontology should be *complete* and *extensible*: expressing the basic concepts that are common across a variety of domains and providing the basis for specialization into the nuanced concepts and vocabularies of individual domains.

Furthermore, a core ontology should be more than the upper level of a terminology. As argued in [22], it should be "property-centric" - providing a common language which

defines how information elements are related rather than just providing typological distinctions. Such a language not only enables the aggregation of equivalent information (e.g., all works of an author in a certain period), but also enables the integration of complementary information (e.g., about meetings between Van Gogh and Gauguin together with images and descriptions of their work in this period). Both of the ontologies that were input to this harmonization process focus on relationships and the classes necessary to define those.

3. Context: Concept spaces from two communities

The CIDOC/CRM and ABC ontologies reflect a commitment to the expression of the common concepts underlying the data structures used by their communities. In both ontologies, these "common concepts" were determined in an experimental way by intellectual analysis of relevant sources [18]. Furthermore, there has been and remains vivid interest in both communities in exchanging and integrating information.

However, Borges states that "all classifications of reality are by nature conjectural and fictional" [13]. In fact, the cultural bias of any classification system has been described in detail by a number of scholars [14, 34]. Both the CIDOC/CRM and ABC groups entered the meetings with models of reality that overlap but, in certain ways, reflect specific perspectives and scope.

Perhaps the most striking example of 'cultural bias' in the source ontologies is the difference in the temporal aspects of both models. While both models attempt to model change over time, the nature of *change* is intriguingly different. The ABC model, driven by digital library requirements, was originally motivated by the need to describe how objects change over time. Examples of this type of change are the versioning of digital objects or the production of derivative works. The CIDOC/CRM model, motivated by cultural artifacts and museum requirements, focuses more on changes in context and ascription rather than object transformation itself. A typical example of this is the movement of an object between museums or the reclassification of an artifact due to changes in scientific understanding. Additional examples of the differences between the two source ontologies and how these differences were resolved are described in sections 4 and 5. A key lesson of these meetings was not only that cultural bias exists even in closely related domains, but that tracing it down to its roots opens the way to find the common ground and common language.

The remainder of this section summarizes the nature of the two models that served as input to the harmonization process. While both efforts are thoroughly described in their respective publications (cited below), the short summaries that are provided here should be sufficient for understanding the remainder of the paper.

3.1. CIDOC/CRM

The "CIDOC object-oriented Conceptual Reference Model" (CRM), originates from earlier standards proposals produced by the ICOM/CIDOC Documentation Standards Group [23]. Since September 2000, the CRM has been progressing as an ISO standard (ISO/AWI 21127) in a joint effort of the CIDOC CRM SIG and ISO/TC46/SC4. It represents an 'ontology' for cultural heritage information i.e., it describes, in a formal

language, the explicit and implicit concepts and relations underlying the documentation structures used for cultural heritage. The primary role of the CRM is to serve as the semantic 'glue' needed to transform disparate, localized information sources into a coherent and valuable global resource. In more concrete terms, it is a conceptual model that can be used:

1. as a global schema in data warehouse-like applications, to provide access to the knowledge of multiple institutions, internally or open to the Internet;
2. as a global schema for query mediation to heterogeneous sources;
3. as a set of concepts which provide a good practice guide for conceptual modeling and system development in the domain;
4. to create common tagging schemes for digital preservation of domain knowledge.

The intended scope of the CIDOC CRM [18] may be defined as all information required for the scientific documentation of cultural heritage collections, with a view to enabling wide area information exchange and integration of heterogeneous sources. The term *scientific documentation* is intended to convey the requirement that the depth and quality of the descriptive information that can be handled by the CRM should be sufficient for serious academic research into a given field and not merely that required for casual browsing.

This does not mean that information intended for presentation to members of the general public is excluded, but rather that the CIDOC CRM is intended to provide the level of detail and precision expected and required by museum professionals and researchers in the field. The term cultural heritage collections is intended to cover all types of material collected and displayed by museums and related institutions, as defined by ICOM [6]. This includes collections, sites and monuments relating to natural history, ethnography, archaeology, historic monuments, as well as collections of fine and applied arts. The scope of the CRM is the curatorial knowledge of museums. The exchange of relevant information with libraries and archives, and the harmonization of the CRM with their models, falls within the CRM's intended scope. Information required solely for the administration and management of cultural heritage institutions, such as information relating to personnel, accounting, and visitor statistics, falls outside the intended scope.

The CIDOC CRM is specifically intended to cover contextual information: the historical, geographical and theoretical background in which individual items are placed and which gives them much of their significance and value. It aims at compiling with high precision the relevant primary knowledge scattered over disparate sources, with all their omissions, contradictions and alternative opinions. It does *not* intend to encode particular social or psychological interpretations or “stories”, but to deliver to the scholar or scientist the best possible collection of material to do such studies. The goal of enabling such information exchange and integration between heterogeneous sources determines the constructs and level of detail of the CIDOC CRM. It also determines its perspective, which is necessarily supra-institutional and abstracted from any specific local context.

The CIDOC CRM aims to leverage contemporary technology whilst enabling communication with legacy systems. It has been formulated in the knowledge representation formalism of TELOS [42] – not because of a preference for this specific formalism, but out of the necessity to use at least one. Essential to the CIDOC CRM are subsumption hierarchies of both, classes and attributes (properties), and the concept of

multiple inheritance (a class or property can have multiple super-classes or super-properties). The CIDOC/CRM uses attributes (properties) on attributes in order to denote a dynamic subtyping of roles, a feature that can be simulated in languages that cannot support this. The scope notes of some attributes (properties) in the CIDOC CRM declare that these attributes can be deduced from other data paths of the model. These deductions could be formalized in several *Knowledge Reasoning* languages by use of logical expressions, or by procedural software code in respective applications. Reasonable encodings of the CIDOC CRM itself can be created in: RDF/RDFS, DAML-OIL, UML, TNM, STEP/Express and others. CIDOC CRM instances can be encoded in virtually any format currently in use.

