The Dream of a Global Knowledge Network—A New Approach

MARTIN DOERR
Institute of Computer Science, Foundation for Research and Technology Hellas (ICS Forth)
and
DOLORES IORIZZO
Imperial College, London

Decades of research have been devoted to the goal of creating systems which integrate information into a global knowledge network, yet we still face problems of cross-repository interoperability, lack of public infrastructure, and a coherent research agenda—both theoretical and practical—to face these challenges. Interest in the semantic Web has revived the dream, but many are sceptical. This article offers a breakthrough to problems of semantic interoperability and defends the feasibility of a global knowledge network against traditional counterarguments. It offers a new approach based on (i) interdisciplinary research of scholarly and scientific discourse, (ii) a generic global ontological model based on relations and co-reference rather than objects, (iii) semi-automatic maintenance of co-reference links, and (iv) public engagement in the creation and development of the network.

Categories and Subject Descriptors: D.2.12 [**Software Engineering**]: : Interoperability; H.1.1 [**Models and Principles**]: Systems and Information Theory—*Value of information*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.2.1 [**Database Management**]: Logical Design—*Schema and subschema*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*; H.3.7 [**Information Storage and Retrieval**]: : Digital Libraries; J.5 [**Arts and Humanities**]:

General Terms: Theory, Design, Standardization, Management

Additional Key Words and Phrases: Research requirements, cultural heritage, interdisciplinary research, ontology engineering, information integration, knowledge networks, global information access

ACM Reference Format:

Doerr, M. and Iorizzo, D. 2008. The dream of a global knowledge network—A new approach. ACM J. Comput. Cultur. Heritage 1, 1, Article 5 (June 2008), 23 pages. DOI = 10.1145/1367080.1367085 http://doi.acm.org/10.1145/1367080.1367085

1. INTRODUCTION

The Web has become an indispensable tool of modern culture. To a degree, its initial promise of creating a global network that offers access to the knowledge of the world has been realized. It supports advanced technological research in the sciences, arts, and humanities, but it also has popular appeal (online news, media, telecommunications) and has drawn wide public engagement (Flickr, del.icio.us, Wikipedia). Powerful, but relatively crude, search engines organize the enormous amount of information on the Internet into simple answers to clear cut, search term-based questions. What is deceptive about

Author's address: M. Doerr (corresponding author); email: martin@ics.forth.gr.

this everyday process is that it flattens rather than deepens and improves knowledge since popular search engines enforce a historical perspective; the Web does not support the long-tail effect. Research questions which require more than immediate information are thwarted. For instance, we can easily find documents on the Web about Lucy, the hominid, but we have no direct way to discover the locations of finds of a similar kind. Even more disturbing is the lack of accessibility to important research which predates Web publishing. Information that is not easily available on the Web is less frequently cited by others and often deemed irrelevant. In a recent seminar on the semantic Web, we found that most researchers were unaware of important advances in database integration carried out in the 70's, although it was directly relevant to their work.

What is the problem? Even though information on the Web is densely linked (the average distance between documents is only 7 successive links [Broder et al. 2000]), the information itself is not related in a meaningful way. Current search engines will never be enough because they fail to provide the epistemological and historical context of a question which gives results meaning, they are designed as a tool for information aggregation not knowledge integration. Recent interdisciplinary research on digital libraries and the semantic Web addresses this precise problem, however, no clear research agenda has yet emerged. Mainstream research seems to stagnate around several key paradigms which are left unquestioned or at least not challenged. Carl Lagoze states that ". . the underlying public key infrastructure that was seen as 'essential to the emergence of digital libraries' remains undeveloped. Despite efforts of the W3C's Semantic Web initiative, the holy grail of semantic interoperability remains elusive" [Lagoze et al. 2005]. So the dream of a global network of knowledge, which goes back at least to early research on machine translation [Sowa 1992], seems still to be as far away as ever.

In this article, we offer a new approach to problems of semantic interoperability and the creation of a global knowledge network by offering a critique of past and current research (Section 2), challenging some current preconceptions about the nature of ontologies and the creation of semantic networks, (Section 3), presenting a nearly generic semantic model for summarizing, structuring, and combining existing data, (Section 4), discussing implementation alternatives (Section 5), and proposing a clear research agenda that will prepare the way for more advanced research (Section 6). This research agenda is focused on, but not restricted to, the management of information for cultural-historical research. Nevertheless, it also seems applicable to the integration of information about scientific experimental records and even some business applications.

2. PAST AND CURRENT RESEARCH

A breathtaking amount of digital information, including formally structured documents and databases, multimedia objects, and documents without formal structure are being added to the Web everyday. Digital libraries offer ways of organizing this data deluge to support scientific and scholarly research and publishing as well as to serve a general public. Those who maintain these repositories strive hard to find generic, affordable, and scalable mechanisms to serve a vast range of disciplines. However, old paradigms of the traditional library have not kept pace with the global shift to an electronic medium. Business models and methods of cataloging remain equally outmoded: subject catalogs may have become electronic but the structured metadata about the creation and form of an object remain the equivalents of old library cards. The quantity of incoming information leaves less time for careful cataloging, and once content becomes accessible in a machine-readable form, automated methods take over to find characteristic keywords in texts by statistical means.

Intellectually, there is a slow shift from classifying the material by the context of creation, for example, the discipline, such as '520 Astronomy' [Dewey 2003], to providing more and more detail about the content itself. Still, the dream of every researcher is to be able to find documents that pertain to specific questions, such as "who were the real editors of the Yalta Agreement" or "where was the last case of smallpox." Everyone expects computer scientists to be able to create systems which support these kinds of searches, yet the real meaning of a text and the real things it refers to still seems to disappear behind polysemy, ambiguity, and missing background knowledge. There is to date no general computational model that logically interprets human argumentation in natural language [Fauconnier and Turner 2002] so, for the time being, texts cannot be understood by machines in the same way as humans.

Artificial Intelligence research in the 60's and 70's fell far behind initial expectations, but it did give rise to important work in semantic networks. Semantic networks were conceived in order to solve the problem of machine translation of natural language. As John Sowa [1992] explains, the term seems to have been coined in 1961, but the idea can be traced back to the Greek philosopher Porphyry in his commentary on Aristotle's *Categories* in the 3rd century AD. It was understood that human knowledge forms a huge network of propositions which are connected via their constituents, typically regarded as entities and the relations between them. In a semantic network, it is possible under certain conditions to infer new knowledge from registered facts. We refer to this phase of research as the beginning of the dream of a global network of knowledge. Huge knowledge bases were formed [Cycorp 2006, Wordnet 2006] to describe the general knowledge behind words in common discourse. Despite these successes, machine translation characteristically transports ambiguities of the original into the translated text. A method which could combine integrated knowledge in our digital resources into a global, machine-readable network of knowledge has never been fully realized. Hypertext links, which form the main structural paradigm of the Web at present, allow one to navigate manually in a network of information, but the links will never form a global semantic network.

There are scientists, scholars, and business people who communicate successfully via millions of databases, which make the meaning of any single bit of information explicit and machine-readable by application-specific data structures in stand-alone systems. Yet new information systems, often employing idiosyncratic structures and new ways of encoding the same information, are boundless. Consequently, encoded information often cannot be accessed, compared, or combined across heterogeneous information systems in a common way and this produces a silos of information problem.

Research in database integration in the 80's, based on a paradigm of algebraic transformations, did not really solve the problem. In the early 90's, the first successful full-scale information integration systems appeared and most were based on extensible knowledge representation models as global schemas (e.g., Lu et al. [1996]; Levy et al. [1996]; Bayardo et al. [1997]). The actual integration of the various idiosyncratic systems in a large institutional setting so that databases could talk to one another turned out to be extremely labor intensive. Standardization to a common schema is often counterproductive since it freezes the state-of-the-art and does not recognize special cases/exceptions.

Since the mid-90's, it has been widely accepted that explicit knowledge representation models of a domain discourse (i.e., how the human mind perceives the possible states of affairs of a domain) are the only way to connect our digital resources in a meaningful way [Gruber 1993]. Only these models have the necessary general validity and flexibility to deal with the diversity and evolution of information. These categories of thought, or ontological structures, are not immediately obvious even to the expert; they need to be engineered by eliciting knowledge from verious experts and consolidated into a formal logic, a process which requires considerable skill and interdisciplinary knowledge.

Berners-Lee and Fischetti's book [1999] and the famous paper by Berners-Lee et al. [2001] on the Semantic Web set forth the grand vision and a new research agenda was on the horizon. Once the designers of information systems could formulate the meaning of data structures in terms of ontologies, it should then be possible to merge these ontological structures with information systems [Smith 2003], including the vast amount of semistructured information which populates the Web, and combine them with natural language processing techniques, thereby making operational for the first time a global network of knowledge. In theory this sounded very promising. Yet 8 years later, many researchers are sceptical and still puzzled by the immense variety of concepts which govern human cognition. Many have concluded that human conceptualization is highly individual and specific to domains of discourse [Aberer et al. 2004]. As a result, a large part of semantic Web research deals with the correlation or mapping of domain-specific terminologies that are used in comparable contexts [Schorlemmer and Kalfoglou 2005]. In our view, it is this almost exclusive focus on terminology rather than ontological structures that has proved a distraction to following other necessary and complementary lines of research.

3. PRECONCEPTIONS AND COUNTERPOSITIONS

There are a number of preconceptions that guide current research which actually pose obstacles to progress.

(1) Semantic integration requires a generic top-down approach, not one dependent on domain-specific needs.
(2) Ontologies are huge, messy, idiosyncratic and domain dependent, so mapping between them is the only generic answer.
(3) Information integration is about the integration of ontologies and individual classes.
(4) More automated reasoning is needed to exploit our digital knowledge.
(5) Manual work is not scalable or affordable.

3.1 The Top-Down Approach

A generic top-down approach, not one dependent on domain-specific needs, is required for semantic integration of digital libraries.

Counterposition. In principle it is true that we need a generic approach. Generic solutions are generally quick and cheap, but they have a limited life span as demonstrated by statistical methods of information retrieval or the hypertext model. The intrinsic failure of any top-down approach is that its initial conceptualization can never fully anticipate future problems and it will never support semantic interoperability.

The pitfalls of a top-down approach for digital libraries can be illustrated by the limits of the Dublin Core (DC) metadata element set (DCMI 2006). It is an excellent simplification of bibliographic information that provides a unified data structure to all kinds of materials. However, after the initial definition in 1995 pertained only to “the essential features of electronic documents” (http://dublincore.org/workshops/dc1/report.shtml), more and more cases were squeezed under the same umbrella, regarding the DC as the suitable framework for wide extensions [Weibel et al. 1997]. So in the end, quite a lot of domain-specific interpretations of seemingly common metadata elements became mutually incompatible, and the usefulness of the DC broke down. (E.g., instance, we have witnessed a serious application of the DC that named an excavating archaeologist as ‘DC.creator’ of a set of artifacts). Attempts to qualify the Dublin Core Elements only increased heterogeneity so qualified DC was officially abandoned by the DC Consortium.

It is simply wrong to think that all generic solutions need to be top-down. We have created a bottom-up ontology that starts with the analysis of real research scenarios and information management practices in different domains. It is based on deep knowledge of the engineering of interdisciplinary work that generalizes domain-specific cases in order to find the most generic ontological structures and generic processes across multiple domains. We encourage more work of this kind.

3.2 Mapping Between Ontologies

Ontologies are huge, messy, idiosyncratic and domain dependent, so mapping between different ontologies is the only generic answer.

Counterposition. From an empirical analysis of ontologies that are used in information systems, there seem to be two distinct kinds of ontologies (see Magkanaraki et al. [2002]): (a) ones that consist mainly of an individual concept, and (b) ones that try to capture the underlying structure of data on a more generic level of analysis.

The ontologies which consist mainly of individual concepts are characteristically used to structure data in data fields such as type, category, object name, or role and refine the characterization of referred entities. Let us call these '*terminological ontologies*'. They are often organized around complex terminologies and thesauri. Characteristically, the WC3 has proposed the RDF schema SKOS Core for dealing with concepts as instances of RDFS classes. These ontologies tend to be huge, and their structure is dominated by the IsA relationship. For instance, UMLS [2006] integrates 5 million medical-pharmaceutical terms from hundreds of source vocabularies into about 1 million concepts.

By contrast, there are other kinds of ontologies that try to capture the generic level of analysis of a data schema or data structure. Let us call these *core ontologies*. For instance, UMLS is structured by a core ontology of roughly 130 semantic types and about 50 semantic relations. These relationships are generic and not domain-specific, but they are very powerful. The CIDOC CRM is a core ontology (Section 4) that abstracts hundreds of schemata used for documentation in various museum disciplines into 80 classes and 130 relationships. Yet we have found that less than five percent of its concepts can be regarded as specific to museums.

Most researchers fail to recognize the distinctive character of core ontologies for schema semantics. Core ontologies are not huge and messy; they are small, compact, and focused on relationships not objects. It is often argued that IsA hierarchies of a domain terminology structure a domain. This is certainly true, but what they do not take into account is the internal information structure itself. The internal structure of information is defined only by relations (or attributes) connecting particulars and not by individual classes or terms [Wittgenstein 1984; Storey 2005], therefore it is quite natural that core ontologies focus on relationships. In our experience, relationships, and the classes they relate, are rarely domain-specific. They pertain to generic kinds of discourse such as location, participation, part-decomposition, and reveal generic structures that integrate both factual and categorical knowledge in a way that is useful even for very specific applications. We need interdisciplinary work to discover and engineer these relationships. It would be cost effective to manually harmonize, merge, or integrate them in the best possible way since they are small and manageable yet widely applicable. Requirements for this work have already been expressed by Guarino [1998] and Smith [2003]. It is important, however, to point out that our core ontologies should not be confused with the *foundational ontologies* as presented in Masolo et al. [2001], which contain generic relationships, but lack relevant empirical support.

3.3 Information Integration

Queries are mainly about things of a given class. The main challenge of information integration is the integration of ontologies and individual classes.

Counterposition. We believe the salience of this simple query paradigm is not sufficiently supported by empirical studies. Yet in a curious way it points to something that is psychologically significant since it expresses an expectation of what databases should be able to provide. A typical research question contains a "why" question, for instance, "why did the Chinese stop seafaring in 1425?". We believe queries pertain more frequently to particulars than to universals, at least this is the case in history and culture.

Surprisingly, there are no reliable studies about the formal structure of typical user questions or long-term research questions. Many user studies are either based on interviews, and hence elicit only intuition, or are based on observing user behaviour with existing information systems, and hence elicit only reactions to the design of particular systems. Dworman et al. [2000] came to a similar conclusion following unpublished user studies (private communication) that *pattern-oriented retrieval* (i.e., search for relationships in documents) can describe a great deal of the more complex questions. In the absence of real user studies, one may argue that there is no scientific basis for the design of our current information systems with respect to the queries users want to be able to process about factual relations. Recently, Cardoso and Sheth [2006] have stressed the extraordinary importance of access to factual relationships for the Semantic Web, in particular with respect to business applications.

We need to systematically analyze the workflow of researchers as well as the original research questions they ask at each phase of their research process. We believe that the ability to track ever deeper paths of relationships is a necessity for grappling with the complex queries that drive primary research questions, such as "is their a reality to Noah's flood?" [Ryan and Pitman 1998]. So, if relationships are actually more relevant for our queries, and those relationships are limited in number, then we now have a chance to create systems with powerful core ontologies that could constitute a conceptual model for global networks of knowledge.

3.4 The Coreferencing Problem

We need more automated reasoning in order to exploit our digital knowledge.