CIDOC CRM is more than a core standard of the most basic entities and relationships. It attempts to adequately capture the semantics behind the most common data structures of the cultural heritage domain and related domains, to the degree that they fall within the intended scope. Through its *isA* hierarchies and deductive data paths, it enables the integration of data structures based on concepts at different levels of abstraction and detail, such as Dublin Core and AMICO.

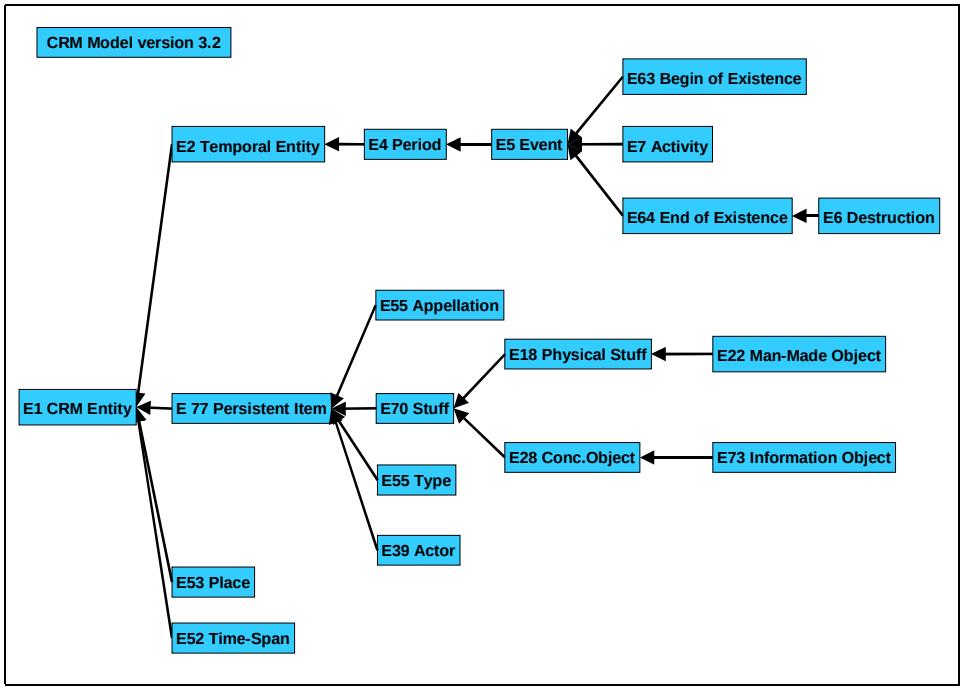

Figure 1: Part of the CIDOC CRM IsA hierarchy of classes

The final edition of the CIDOC CRM, which is due for completion in Fall 2002, comprises 83 classes and 130 properties [19]. Figure 1 shows the upper part of the class hierarchy relevant for the presented work. It is “event aware” in the sense that models events explicitly wherever possible: Time is only connected to temporal phenomena (Temporal Entity), also called *perdurants*, [38]. Location (Place), people (Actors), material and immaterial items (Stuff) are primarily connected to temporal entities. For example, that a thing resides at a place is seen as a result of a temporal entity e.g., the building of a house. In addition to that, the CIDOC CRM builds into the model the

historic discourse about names (Appellation) and identification, and about classes (Type) and classification.

3.2. ABC Ontology

The ABC ontology is an outcome of the Harmony Project [5], which is funded cooperatively by the Distributed Systems Technology Cooperative Research Centre (DSTC) (Australia), the Joint Information Systems Committee (JISC) (UK), and the National Science Foundation Digital Libraries Initiative (NSF DLI2) (US). The original goal and continuing motivation of the ABC work arises from the need to integrate information from multiple genres of multimedia information within digital libraries. The researchers working on the Harmony Project have each been involved in a number of metadata initiatives including Dublin Core and MPEG-7 and from this experience recognized a number of key issues:

- The number of metadata vocabularies will continue to grow as individual communities seek to structure their own information for their own purposes;
- Attempts to develop universal metadata vocabularies are misdirected, since "spoken" languages (those used by communities to actively describe content) will inevitably diverge (history is replete with failures to find common spoken languages [25]);
- A more useful effort is to attempt to formulate a language as a base for "understanding" - a core ontology - incorporating basic entities and relationships common across the diverse metadata vocabularies;
- Such a core ontology might then be useful for a number of purposes including integrating information from heterogeneous vocabularies, uniform processing across heterogeneous information sources, and providing base concepts that future metadata initiatives could build on when developing domain specific vocabularies.

Complete details of the ABC ontology are described in a 2001 Journal of Digital Information paper [33]. Figure 2 shows the class hierarchy of ABC. In the remainder of this section we give only a short summary of ABC features.

As noted above, the ABC ontology is intended to model the multiple entities that are essential to the description of digital (and, indeed, physical) objects. Many metadata vocabularies are largely resource-centric, inadequately expressing entities such as people, places, ideas, and the like. For example, both Dublin Core and MARC provide a set of properties that are associated with a primary resource, the "library object". The values of some of these properties – for example, "creator" in Dublin Core and "100" in MARC – are entities themselves, the nature of which are not fully modeled in these formats. As a result, the properties of these other entities (e.g., agents) are either inadequately modeled, as in Dublin Core, or segregated into other models (e.g., authority records) with associations that are implementation-specific, causing problems associated with poor interoperability across systems.

ABC also describes temporality in a first-class manner. As we noted elsewhere [32], modeling change over time is critical to the description of digital content due to its inherent fluidity and the linkages of provenance to integrity or trust. ABC includes both the notions of "events" and "situations", which respectively model transitions (i.e., verbs) and existential properties. The inspiration for these concepts lies in process models such as Petri Nets [44] and temporal extensions for first-order logic such as Situational Calculus [39]. Due to these temporal concepts, ABC is able to definitively model time periods during which certain properties of an object are static. It is also able to model events or transitions marking property modification, for example during the change of a version of a digital object.

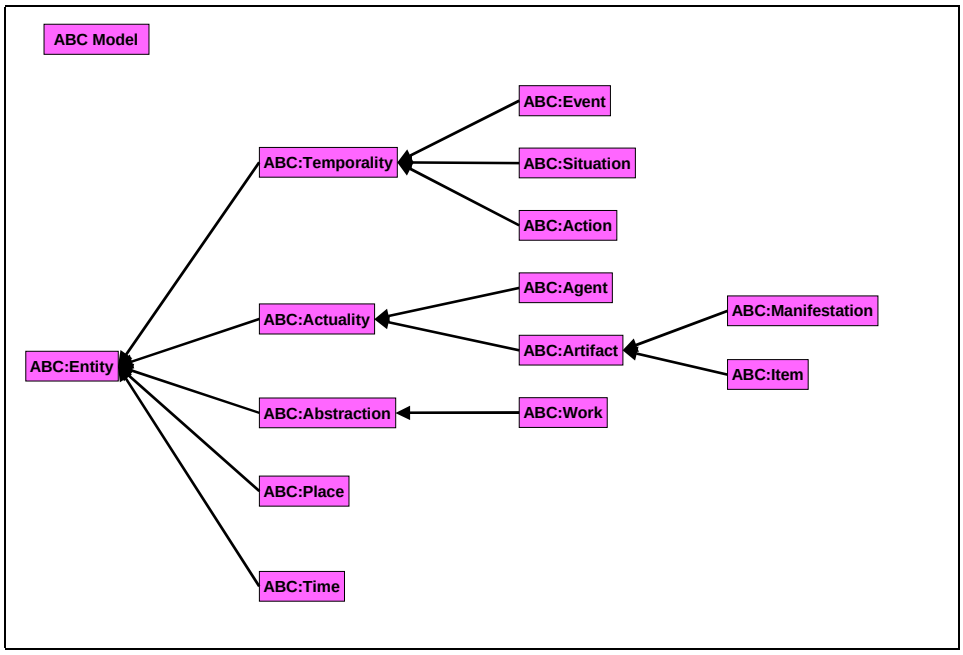

Figure 2: The complete ABC IsA hierarchy of classes

Finally, ABC builds on the concepts developed in the FRBR model [4]. These concepts – *works, expressions, manifestations,* and *items* – give ABC the ability to link entities that have common intellectual property origins. Work in the library community has proven the utility of these concepts [36].

While not dependent on RDF and RDFS, the ABC model is strongly influenced by the core aspects of the RDF data model; that is, the definition of classes and sub-classes for entities and properties and sub-properties for relationships among those entities. The ABC paper [33] includes an expression of the ABC ontology in RDF Schema language. In addition, the Harmony web site [5] includes an ABC model builder that provides a graphical user interface for constructing instances of the ABC model and writing them out as RDF instance documents. It also provides an example of a XML query interface across ABC instances mapped from metadata provided by four different CIMI member organizations.

Given an RDF schema representation of the ABC model, it is then possible to evaluate its usefulness as a tool for information integration by applying it to the integration of ontologies from other domains. For example, in [30] Hunter has evaluated it as basis for

combining the MPEG-7 [9]ontology (for multimedia content description) and the MPEG-21/<indecs> [8] ontology (for rights management) to enable both the discovery and rights management of multimedia content.

4. Process: Building an agreement across the concept spaces

The DELOS harmonization meetings began with detailed explications of the models by representative parties in order to set the context as summarized in section 3. This process laid the foundation for the harmonization task, which consumed the bulk of the remaining meeting time. The group decided to follow a formal, analytical, deterministic method to compare and converge ontologies based on the OntoClean approach [29]. This method analyses top-level ontological distinctions, in particular metaproperties, related to:

1. instantiation versus membership
2. part-of and mereological axioms
3. extensionality
4. connection
5. location and extension
6. co-extension, co-connection
7. unity, singularity and plurality
8. dependence/independence

Furthermore, the OntoClean approach enables the detection of concept definitions that are lacking in clarity or rigidity, the justification of valid subsumption relations, and the detection of invalid subsumption declarations.

In addition to comparing the concept definitions of each model, the meetings invested substantial time and effort into investigating how well the declared concepts fit the specified functionality and scope, with the goal being to better clarify the intended functionality and to improve both models. From the beginning, it was assumed that agreement on a coherent common ontology would be possible following this process. This assumption was based the following beliefs, which were later confirmed:

- that the groups shared common conceptualizations between both domains,
- that diverging definitions of seemingly equivalent concepts could be traced back to different intended functionality, and
- that the groups could learn useful concepts from each other (perhaps originating from the uniqueness of their respective domains), which could consistently be integrated.

The application of the OntoClean approach had a catalytic effect. Differences between the devised concepts that had previously been poorly understood became clear. Several concept definitions were refined; for example, the ABC concept “State” was transformed and clarified to “Situation”. However it did not bring us any nearer to understanding the reasons for the differences, almost immediately raising questions over how each team justified their divergent definitions.

One example of non-congruent concepts that needed to be resolved was the difference between ABC *Agent* and CRM *Actor*. From the ABC perspective the notion of an *Agent*

is tied to the phase of a person or machine acting during an event. CRM, on the other hand, defined the notion of an *Actor* that was bound to a person or group over its life-time in the legal sense. Both are perfectly logical definitions, so why choose the one or the other? In what scenarios is one definition better? Is there a common concept that covers the differences?

In the course of these discussions we concluded that in some cases the level of abstraction of some concepts is so high that there is no objective third-party source by which the correctness of a concept can be measured. Rather it must be admitted, that concepts like “event”, “actor”, “agent”, can be defined in different ways, as appropriate for each context. In order to understand each context and the intended functionality behind the decisions made in the respective context, the meetings focused on specific example scenarios, using both models to describe the assumed or real facts in the scenarios.

The nature of this process seems to suggest the more philosophical question: to which degree are original user conceptualizations well-defined at all. Ontologies are often assumed to be the result of user conceptualizations derived from some pre-existing ‘nature of reality’. Instead, our impression is that they possess to a certain degree a character of invention, derived from originally weakly defined concepts. They are a hybrid intellectual-technological solution to the need to impose conceptualizations that meet a set of desired functionalities. The situation resembles Wittgenstein’s extendable boundaries of categories, as presented by George Lakoff in [34]. If we accept that user conceptualizations have a certain degree of freedom, this can be exploited to craft models that fit better together. That seemed to be the essence of the harmonization process, we describe in more detail in the remainder of this section..