Counterposition. This is true. But before any reasoning can be done, data must be correctly identified and connected. Therefore, we must define global relationships and not just individual classes. We can do this by first defining identity and coreference conditions since otherwise it is unclear what we are actually connecting. This problem is often overlooked by researchers, especially in the Semantic Web community. The tendency is to regard this problem as a question of dirty data, as the term data-cleaning suggests, which can be overcome in an intermediate step of data processing or by the good practice of using correct or unique identifiers. At present, there is no theory of what ultimately allows us to conclude *coreference*, i.e., that two disparate resources refer to a common item. One may argue that questions of identity can be solved by comparing properties (see Guarino and Welty [2001]). But, since a computer cannot witness reality, reasoning can only be based on relations to other items for which identity has already been determined. So we remain at the same point and no progress is made. It is high time to develop a theory of coreference negotiation. In Section 4, we elaborate in detail on our solution to the coreference problem.

3.5 Harnessing Manual Work

Manual work is not scalable or affordable.

Counterposition. Billions of people produce content manually on the Web everyday and we need to think of ways to harness this effort as a resource. Wikipedia demonstrates that manual work is indeed scalable and cost effective. The question of how to employ manual work in the integration and maintenance of digital knowledge is a social problem not a technical one. We need to design reliable interactive processes and reward mechanisms that draw in researchers, experts, virtual communities and organizations on a massive scale for cataloging, coreference detection and ontology development.

The idea is not to do most of the work manually. We suggest the development of semiautomatic algorithms that support automated methods which would flag two items that might be identical and register an association attribution with a belief ranking which could then be checked and verified manually either by an expert or by someone else with a reliability ranking. As we have learned from machine translations that support human translation (as pretranslation), systems become more effective when trained manually for a special domain, ontology creation, ontology mapping, and coreference negotiation of particulars. The best method would be one which combines manual and automated processes.

In summary, these counterpositions lead us to believe in the feasibility of large-scale knowledge management systems that have the potential to transform the technical, social, and political structure of the Internet which would then make a global knowledge network possible.

4. A NEARLY GENERIC INFORMATION MODEL

In this section, we present an application of the counterpositions presented in Section 3 to a concrete example from the CIDOC CRM, a core ontology of the kind proposed in Section 3.2, as a nearly generic model. We describe the context and process of its creation and the motivation behind its development as a generic model, then we go on to propose a powerful information architecture that could emerge based on the CRM, and finally, we discuss the general processes necessary to maintain and support a global knowledge network.

4.1 The CIDOC CRM—Engineering Core Information Structures

The CIDOC CRM is a formal ontology [Crofts et al. 2005] intended to facilitate the integration, mediation, and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields as diverse as computer science, archaeology, museum curation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It started bottom-up, by reengineering and integrating the semantic contents of a multitude of database schemata and documentation structures from all kinds of museum disciplines, archives, and more recently libraries. The very first schema analyzed in 1996, the CIDOC Relational Data Model with more than 400 tables [Reed 1995], was reduced in 1996 to a model of about 50 classes and 60 properties with a far wider applicability than the original schema. Now, the model contains 80 classes and 132 properties, representing the semantics of hundreds of schemata. The development team applied strict principles to admit only concepts that serve the function of global information integration, and other more philosophical restrictions about the kinds of discourse to be supported (for more detail, see Doerr [2003]).

The application of these principles was successful in two ways. First, the model became very compact without compromising adequacy. Second, the more schemata were analyzed, the fewer changes were needed in the model (see version history CIDOC CRM [2006]. This experience convinced CIDOC in 2000 to begin the ISO standardization process and the model was accepted in September 2006 as ISO21127.

A competitive core model, the ABC Harmony model, developed independently by the digital library and multimedia communities, was harmonized with the CRM in 2001 [Doerr et al. 2004], enriching the CRM with some abstractions of material and immaterial things. Currently, the model of library concepts maintained by the International Federation of Library Associations (IFLA), the FRBR and FRAD model [LeBoeuf 2005], has been formulated as a specialized version of the CRM without any changes. [Doerr and LeBoeuf 2006]. This ease of convergence, even with models from new domains, is encouraging evidence that the CRM captures nearly generic concepts beyond its original limited scope.

Three ideas are central to the CRM.

(a) The relationship between entities and the identifiers that are used to refer to the entities, and the ambiguity of reference, are part of the historical reality that is to be described rather than to be resolved in advance. Therefore, the CRM distinguishes nodes representing real-world items from nodes representing names per se (see Section 4.2).

(b) Types and classification systems are not only a means of structuring information about reality from an external point of view, but they are also part of the historical reality itself as a human invention. Similarly, all documentation is seen as part of a reality and may be described in conjunction with the documented content.

(c) A characteristic way to analyze the past is to divide it up into discrete events. The documented past can be formulated as events involving *persistent items* (continuants or endurants) [Crofts et al. 2005], both material (Ceasar, Lucy) and immaterial (the Empire, Hominid). Material and immaterial items can be present in events either through physical information carriers or as concepts.

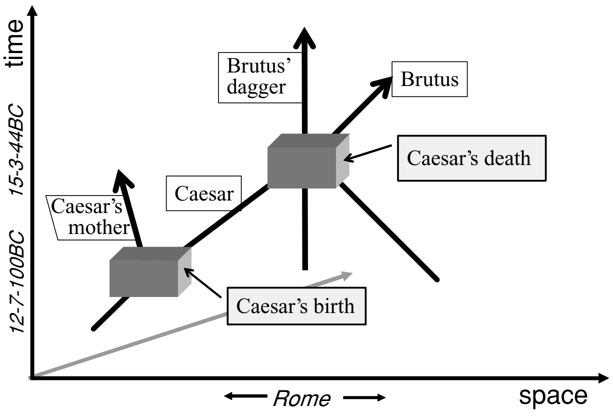


Fig. 1. Historical events as meetings of things and people.

From this point of view, a picture of history as a network of lifelines of persistent items meeting in events in space/time emerges (Figure 1). This abstraction turns out to be extraordinary powerful. Many intuitive relationships are analyzed in terms of events, such as "has creator" or "has origin." With a minimal schema, a surprising wealth of inferences arise and any event can be described by the CRM. For instance, complex genetic family relations can be represented by birth events from a father and a mother. The friend-of-a-friend application (FOAF) [Dodds 2004] can be based on co-authoring and other common events between people. Influences on lives and achievements can be traced to people meeting or communicating with other people and the development of ideas, theories, and discoveries that lead back to them. Chronologies can be justified by the causal ordering of events [Doerr et al. 2004]. Experimental knowledge in the sciences is gained by actual human experiments that are carried out by individuals and teams of researchers in space/time; they can be documented as events, independent of subject matter, calculating statics of bridges or climate models that are not covered by ontologies but can be documented as events. Descriptive sciences, like geosciences and biodiversity studies, gain knowledge by collecting an immense number of observations from individual scientists and research teams, that can be described as events on a human scale connected to people and ideas. Embedded in all metadata stored in libraries, including digital libraries, is a historical perspective which can be described as events.

The key ideas of the CRM have been compressed into an extraordinarily small metadata schema, the CRM Core (see Appendix) [Sinclair et al. 2006; Doerr and Kritsotaki 2006] with 20 fields which maintain the potential to create huge meaningful networks of knowledge about correlated events that support powerful queries. We assume, that it is not the simplicity of the applications we have studied but the quality of the abstractions we have developed over more than a decade that uphold the integrity of the compressed model.

In summary, we have shown that at least in the case of the CRM, a bottom-up engineering approach can lead to a nearly generic information model of extraordinarily wide applicability, robust against further extensions of scope, and superior in its expressive power to top-down approaches.

In the following sections, we investigate what is necessary to create global knowledge networks based on the CRM or similar models.