As a means of moving past such philosophical conundrums and towards successful harmonization, the group agreed that a core ontology must be adequate for the following desirable functionalities in an information integration scenario, such as described in section 2:

1. The ability to transform typical data examples from the domain of discourse into semantically equivalent instances of each model. Such an ability requires that the knowledge expressed in the data examples together with general background knowledge is sufficient to instantiate concepts of the model. This process raised problems related to the character and the level of detail used in each model. Certain facts can always be inferred; e.g., the existence of a creation process can be inferred from the existence of an artifact, whereas other facts may be based on weaker assumptions. Certain questions related to the cause of an event or the equivalence between two identities, may not be resolvable.
2. The ability to integrate redundant, overlapping and complementary information into one consistent data set without the need to delete nodes created in between. For example, an assumed state of stability may need to be split into two, if a further event is encountered from another data set.
3. The ability to handle and locate alternative opinions about the past in a predictable way in the integrated data set.
4. The ability to transform integrated data sets into narratives about objects and to trace individual human contributions to creative, transitional and usage processes.

The effect of this agreement on desired functionality was three-fold. Firstly, further differences between the intended functionalities of each model were identified. Secondly, some of the existing concept definitions needed to be revised to include cases not previously thought of by either team. Thirdly, both teams introduced new constructs or changed existing ones in order to cover cases and aspects they regarded as pertaining to their common set of intended functionality e.g., CIDOC CRM introduced the concept of "Persistent Item", an anchor for properties common to objects, actors, names etc.

This process prepared the ground for the merged ontology presented in Section 5 below. Certain details of the merged ontology still need to be clarified, but the fundamental achievement is a common understanding of the reasons for the modeling choices, an acknowledgement of both solutions adopted, and an understanding of their intellectual relationship. Within this process, another iteration of the OntoClean methodology to the merged ontology may be helpful to ensure semantic consistency.

5. Results: A merged ontology for information integration

This section describes the ontology resulting from the merge. We begin with an overview of the agreement on class definitions, and then follow with the agreement on and discrepancies between properties. We then comment on the key problems encountered, and on the key ideas that led to a successful agreement.

There was agreement between the two models on the following classes:

ABC:Entity	is equivalent to	CRM:Entity
ABC:Temporality	is equivalent to	CRM:Temporal Entity
ABC:Event	is equivalent to	CRM:Event
ABC:Action	is equivalent to	CRM:Activity
ABC:Artifact	is equivalent to	CRM:Man-Made Object
ABC:Place	is equivalent to	CRM:Place
ABC:Time	is equivalent to	CRM:Time-Span

A comparison of the entity definitions in the two models revealed the following equivalences, subsumption relationships and differences:

ABC:Agent - Someone who performs an Action	is similar to	CRM:Actor – Someone capable of performing an Activity
ABC:Agent	is expansion of	CRM:carried out by – in the role of
ABC:Abstraction	is subclasss of	CRM:ConceptualObject
ABC:Actuality	is equivalent to (the union of)	CRM:Physical Stuff ∪ CRM:Information Object ∪ CRM:Actor

ABC:Situation	is superclass of is subclass of	CRM:Event CRM:Temporal Entity
ABC:Manifestation (= FRBR:Manifestation)	is subclass of is subclass of	CRM:InformationObject (not physical) CRM:Type
FRBR:Item	is subclass of	CRM: Man-Made Object
FRBR:Expression	is equivalent to	CRM:InformationObject

A comparison of property definitions in the two models revealed the following equivalences, subsumption relationships and differences:

ABC :hasParticipant	is equivalent to	CRM :hadParticipant
ABC :hasParticipant	is subproperty of	CRM :occurred in the presence of
ABC :hasPatient	is subproperty of	CRM :occurred in the presence of
ABC :inContext ABC :involves ABC :hasPresence	is sub-property of is sub-property of is equivalent to	CRM :occurred in the presence of
ABC :destroys	is equivalent to is super-property of	CRM :took out of existence CRM :destroyed
ABC :creates	is equivalent to	CRM :brought into existence
ABC: hasPatient	is super-property of	CRM :has modified
ABC :usesTool	is equivalent to	CRM :used specific object
ABC :phaseOf	has no equivalent property	

Figures 3 and 4 below illustrate diagrammatically, the equivalences, overlaps and differences between the two models.

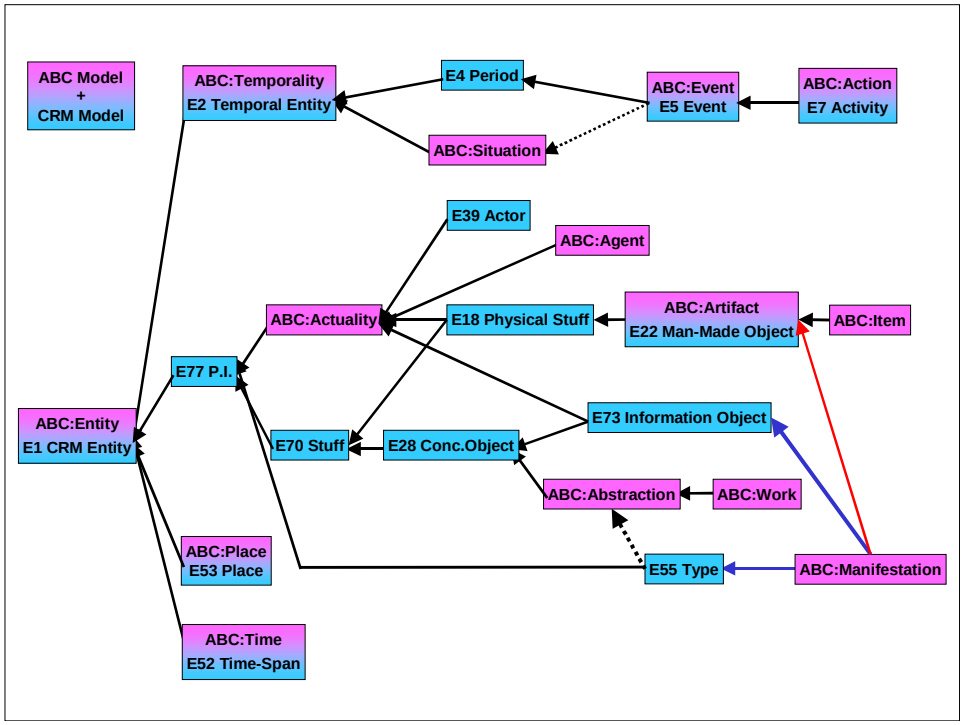

Figure 3: The merged ABC and CIDOC CRM class hierarchies

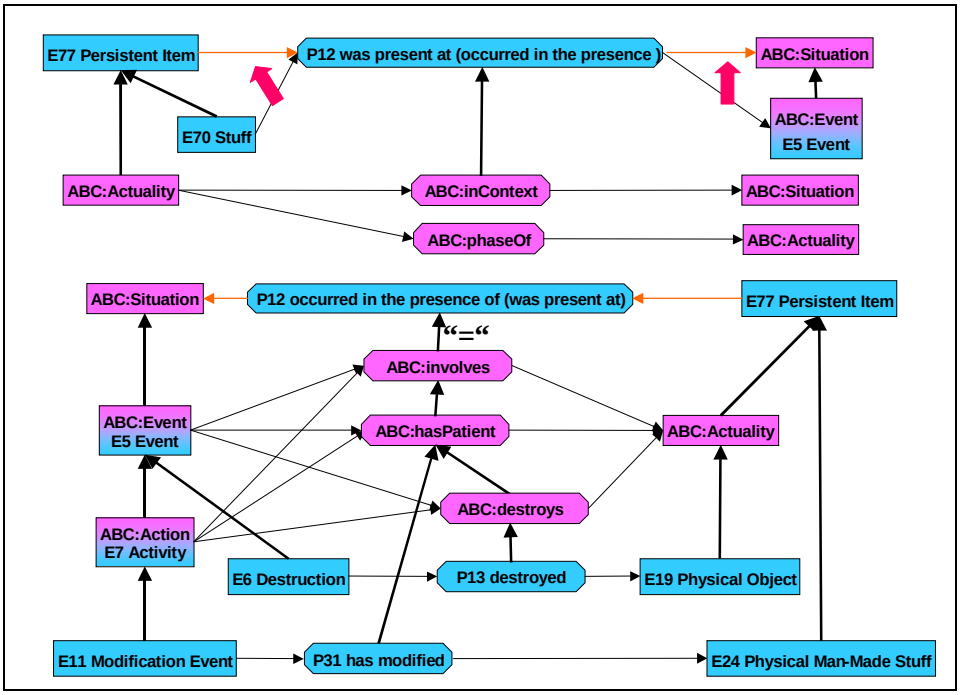

Figure 4: Relationships between ABC and CIDOC CRM property hierarchies

The major areas of conflict and discussion which were raised during the harmonization process were:

- the *Situation* class which was present in the ABC model but missing from the CIDOC CRM;
- whether the notion of an atomic Action (as provided by ABC) is actually necessary and useful?
- the fundamental distinction between *material* and *immaterial* objects in the CRM, versus the *concrete* and *abstract* dichotomy in ABC;
- problems and inconsistencies within the IFLA FRBR model.

In the following four sub-sections, we discuss each of these issues in detail.

5.1 About Situations and Events

The ABC model includes both *Events* and *Situations*, which respectively model transitions (i.e., verbs) and existential properties. The inspiration for these concepts lies in process models such as Petri Nets [44] and temporal extensions for first-order logic such as Situational Calculus [39].

The CIDOC CRM on the other hand does not include a general construct to model time periods during which certain properties of an object are static. Only the more specific notion of a *Condition State* is provided for describing phases during which an object's condition is assumed static e.g., a building was in ruins. The CRM encourages the description of transitions rather than the states or situations in between them. This approach was inspired by considerations of modern physics, which considers *Events* as complex interactions with non-negligible temporal extent, leading into phases of stability with fuzzy (if any) temporal boundaries. The observations necessary to acquire knowledge can themselves be regarded as Events that are part of the observed system. It seems that an analogous argument can be applied to historical research.

Both the ABC and the CIDOC/CRM approaches can be regarded as formally correct. A complete description of all transitions enables the inference of all states. So ABC can be regarded as a compatible extension to the CIDOC CRM, in which all states are made explicit. ABC recommends the explicit creation and description of all states/situations between the events. Extensive discussions on the relative advantages of the two approaches, generated the following key arguments:

- Events are normally primary knowledge but true witness of states is rare. States are usually inferred and inferred states are not robust against increasing knowledge and can render integration non-monotonous.
- Replacing a truly witnessed state by transitions may discard the knowledge, that there were no other transitions in between.
- Transitions may be inferred from two different observed states, such as: "intact in 1932", "in ruins in 1951".
- States are subjective and relative to a context because if any property change is below the threshold of our interest, the property is regarded as static.
- Knowledge about states, even inferred and subjective states, is a relevant part of our discourse.

One conclusion that was a key to the whole harmonization process was the realization that information integration can be divided into two phases: a *compilation phase* and a *consolidation phase*:

- During the compilation phase, no inferred knowledge should be generated. Observations of static properties should explicitly include the duration of the observation. Alternative or even contradictory information is not resolved during the compilation phase.
- However during the consolidation phase, opinions about the described reality may be produced. The consolidation phase includes the generation of narratives, the inference of states and the resolution of alternatives or contradictions.

Our analysis revealed that the CIDOC CRM is aimed at the compilation phase, and at preparing the ground for data consolidation. The ABC model, on the other hand, aims to cover both phases. Characteristic differences between the models could be traced back to these intended functionalities.