4.2 The Linking Paradigm—Documents and Knowledge

In this section, we develop a model of how to semantically connect documents in a way that is diametrically opposed to the current hypertext paradigm. Using a minimal but central part of the CIDOC

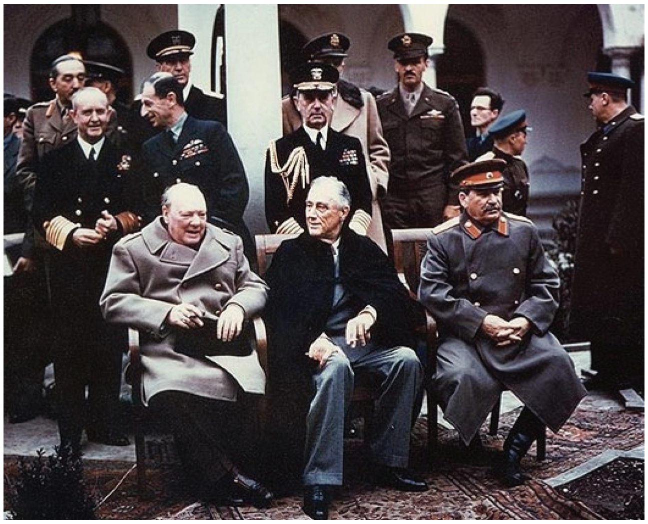


Fig. 2. Allied leaders at Yalta.

CRM as an example, we elaborate the problem of extracting knowledge from the contents of documents, metadata and links between documents into a coherent semantic network.

4.2.1 *A Minimal Network Model.* Let us examine the following model of three classes and two properties from the CIDOC CRM:

E5 Event. *P12 occurred in the presence of*: E77 Persistent Item
E5 Event. *P7 took place at* : E53 Place

The CRM class E77 Persistent Item comprises material and immaterial things, including people. Consider now the following data and metadata records. The State Department of the United States holds a copy of the Yalta Agreement. One paragraph begins, "The following declaration has been approved: The Premier of the Union of Soviet Socialist Republics, the Prime Minister of the United Kingdom and the President of the United States of America . . . jointly declare their mutual agreement to concert . . . " [Halsall 1997]. A Dublin Core record about this text might read as follows.

Type:	Text
Title:	Protocol of Proceedings of Crimea Conference
Title.Subtitle:	II. Declaration of Liberated Europe
Date:	February 11, 1945.
Creator:	The Premier of the Union of Soviet Socialist Republics The Prime Minister of the United Kingdom The President of the United States of America
Publisher:	State Department
Subject:	Postwar division of Europe and Japan

The Bettmann Archive in New York holds a world-famous photo of this event (Figure 2). A Dublin Core record for Figure 2 might be as follows.

Type:	*Image*
Title:	*Allied Leaders at Yalta*
Date:	*1945*
Publisher:	*United Press International (UPI)*
Source:	*Wikipedia*
References:	*Churchill, Roosevelt, Stalin*

Another piece of information comes from the Thesaurus of Geographic Names [TGN], which may be captured by the following data.

TGN Id:	*7012124*
Names:	*Yalta (C,V), Jalta (C,V)*
Types:	*inhabited place(C), city (C)*
Position:	*Lat: 44 30 N,Long: 034 10 E*
Hierarchy:	*Europe (continent) <– Ukrayina (nation) <– Krym (autonomous republic)*
Note:	*Located on S shore of Crimean Peninsula; site of conference between Allied powers in WW II in 1945; is a vacation resort noted for pleasant climate, & coastal & mountain scenery; produces wine, canned fruit & tobacco products.*
Source:	*TGN, Thesaurus of Geographic Names*

It has long been recognized that the only element that is common to all of these records in the date 1945, and that is why a Google search for the Yalta Agreement will not be adequate.

We can represent most of the information from these three sources as instances of 3 Classes and 2 Properties of the CIDOC CRM.

(1) *Crimea Conference (E5)*
P12 occurred in the presence of
The Premier of the Union of Soviet Socialist Republics (E77)
The Prime Minister of the United Kingdom (E77)
The President of the United States of America (E77)
Protocol of Proceedings of Crimea Conference (E77)

(2) *Allied Leaders at Yalta (E5)*
P12 occurred in the presence of
Stalin (E77)
Churchill (E77)
Roosevelt (E77)
Photo of Allied Leaders at Yalta (E77)
P7 took place at
Yalta (E53)

(3) *Yalta Conference (E5)*
P12 occurred in the presence of
Allied Powers (E77)
P7 took place at
Yalta(E53)

If we can resolve in sequence the different ways of referring to the same items, the uncorrelated parts will collapse into a single network, which connects the text, the image, the place, and the people through the historic event.

(4) *Yalta Conference (E5)*
P12 occurred in the presence of
Stalin, Premier of the Union of Soviet Socialist Republics (E77)
Churchill, Prime Minister of the United Kingdom (E77)
Roosevelt, President of the United States of America (E77)
Protocol of Proceedings of Crimea Conference (E77)
Photo of Allied Leaders at Yalta (E77)
P7 took place at
Yalta(E53)

4.2.2 *Relations and Documents.* If we are able to collect enough related events even this rudimentary schema would create a powerful network for collecting biographical and contextual data about people and documents, objects, and places. The compressed CRM Core metadata schema allows for classifying all referred items, roles in events, and part-whole relationships, resulting in readable descriptions of events and things. The full CIDOC CRM has approximately 130 relationships and covers a high level of detail with respect to elaborate database schemata. However, the full and compressed versions both rely on the same principles. What we learn from this example is the following.

(a) A knowledge network must be built on suitable ontological abstractions that support relevant relationships. These can be surprisingly simple yet powerful. Explicit event representation seems to greatly reduce the complexity of necessary relationships.
(b) Little advanced reasoning can take place if the elements of the network are not connected. They connect through the domain and range values of the relations that identify items in a domain of discourse. The identifiers are normally not unique and therefore don't match. This duplicate removal or co reference detection is a process widely underestimated in importance for information integration (see Section 4.3).
(c) Knowledge about relationships comes from the document, either from its proper contents or its metadata, or it is derived from another source. What actually relates the documents is not a hyperlink, but the fact that they refer to the very same items. These may be events, dates, places, people, material or immaterial things such as texts, images, names etc. Even terms can often be seen as (conceptual) items of discourse, rather than as expressions of classification or reasoning.

Since the connecting facts are not revealed in the hyperlink, the hypertext model is fundamentally limited to manual navigation. Equally misleading is the paradigm of a document as a digital surrogate of a real-world item, which is one of the motivations for the RDF syntax. There is a problem, however, about which of the documents, out of all the documents about a real-world item, should become the surrogate; how should the competition between the properties of the surrogate and the thing itself be resolved?

We suggest that appropriate digital surrogates of real-world things should be modeled as nodes external to the documents, bare of any necessary property besides an identity as described in Section 4.3. Let us call them in the sequence *surrogate nodes*. The *relations* between nodes should be seen as *extractions* or summarizations from the documents (see Figure 3). Let us call these *facts*. By facts [Degen et al. 2001] we mean the instances of relationships (or properties in the terminology of OWL and RDFS). Constructs like the reification in RDF and other argumentation models [Roux and Blasco 2004] make explicit the link between the source provided and the relation explicit.