Situations are without doubt Temporalities or Temporal Entities. In a wider sense a Situation may be regarded as a constellation where certain things are present over a time-span. This is also true for the CRM perception of an Event. In a narrower sense one may like to exclude something that happens to this constellation (an Event). The wider definition can be seen as a superclass of Event, and the narrower definition as the complement of the Event as seen from the wider perspective. For reasons of monotonicity, the CRM would prefer the wider definition. In both cases, the harmonization is based on an extension of the CIDOC CRM, and a modification of the initial positions of both models.

5.2 About Activities and Actions

Legal ontologies like indecs [8], E-Court and others [8, 10] seem to decompose the human activities within an event into units called *Actions*, which have sufficient granularity to describe and reason about an individual contribution to some change of the state of affairs. Although this view is of ultimate cultural relevance, it may be context dependent or subjective and is difficult to establish during the *compilation phase*. A good analogy to this problem is the separation between the investigative work of the police and determination of the truth by the justice court. This analogy highlights the necessity of both the compilation and consolidation phases and the methodological differences between them.

So, prior to describing *Actions*, one can often describe collective activities without the knowledge of finer details or deciding about individual roles. On the other hand, there are investigative cases where even very fine-grained “subatomic” parts of events can be relevant e.g., time differences when deciding who is the winner at the Olympic Games. A reasonable generalization of *Action* seems to be the notion of events with human participation, without limitations to their aggregation or decomposition, as provided by the CIDOC CRM *Activity* concept, which is a subclass of *Event*. Because there are no relevant additional properties associated with the legal notion of Action, ABC adopted the CRM notion as the more generic one. Indeed, in ABC, some properties had both,

Event and Action, as domain. Inheritance of the properties of the Event class by the Action class simplified the respective ABC definitions. In this case, harmonization is based on a generalization of the ABC model – and *Action* was made a subclass of *Event*.

5.3 About Concreteness of the Immaterial

Initially the CIDOC CRM effort focused on descriptions of material objects. Therefore one of the most productive outcomes for the CIDOC CRM from this collaboration was a better understanding of the nature of immaterial human creations. This improved understanding resulted in several amendments to the CIDOC CRM – most notably changes to the "Stuff", "Information Object" and "Persistent Item" entities.

Material objects have a striking characteristic: they have a virtually infinite number of properties (scratches, microscopic patterns, DNA, etc) existing at least during some limited time-span. This characteristic makes them "concrete" and generally provides a method of identification from one observation to the next. (There are certain complex exceptions, such as the replacement of all of the parts in my computer e.g., Nicola Guarino's discussion of diachronic identity [46]). Immaterial objects such as ideas, categories, plans etc. lack those "concrete" features. E.g. archaeological types are notoriously difficult to be defined in objective terms. Even biological species cannot be clearly identified without a prototypical instance.

Material objects can also be destroyed in one event. Immaterial objects cannot be destroyed, only forgotten – more precisely, the last carrier may be lost, be it a book or a human memory. This was the motivation for CIDOC CRM's strict division of the properties relating to creation, destruction, use and presence with respect to material (Physical Stuff) and immaterial things (Conceptual Objects), "…the non-material products of our minds…".

On the other hand, human creations such as texts and digital objects, although immaterial, can be clearly identified on the basis of their (finite) content. In contrast to ideas, categories, plans, etc., their content endows them with a legal and practical status similar to material items. Consequently, ABC distinguishes between *Actuality,* "...a category…of entities that have a tangible existence in some worldview" and *Abstraction.* ABC's *Actuality* also subsumes the ABC model's *Agent* class.

Hence there seem to exist three dominant distinctions between most abstract entities:

1. To be identifiable from event to event - to be *endurants* [43] or not;
2. To be *material* or not;
3. To be *tangible* (*concrete*) or not;

Each of these distinctions has an impact on the relationships to other entities. So relationships governing identification, use, creation, destruction, location, and presence take on different forms depending on the above distinctions.

A harmonization can be achieved by describing classes of one model as a combination of unions, intersections and complements of classes from the other model. We found that the union of the ABC's *Abstraction* and *Actuality* classes corresponds to the CRM's *Persistent Item* (or endurants), an entity introduced during the course of this

collaboration. This analysis also identified that the intersection of *Actuality* with the immaterial *Conceptual Object* corresponds to *Information Object*, and possibly includes CRM's *Appellation* and *Contact Point*. ABC's *Actuality* corresponds to the union of CIDOC CRM's *Physical Stuff*, *Information Object* and *Actor*.

In this case, harmonization was based on accepting the distinctions made by both models and a solution was found by determining the unions and intersections of the overlapping concepts.

5.4 About Manifestations

The FRBR model [4] is a careful analysis of the conceptualizations of the library communities that has attracted international attention. Because FRBR uses an Entity-Relationship formal description method, any integration into the more semantically rigid Knowledge Reasoning (KR) formalism used by the CIDOC CRM and ABC, requires a certain amount of interpretation. Patrick Le Boeuf, a member of ISO and the CIDOC CRM development team, has carried out an extensive survey of reactions to the FRBR and documented comments and suggestions for additions and modifications [36]. The FRBR also provoked extensive discussions during the DELOS Working Group on Ontology Harmonization meetings and in the joint meetings of the CIDOC CRM Special Interest Group and ISO/TC46/SC4.

In particular, the FRBR definition of the *Manifestation* concept was found to be particularly problematic from a logical point of view, with a number of somewhat inconsistent characteristics:

- Manifestations can be seen as the focus of current bibliographic documentation e.g., a particular edition of a book or a manuscript.
- Manifestations can be identified by an ISBN number and represent the stage at which ideas become "tangible".
- A Manifestation can be either a set of like items or a single item. FRBR characterizes both, the sets and single items as "material". FRBR further declares that an "Item" is the copy I hold in my hand. Sets however, are not material, but conceived, contrary to how FRBR defines it
- On the other hand, *aggregates* [11, 43] or *collections* [26] of material items, could be regarded as material, with a definite mass etc. So, if Manifestations were aggregates, the FRBR definition makes sense. However, editions of a book are neither aggregates nor collections - because the number of items is not fixed, when the edition is being produced. Moreover, even a manuscript or painting may be "reproduced" by the author himself.

So it seems to be more consistent to regard Manifestations as of a *set of like items*, and the *unique item* as the exception, " being the only one of its kind" (see Microsoft Encarta World English Dictionary ").

Hence we interpret *Manifestation* as "kind" –such as car models, "species" in biology or E55 Type in the CIDOC CRM – even though there may exist only one instance. A Manifestation represents the stage at which relevant physical features of the product are

determined, but it is not the product itself. In this interpretation, the concept *Manifestation* is itself regarded as a metaclass, a class of classes ("species"), and the relationship of an *Item* to a *Manifestation* is that of an instantiation.

Apart from a few exceptions (such as the KR language TELOS [42] and [24, 41]), the importance of metaclasses is ignored by the practice of conceptual modeling and most KR formalisms. A simple workaround is to make the *metaclass* a class, to declare the instances of the metaclass as data, and to simulate the instantiation of the metaclass instances via an explicit relationship at data level, such as the "P2 has type (is type of)" property in the CIDOC CRM. On the other hand, a Manifestation defines intellectual content, similar to E77 Information Objects. Hence two approaches are possible: either a Manifestation is a specialization of both, Types and Information Objects, or it is only a specialization of Types, and the information content is part of a Manifestation instance. This issue requires further discussion.

6. Conclusions

The effort described here for harmonizing relatively small ontologies may appear daunting and generate questions about scalability. Although the effort has been substantial, we believe that it has been worthwhile. Core ontologies are small and they represent the key for merging a considerably larger amount of specialized concepts [28]. In addition, there is also no real alternative, and the merging process itself raises new scientific insight into the nature of cultural bias, the functionalities determining specific modeling choices and how to integrate those.

Merging core ontologies requires considerable intellectual effort. It is a learning process for both sides, which requires a strong commitment to the common goal and a willingness to accept different perspectives and understand one's own bias. This is not only an intellectual problem, but requires careful coordination in order to satisfy the interests of the communities behind those models and their need for stable and valid solutions. Although only seven days was spent in actual meeting time, the learning and coordination processes were very time consuming with over a year spent in rethinking, verifying and coordinating agreement on solutions.

Throughout the harmonization process, a significant component of the intellectual effort was spent on clearly identifying the ontological commitment of all concepts. The textual descriptions and formal definitions of the concepts in both ontologies were extremely useful for this. However, a more useful route to agreement on common concepts was a thorough analysis of the intended functionality via prototypical scenarios. This analysis led to an understanding of two key issues. Firstly, there were the different biases resulting from the two groups' opposing foci on volatile, immaterial (digital) objects (ABC) versus rigid, well-preserved material objects (CIDOC CRM). Secondly, there was the need to distinguish carefully between the compilation and consolidation phases in information integration.

Both models emerged from this process enriched, better founded and more validated than would have resulted from a merge based solely on initial definitions. Beyond accepting the solutions of both models and identifying the unions and intersections of the

overlapping concepts, we found it beneficial to extend or generalize the models and even to modify the initial positions of both models.

We regard both the methodological experience and the final ontological product as useful outcomes. Nothing in the process seemed to be in any way specific to the addressed domains, except for the belief in a common underlying conceptualization. The final product is a common core ontology for the integration of data and metadata which describes digital library objects, multimedia objects, museum, library and archives information, from a historical, record-keeping perspective. It applies to a very wide set of domains and a vast range of information resources.

The partners of this collaboration are all heavily involved in interoperability and standardization efforts. We expect that both the impact of this work and the productive collaboration between the respective communities will continue. In particular, the work described here has triggered further investigations into a number of ontological issues including the nature of *types* as classes and objects of discourse and the incorporation of models for rights management, legal scenarios and future or potential events.

Acknowledgements

Support for the harmonization meetings described in this paper came from the DELOS Network of Excellence on Digital Libraries an initiative funded by the European Commission's Information Society Technologies 5th Framework Programme (IST-FP5). Work on the CIDOC CRM is supported by The Research Libraries Group, ICS-FORTH and the European Union. Work on the ABC ontology is supported by NSF Grant 9905955, JISC Grant 9906 and the Cooperative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme. The authors also wish to thank the other participants in the harmonization workshops for their valuable contributions to this effort: Thomas Baker, Dan Brickley, Donatella Castelli, James Cheney, Nicholas Crofts, Aldo Gangemi, Tony Gill, Nicola Guarino, Stephan Koernig, Carlo Meghini, Wolfgang Meier, Libby Miller and Nikolay A.Skvortsov.

References

[1] *DELOS Network of Excellence for Digital Libraries*, http://delos-noe.iei.pi.cnr.it/.

[2] *Dublin Core Metadata Initiative*, http://dublincore.org.

[3] *Extensible Markup Language (XML)*, 2002 http://www.w3.org/XML/.

[4] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions March 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

[5] *The Harmony Project*, http://metadata.net/harmony.

[6] "ICOM Statutes," International Council of Museums 2001. http://icom.museum/statutes.html.

[7] "IMS Learning Resource Meta-data Best Practices and Implementation Guide," IMS Global Learning Consortium, Version 1.0 - Final Specification August 20 1999. http://www.imsproject.org/metadata/mdbest01.html.

[8] *INDECS Home Page: Interoperability of Data in E-Commerce Systems*, http://www.indecs.org/.

[9] "MPEG-7 Overview," International Organization for Standardisation, Klangenfurt ISO/IEC JTC1/SC29/WG11, July 2002. http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

[10] *OntoWeb Working Group on Legal Ontologies*, 2002 http://saussure.irmkant.rm.cnr.it/onto/legontoweb.html.

[11] C. Bekiari and M. Doerr, "Documentation and Reasoning on Parts and Potential Wholes," presented at Computer Applications in Archaeology Conference, Dublin, Ireland, 1999.

[12] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*(50), May, 2001.

[13] J. L. Borges, *Other inquisitions, 1937-1952. Translated by Ruth L.C. Simms. Introd. by James E. Irby*. Austin: University of Texas Press, 1964.

[14] G. C. Bowker and S. L. Star, *Sorting things out : classification and its consequences*. Cambridge, Mass.: MIT Press, 1999.