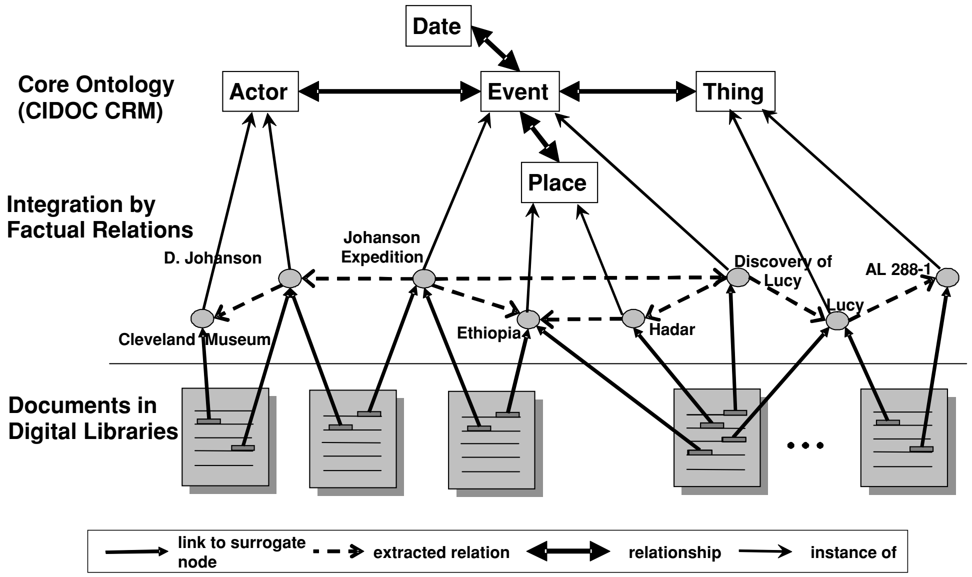


Fig. 3. Relations as document summarization.

4.3 Knowledge Management Based on Coreference Links

As already shown, even if we have a global schema and the means to provide factual relations, one important feature that is necessary to build a network is missing; we do not know if two relations relate to the same item. This important problem has not received much attention but has been described by Levesque [1984] as coreference. In the next section, we propose a novel model to achieve long-term scalable integration of facts.

In the context of our daily lives, we communicate through the consistent use of names and identifiers, and we resolve ambiguities through dialogue. It is normal practice for people entering data into a well-maintained database to make sure that exactly one identifier is used for each real-world item so that the identifiers reflect reality. Normally, those who maintain databases form a community, so that doubts about the correct use of identifiers can be resolved between the users. For instance, we would expect that an administration database of a business has unique identifiers for each of their employees. However, even under conditions of good practice, we do not know if two independently maintained databases refer to the same real-world items or even if they use the same identifier. Given the variability of human language, there is nothing that would allow us to be sure about the meaning of any word and, for most particular items (i.e. things and events), there are not even specific names in use.

Librarians and others have invested heavily in so-called authority files or knowledge organization systems (KOS) [Patel et al. 2005], which register names and characteristics of authors and other items and associate them with a preferred representation in a central resource, and then advise colleagues to use the central resource as a reference to obtain unique identifiers. In one respect, this does not solve anything since we still cannot determine if a local source refers to an item also listed in the KOS. In another respect, the approach has been partially successful. The descriptions increase the chance that an expert of the local source can recognize the item, use the identifier, and then pass this on to colleagues who will also use the identifier for the same item. But using a central resource causes serious scalability problems. The standardization process always lags behind reality. Even worse, different communities in different countries tend to create their own authority files with overlapping content so once again we must start from the beginning.

Computer scientists tend to regard the recognition of coreference (duplicate detection) as a question of probability that two items are referred to by similar names or similar properties, for example, Bilenko and Mooney [2003]. They employ various heuristics that compare the attributes by which items are characterized in different data sources and compute a probability that they mean the same thing (e.g., it is reasonable to assume that two people with the same name, same birth date, and birth place are actually identical, but they leave it to the user to believe the association or not). Unfortunately, databases tend to register different properties of the same things or properties that may change over time so that it is rare for properties to be closely compared. Even worse, duplicate detection requires a certain preexisiting degree of integration: for instance, does the place name for a birth place in one record refer to the same place as the placename in another record? In the end, the pure probability of two different items having the same properties is messed up with all kinds of systematic errors which are very difficult to treat systematically.

What is common to both approaches is the fact that they do not preserve actual knowledge that an identifier *a1* in source *s1*, and an identifier *a2* in source *s2* refer to the same real-world item. If we make the assumption that the maintainer or creator of *s1* knows what *a1* means, and the maintainer or creator of *s2* knows what *a2* means, both could convene and record the fact of coreference without any common attribute or authority file. We regard a coreference statement as an elementary piece of scientific or scholarly knowledge, regardless of any heuristic-based software assisting in the identification process, and not as a question of data cleaning. Each coreference statement allows for the connection of all factual relations to the two identifiers involved.

So, if we publish a coreference statement and preserve the referential integrity, we have achieved more than any authority file: we have connected facts from two information assets to the best of our knowledge. In order to create a global knowledge network, we need only publish and preserve each and every detected coreference together with its sources. Then the network will grow simply through the efforts of the users. Nothing like this exists on the Web at the moment, but it is potentially a way of engaging the public, as in Wikipedia, to play a large role in building a global knowledge network. Intuitively, coreference should be transitive and form equivalence classes [Levesque 1984] that could scale up to any size. In order to relate the elements of an equivalence class of cardinality ν, a minimal number of $(\nu\text{-}1)$ primary equivalences is needed to derive all $\nu(\nu\text{-}1)/2$ equivalences. This demonstrates the economic power of preserving coreference knowledge once the networks grow tighter. Each equivalence class can be identified with a surrogate node as described previously. Coreference links can also be implemented indirectly as links to a common surrogate node.

Let us regard a pool of documents. Before any coreference has been detected, each occurrence of an identifier corresponds to another surrogate node. As long as an activity of finding coreferences is maintained, the network has the potential to converge to a state in which each real-world item is represented by exactly one surrogate node, and then complete integration of facts will have been achieved. One can regard the average ratio of surrogate nodes to referred items as a kind of "entropy" measure for the degree of factual network integration. The entropy will increase with each new document introduced into a network pool. Some coreferences may never be resolved due to lack of knowledge. So a competition situation will arise between coreference detection, the *integrative* process, and the introduction of new information, the *divertive* process. The degree of integration is a measure of the quality of the network.

We suggest that coreference detection must be a semi-automatic process. Massive participation of scholars qua experts in this process will be essential since it requires specialized knowledge and should not be left to automated guesswork. As a matter of good practice, it should become a product of scholarly research that is properly documented. Mathematical models could be developed that would estimate the time it took to carry out integration activity and offer a cost-benefit analysis. Further, formal foundations of data-cleaning methods could be investigated, such as the extent to which the propagation of coreference knowledge allows for inferences or assumptions about other coreferences via related facts etc. Finally, mathematical models could be used to develop effective strategies in peer-to-peer networks of coreference detection and monitoring of global consistency.

In this section, we have introduced the idea of a knowledge economy and the long-term integration of digital repositories by preserving knowledge about coreference. This idea is radically new, in four respects.

The ultimate authority for identifier equivalence are people—the witness or the expert—with knowledge of the two contexts that are to be connected. Coreference is a valuable element of knowledge that comes at high cost, therefore it should be curated and preserved and future information systems need to take account of this fact.

The model suggests that several current approaches of ad-hoc data cleaning and central authorities are ineffective and miss an important part of the problem, namely, the preservation and control of real-time detected coreferences.

The co-reference model can be implemented in a completely distributed democratic manner. Therefore, in contrast to other approaches, it is completely scalable and imposes minimal constraints on the kind of organization in which it will be implemented. Much research is ongoing on the management of networks for social tagging or decision making. For instance, Rodriguez et al. [2007] have implemented *Smartocracy*, a trust-based social network combined with a vote-based decision network. This approach shows an interesting way to effectively manage the problem of varying expertise between participants in a social decision network without authoritative control. The suggested model has the potential to converge on the "best knowledge to date."

So far we have deliberately presented a simplistic view of the identity question for the purposes of argument. The concept of identity is philosophically demanding [Wiggins 2001]. We expect this simplistic view to hold in those cases where, for any two identifiers *a1*, *a2*, one can decide whether *a1* is equivalent to *a2* or not, at least if there is common agreement about the interpretation of both sources. This is normally true for references to particulars such as people, geopolitical units at a particular time, vehicles, buildings etc. Yet this may not be the case for concepts [Lakoff 1987] or things with fuzzy boundaries, such as mountains etc. In these harder cases, it is more difficult to agree about transitivity of coreference. Therefore, it is clear that fundamental theoretical research needs to be carried out regarding how the identity of things, described in different contexts, can be compared.

5. IMPLEMENTATION AND ADOPTION

The following section is based on the experience of the authors and others with applications of the CIDOC CRM or comparable systems. A series of applications have already been based on the CIDOC CRM in one form or another. Particularly successful were information integration environments in the cultural domain, which employ relevant parts of the model or its precursors either as common schema (e.g., the musinfo system [Crofts 1999] or as global schema [Kim et al. 2004; Nussbaumer and Hashofer 2007]. These confirm the adequacy of the model. In the CLIO system, [Dionissiadou and Doerr 1994], we found entering data directly into a semantic network ergonomically ineffective. Most applications use the standard as a guide to good practice for conceptual modeling of dedicated information systems (http://cidoc.ics.forth.gr/uses_applications.html). They take parts of the model and customize it according to their needs with the built-in ability to export or query content in terms of the global model. The ubi-erat-lupa project [Doerr et al. 2004] integrates complementary rather than analogous resources under the CIDOC model: there are inscriptions of the names of people on stones on sites, and each relationship is listed in another autonomous resource. Because most identifiers do not match between resources, we had to start implementing a system for semi-automatic coreference detection and correction. This experience was a major motivation for this article. To our knowledge, no information integration system for autonomous resources employs a global model and solves the coreference problem.

So, how can facts be created in an efficient way? Apart from using free text, people like to document and create metadata by filling in forms that are highly specific to the context of their work, and they do it in very large quantity. The role of these forms is often that of a questionnaire which suggests what kind of statements should be made with respect to any given thing. The CIDOC CRM, on the other hand, offers only abstract concepts and by virtue of this it has wider applicability. In our experience, when data is presented to people, they understand quite well if information consisting of names and terms is connected only by relatively abstract relationships and they appreciate graphical or schematic representations, but they tend to get confused, even as domain specialists, when asked to enter data directly into a generic model.

The problem is that a generic model does not suggest what to document in the specific case, it only sufficiently explains what has been documented. It requires constant abstract thinking to match generalizations to specific problems even though the generalizations are quite obvious after one sees them. For instance, finding an object (as in archaeology) in the CRM would be represented as activity in which an object is present. This abstraction is sufficient for most inferences about an archaeological find. The activity type “finding” would be a term entered as data but not as part of the core model. An archaeologist entering data, however, would like to see a field reminding him to enter where and when an object was found. Similarly, other disciplines will have other special data needs. Hence, data entry forms should normally be more application-specific than the generic model even if they are designed to capture data for instantiating the generic model.

It is also good practice for a researcher or documentation specialist to preserve the information unit (s)he has elaborated on as a whole both in order to maintain her/his authorship of the information unit and for future revisions. If data is directly entered into a global semantic network and all knowledge is merged, then the original units are lost. Many traditional relational database schemata are not immune in this respect. Preserving the information unit allows an association to be made between them and the people who understand their interpretation and other relevant knowledge, thereby verifying the quality of the contents.

Finally, large monolithic resources are more sensitive to complete corruption and therefore cause more problems for digital preservation than distributed units. Therefore, we propose to logically separate documentation units and primary sources from the network level, and instead to derive data for the network level from the documentation units and primary sources. Duplication of information establishes good practice for digital preservation.

We distinguish three possible architectures to achieve this separation. They have different performance characteristics but can easily be combined to make optimal use of all of them.

(1) In data warehouse-style, facts can be extracted from sources and physically aggregated in a semantic network. The extracted facts directly connect the surrogate nodes (Figure 4). In order to update the network when sources change, it may be necessary to introduce reification statements [Hayes 2004] or similar mechanisms linking facts to their sources. This strategy makes querying, across resources very fast, especially joins and deductions. Updating is more difficult since individual facts may have multiple sources. Maintaining reification links is relatively expensive. It becomes even more complex when coreference statements are added and linked to the surrogate nodes (see Section 4.3). On the other hand, physically (on a dedicated system) creating the network provides

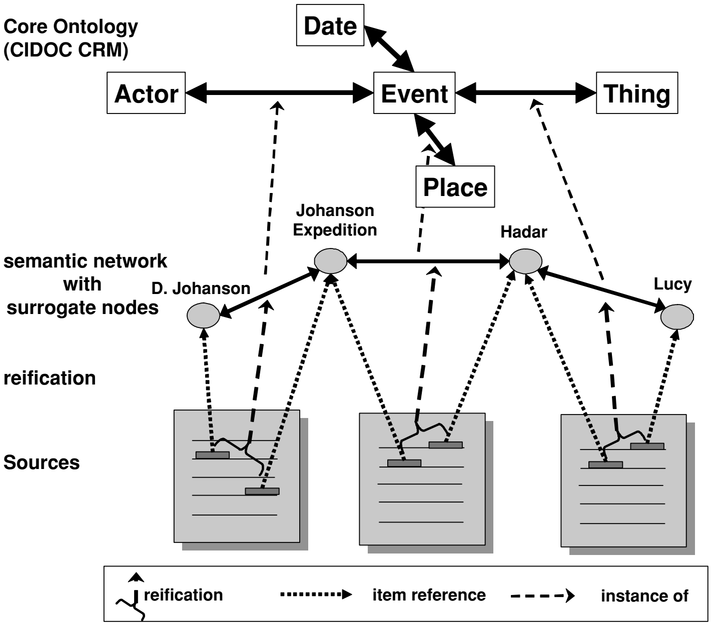


Fig. 4. Semantic network linked back to sources via reification links.

more flexibility to actually detect coreference relations [Doerr et al. 2004] because extraction and aggregation can be done in complex processes. Finally, semantic networks are not scalable or at least no scalable architecture has yet been proposed.

(2) Sources are interpreted by a mediation service [Wiederhold 1992]. For instance, queries are formulated in terms of the global model and transformed according to the different source models to bring back results conforming to the global model (e.g., Calvanese et al. [1998]; Figure 5). Assuming mainly a local-as-view (LAV) approach [Cali 2003], this is only possible if the sources have a data structure which can be mapped to the global model. The performance may depend on the degree of heterogeneity of the local source to the global model. For mediation services, it is more difficult to resolve coreference relations because queries are expected to be answered in real time. This would change completely if explicit coreference relations were available. Joins and deductions are more costly and require larger temporary computer memory. There is no update problem at all.

(3) Whereas the previous solutions were widely discussed in the past decade, we propose yet another variant. Extracted local facts are represented in terms of the global model as summarization metadata units, which are preserved and remain connected to their sources. Then, coreference relations could be described by linking to the surrogate nodes the corresponding local nodes, which in turn are linked by local facts (Figure 6). The surrogate nodes could, for instance, be implemented by one-to-many XLinks. This strategy doubles the path lengths in the network and makes querying slower, but it has the advantages of avoiding both heterogeneity and reification and of offering a scalable solution without central update problems.

We suggest that the architectures described deserve more research about the precise conditions under which they would be most effective, both singularly and in combination. Obviously, querying data paths in solution (3) is more effective the lower the density of coreference relations relative to the number of local nodes and the lower the multiplicity of identical facts between the metadata units

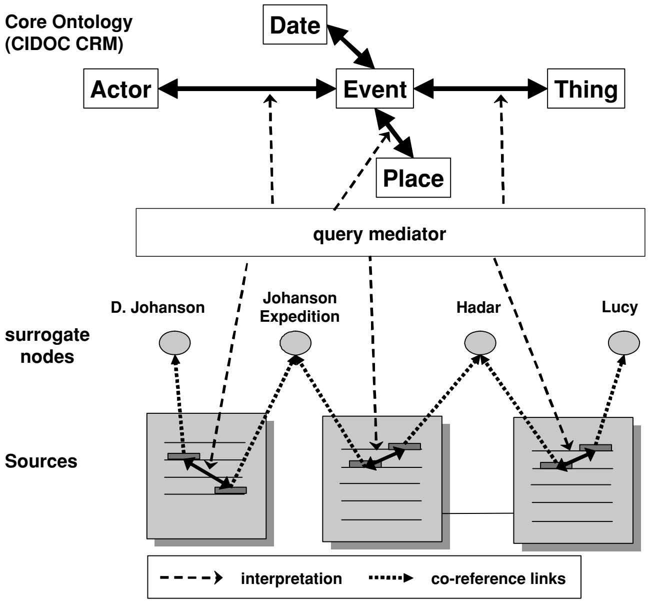


Fig. 5. Query mediator interprets source relations.