[15] D. Brickley and R. V. Guha, "Resource Description Framework (RDF) Schema Specification," World Wide Web Consortium, W3C Candidate Recommendation CR-rdf-schema-20000327, March 27 2000. http://www.w3.org/TR/rdf-schema.

[16] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Description Logic Framework for Information Integration," presented at 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98), Trento, IT, 1998.

[17] J. Clark, "XSL Transformations (XSLT)," World Wide Web Consortium, W3C Recommendation REC-xslt-19991116, November 16 1999. http://www.w3.org/TR/xslt.

[18] N. Crofts, I. Dionissiadou, M. Doerr, T. Gill, S. Krause, M. Stiff, and S. Stead, "CRM Scope Definition," CIDOC Conceptual Reference Model Special Interest Group, Proposal of the Steering Committee of the CIDOC CRM SIG 2001-07-7 2001. http://cidoc.ics.forth.gr/crm_scope_definition.html.

[19] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff, *Definition of the CIDOC object-oriented Conceptual Reference Model and Crossreference Manual*, 2002 http://cidoc.ics.forth.gr/crm_definition_documents/cidoc_crm_3.3.2/cidoc_crm_title.htm.

[20] M. Dean, D. Connolly, F. van Harmelen, M. R. Henzinger, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, "OWL Web Ontology Language 1.0 Reference," World Wide Consortium, W3C Working Draft WD-wl-

ref-20020729, July 29 2002. http://www.w3.org/TR/2002/WD-owl-ref-20020729/.

[21] L. Dempsey, “Divided by a Common Language: Digital Library Developments in the US and UK,” presented at JISC/CNI Conference, Edinburgh, 2002.

[22] M. Doerr, “The CIDOC CRM - an Ontological Approach to Semantic Interoperability of Metadata,” *AI Magazine - Special Issue on Ontologies*, March, 2002.

[23] M. Doerr and N. Crofts, “Electronic Esperanto: The Role of the Object-Oriented CIDOC Reference Model,” presented at ICHIM'99, Washington, DC, 1999.

[24] M. Doerr, D. Plexousakis, and C. Bekiari, “A Metamodel for Part-Whole Relationships for Reasoning on Missing Parts and Reconstruction,” presented at ER-2001, Yokohama, Japan, 2001.

[25] U. Eco, *The search for the perfect language*. Oxford, UK ; Cambridge, Mass.: Blackwell, 1997.

[26] P. Gerstl and S. Pribbenow, “A Conceptual Theory of Part-Whole Relations and its Applications,” *Data and Knowledge Engineering*, 20 (3), pp. 305-322, 1996.

[27] P. Ginsparg, “Creating a global knowledge network,” presented at Second Joint ICSU Press - UNESCO Expert Conference on Electronic Publishing in Science, Paris, 2001.

[28] N. Guarino, “Formal Ontology and Information Systems,” presented at FOIS'98, Trento, Italy, 1998.

[29] N. Guarino and C. Welty, “Evaluating ontological decisions with OntoClean,” *Communiations of the ACM*, 45 (2), pp. 61-65, 2002.

[30] J. Hunter, “Enhancing the Semantic Interoperability of Multimedia through a Core Ontology,” *IEEE Transactions on Circuits and Systems for Video Technology*, January, 2003.

[31] ICOM/CIDOC Documentation Standards Group, *CIDOC Conceptual Reference Model*, 1998 http://cidoc.ics.forth.gr/.

[32] C. Lagoze, “Business Unusual; How "event awareness" may breathe life into the catalog,” presented at Bicentennial Conference on Bibliographic Control in the New Millennium, Library of Congress, Washington DC, 2000.

[33] C. Lagoze and J. Hunter, “The ABC Ontology and Model,” *Journal of Digital Information*, 2 (2), 2001.

[34] G. Lakoff, *Women, fire, and dangerous things : what categories reveal about the mind*. Chicago: University of Chicago Press, 1987.

[35] O. Lassila and R. R. Swick, “Resource Description Framework: (RDF) Model and Syntax Specification,” World Wide Web Consortium, W3C Proposed Recommendation PR-rdf-syntax-19990105, January 1999. http://www.w3.org/TR/PR-rdf-syntax/.

[36] P. Le Boeuf, "FRBR and Further," *Cataloging & Classification Quarterly*, 32 (4), pp. 15-22, 2001.

[37] D. B. Lenat, "Cyc: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM*, 38 (11), 1995.

[38] C. Masolo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider, "The WonderWeb Library of Foundational Ontologies," LADSEB-Cnr, Padova, IT, Preliminary Report D17, June 30 2002. http://wonderweb.semanticweb.org/deliverables/documents/D17.pdf.

[39] J. McCarthy, "Programs with common sense," in *Semantic Information Procesing*, M. Minsky, Ed.: MIT Press, 1968, pp. 403-418.

[40] Metadata Ad Hoc Working Group, "Content Standard for Digital Geospatial Metadata," Federal Geographic Data Committee, Washington DC FGDC-STD-001-1998, 1998. http://www.fgdc.gov/standards/documents/standards/metadata/v2_0698.pdf.

[41] R. Motshnig-Pitrik and J. Kaasboll, "Part-Whole Relationship Categories and their Application in Object-Oriented Analysis," *IEEE Transaction on Knowledge and Data ENgineering*, 11 (5), 1999.

[42] J. Mylopoulos, A. Borgida, M. Jarke, and M. Koubarakis, "Telos: Representing knowledge about information systems," *ACM Transactions on Information Systems*, 8 (4), pp. 325-362, 1990.

[43] A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo, "Restructuring WordNet's Top-Level: The OntoClean approach," presented at LREC2002, Las Palmas, Spain, 2002.

[44] J. L. Peterson, *Petri net theory and the modeling of systems*. Englewood Cliffs N.J.: Prentice-Hall, 1981.

[45] S. Weibel and J. Hakala, "DC-5: The Helsinki Metadaa Workshop," *D-Lib Magazine*, 1998 (February).

[46] C. Welty and N. Guarino, "Supporting Ontological Analysis of Taxonomic Relationships," *Data and Knowledge Engineering*, 39 (1), pp. 51-74, 2001.