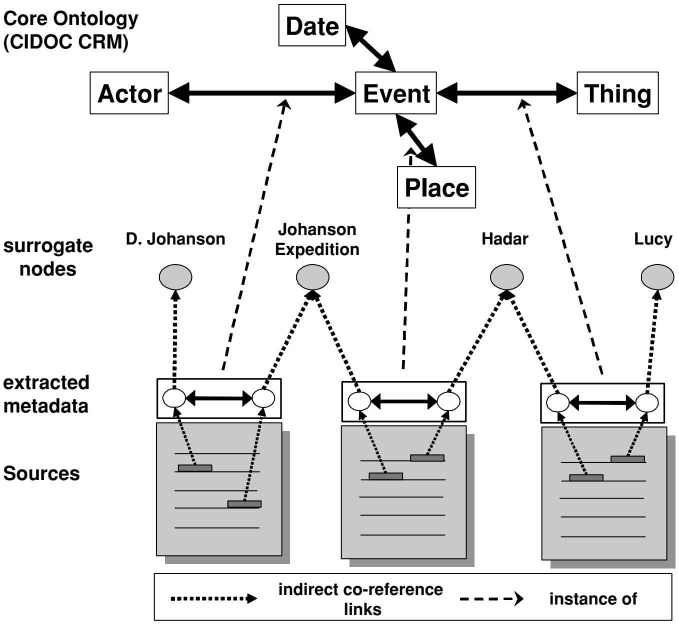


Fig. 6. Metadata connected to sources and indirect coreference links.

because it requires less joins across metadata units. The multiplicity of coreference and facts can be quite characteristic for certain applications. For instance, in descriptions in typical catalogs of art museums, certain facts about artists reoccur many times, whereas other facts about objects are quite unique.

There is nothing to prevent introducing some limited heterogeneity in solution (3) so that solution (3) and (2) become more similar. In new systems, one could design the data structures of local sources right away with minimal heterogeneity and known mappings to the global model. In local environments with a low update rate, solution (1) may be most effective if reification can be simplified. Then, a complete semantic network could take the logical place of a metadata unit in solution (3) Under this circumstance, solution (3) could provide a way to make semantic networks distributed. These are only examples of how the architectures could be combined to produce far more flexible and generic solutions for information integration. For any distributed solution, research about effective indexing would also be a major issue that needs attention (see, e.g., Podnar et al. [2006]).

Solution (3) is particularly suited to natural language processing techniques for knowledge extraction from free text. The CIDOC CRM has a nearly linguistic structure and makes this task relatively easy [Genereux and Niccolucci 2006]. In particular, the event model maps easily to phrases containing action verbs. We suggest that more research be invested in extracting event-based metadata by semi-automatic methods from free text. Far too little attention has been focused on this important problem. [Vincent 2005]. We suggest that peer-to-peer networks and GRID technology can provide an effective infrastructure to run various utilities to build the network create distributed indices, control and monitor its consistency, and manage its convergence to higher states of integration.

6. FUTURE RESEARCH AGENDA

Summarizing, we believe it is time to revive the dream of a global network of knowledge, for the fourth time now, after semantic networks, the Web and the Semantic Web. We have shown that the following factors have not received particular attention in the past and could bring about a breakthrough.

(1) A global model must be relevant, small, and manageable. We achieve this for the CIDOC CRM by an objective relevance criterion, that is, empirical evidence that data are actively structured by experts in these terms, and by a clear functional focus, that is, the aggregation of factual, mesoscopic, discrete knowledge about the past.

(2) The coreference problem needs to be addressed with a scalable solution, bringing precision to the best stage of knowledge.

(3) Global consistency is not a yes or no question but a quality measure. Inconsistency must be effectively managed so that it can be systematically reduced. This holds in particular for coreference and terminology.

Therefore, we believe it is feasible, and no more a dream, to create at reasonable cost factual relations to feed global knowledge networks in the way we have outlined. They will never be globally consistent, but it is possible to initiate processes that will insure the continuous improvement of consistency and integration of results into an ever growing network where success is measured by the number of surrogate nodes compared to the number of real-world items for which they stand. These networks should have the expressive power to provide and aggregate information necessary for the building and validation of scholarly and scientific hypotheses. More than any other discipline, cultural heritage research needs to integrate factual knowledge. If research takes this direction, we can expect a multitude of effective methods to emerge. In order to achieve this goal, we propose the following research agenda.

6.1 Ontology Engineering and Global Models

More systematic research needs to be undertaken involving empirical evidence for global models relating to generic kinds of discourse. Generic principles candidates should be validated in multiple domains. We need a better understanding of the dependency of global models on the kinds of discourse they are meant to support. Which models can be merged? How formal should they be? Many researchers regard ontologies as discipline and discourse neutral. This preconception may prevent us from better understanding the relationship between classes properties, and entities created following different ontological choices.

6.2 Identity Negotiation

There should be more research on problems of coreference negotiation between distributed systems. Processes should be developed to improve the states of knowledge reflected in distributed systems with respect to an assumed reality, taking into account missing and erroneous knowledge. The authors currently work on these problems. More elaborate theories of identity conditions and states of knowledge should inform duplicate detection methods. New architectures for coreference networks should generalize over authority files. Social models should be developed to engage scholars in large-scale maintenance of coreference networks and, more generally, to preserve the personal relation of scholarly work to digital information. There should also be more research on the different notions of identity one would expect in distributed environments and how they could be compared or reconciled for information integration.

6.3 Information Mapping

There should be an increased effort in information mapping. In general, it is impossible to solve all cases of heterogeneity between information sources. We have observed that a carefully crafted global model helps to avoid many cases of heterogeneity; tools tuned to a specific target model and to typical cases of remaining heterogeneity are needed. Mapping presents a major bottleneck since it requires domain experts; therefore, user-friendly interfaces are needed to fully engage and motivate the research community. It should be possible to derive mediation algorithms and data-transformation algorithms automatically from mapping specifications by domain experts and they should be able to validate the mapping by evaluating the results of data-transformation examples.

6.4 Discourse Analysis

All of this research will continue to be speculative or naïve in its aims if we do not begin to analyze systematically the way scholars and scientists work and argue and what kinds of real questions arise at different stages of research, hypotheses building, and validation. We anticipate that the availability of nearly global ontologies geared towards a specific discourse will open up new possibilities of combining argumentation models [Toulmin 1958; Gardin 1990; Streitz 1992; Roux 2004] with the subject matter of argumentation, and thereby gain new insight into the structures and processes of discourse. Real user scenarios and research questions from user studies could be analyzed as operations with an ontology which could be compared to argumentation models. This knowledge could be used to derive requirements for the research previously mentioned. It should, for instance, answer questions about necessary query mechanisms and automated reasoning, the benefit of granularity of ontologies, and the relevance of the ontological constructs we use in particular argumentation processes.

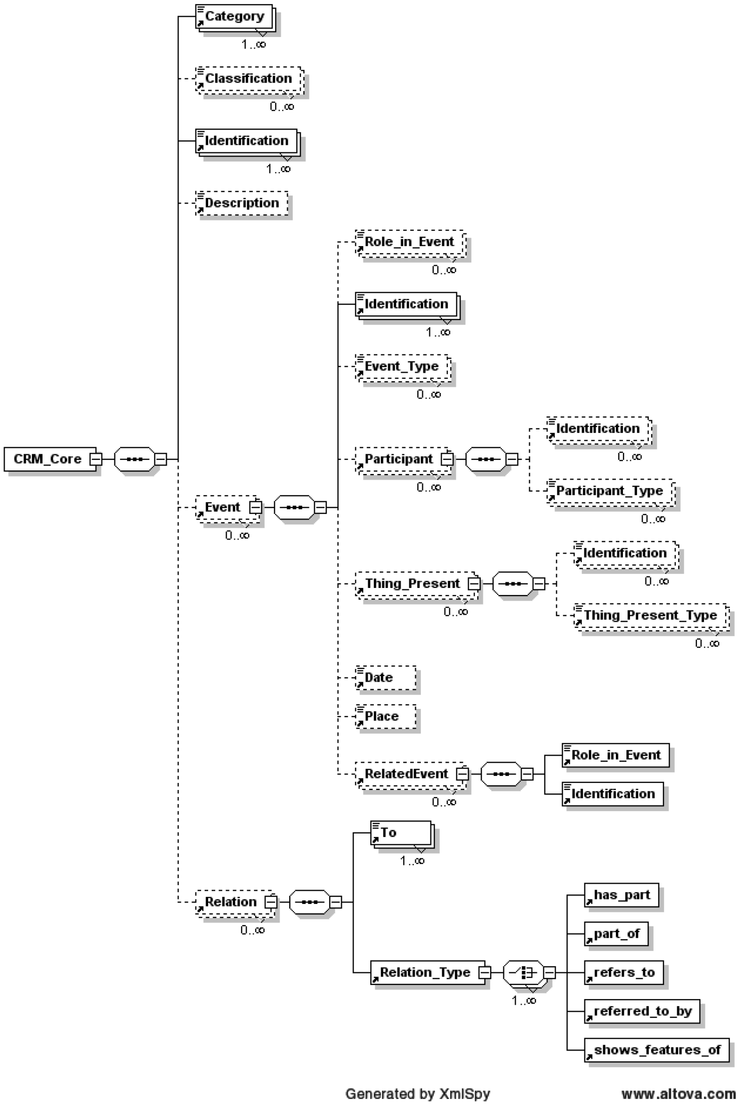
After advocating that it is feasible to create meaningful global networks of knowledge we must clearly state that we do not envisage one single global network of knowledge but rather several very large networks. Nevertheless, one may come to dominate, like the current Web, and others may be partially integrated into it.

6.5 Obstacles

Basically we see two kinds of obstacles, social and technical. The social obstacle is that tight integration requires a commitment to curate data and adhere to some common standards. We would expect that user communities committing to such activities will grow and reach some natural limits related to the effort versus the benefit drawn from the subject matter and quality of information.

A technical obstacle is the nearly generic information model, which will have natural application limitations. But even though the cultural heritage domain is notorious for offering competing points of view and putting forward a proliferation of concepts, the cultural/historical community has agreed upon a compact formulation of common concepts, the CRM. This is cause for optimism, since the ideas outlined could be adopted by this large community and lead to knowledge networks far larger and more powerful than anything we have seen so far. If research in this direction is taken up, we would expect the realization of large full-scale knowledge networks within a decade, and these would give rise to a multitude of research questions that could exploit advanced reasoning methods.

APPENDIX. CRM CORE



REFERENCES

Aberer, K., Catarci, T., Cudre-Maruroux, P., Dillon, T. S., Grimm, S., Hacid, M.-S., Illarramendi, A., Jarrar, M., Kashyap, V., Mecella, M., Mena, E., Neuhold, E. J., Ouksel, A. M., Risse, T., Scannapieco, M., Saltor, F., Santis, L., Spaccapietra, S., Staab, S., Studer, R., and Troyer, O. 2004. Emergent semantics systems. In *Proceedings of the International Conference on Semantics of a Networked World (ICSNW'04).* Lecture Notes in Computer Science, vol. 3226, 14–43.

Bayardo, R. J., Bohrer, W., Bnee, R., et al. 1997. InfoSleuth: Agent-based semantic integration of information in open and dynamic environments. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. vol. 26, 2, 195–206.

Berners-Lee, T. and Fischetti, M. 1999. *Weaving The Web: The Original Design And Ultimate Destiny Of The World Wide Web by its Inventor*. Harper Collins, New York, NY.

Berners-Lee, T., Handler, J., and Lassila, O. 2001. The Semantic Web. *Scientific American*. May.

Bilenko, M. and Mooney, R. J. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*. Washington DC, 39–48.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. 2000. Graph structure in the Web. *Comput. Netw.: Int. J. Comput. Telecomm. Netw. 33*, 1–6, 309–320.

Cali, A. 2003. Reasoning in data integration systems: Why LAV and GAV are siblings. In *Proceedings of the International Symposium on Mothodologies for Intelligent System (ISMIS'03)*. Lecture Notes in Computer Science, vol. 2871, 562–571.

Calvanese, D., Giacomo, G., Lenzerini, M., Nardi, D., and Rosati, R. 1998. Description logic framework for information integration. In *Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98)*. 2–13.

Cardoso, J. and Sheth, A. Eds. 2006. *Semantic Web Services, Processes and Applications.* Springer.

Cidoc, Crm. 2006. The CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr/.

Crofts, N. 1999. Implementing the CIDOC CRM with a relational database. *MCN Spectra. 24*, 1.

Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff M. 2005. Definition of the CIDOC conceptual reference model. http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc.

Cycorp, Inc. 2006. What does Cyc know? http://www.cyc.com/cyc/technology/whatiscyc_dir/ whatdoescycknow.

Dionissiadou, I. and Doerr, M. 1994. Mapping of material culture to a semantic network, In *Proceedings of the 1994 Joint Annual Meeting of the International Council of Museums Documentation Committee and Computer Network*. Washington DC.

Degen, W., Heller, B., Herre, H., and Smith, B. 2001. GOL—Towards an Axiomatized Upper-Level Ontology. *Electron. Comput. Sci*.

Dewey, M. 2003. *Dewey Decimal Classification and Relative Index*. Ed. 22. Vol. 1–4, OCLC Forest Press.

Dodds, L. 2004. An Introduction to FOAF. http://www.xml.com/pub/a/2004/02/04/foaf.html.

Doerr, M. 2003. The CIDOC CRM—An ontological approach to semantic interoperability of metadata. *AI Magazine 24*, 3.

Doerr, M., Hunter, J., and Lagoze, C. 2003. Towards a core ontology for information integration. *J. Digital Inform. 4*, Article 169.

Doerr, M., Plexousakis, D., Kopaka, K., and Bekiari, C. 2004. Supporting chronological reasoning in archaeology. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology Conference (CAA'04)*. Prato, Italy, http://www.ics.forth.gr/isl/publications/paperlink/caa2004_supporting_chronological_reasoning.pdf.

Doerr, M., Schaller, K., and Theodoridou, M. 2004. Integration of complementary archaeological sources. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology Conference (CAA'04)*. Prato, Italy. http://www.ics.forth.gr/isl/publications/paperlink/doerr3_caa2004.pdf.

Doerr, M. and Leboeuf, P. 2006. Modelling intellectual processes: The FRBR—CRM harmonization. In *Conference Proceedings of ICOM-CIDOC Annual Meeting.* Gothenburg, Sweden. 10–14.

Doerr, M. and Kritsotaki, A. 2006. Documenting events in metadata. In *The e-volution of Information Communication Technology in Cultural Heritage,* 56–61.

Dcmi. 2006. Dublin Core metadata initiative, Making it easier to find information. *http://dublincore.org/*.

Dworman, G. O., Kimborough, S. O., and Patch, C. 2000. Pattern-directed search of archives and collections, *J. Amer. Soc. Inform. Sci. 51,* 1, (Special issue. When museum informatics meets the World Wide Web), 14–23.

Fauconnier, G. and Turner, M. 2002. *The Way we Think: Conceptual Blending and the Mind's Complexities*. Basic Books, New York, NY.

P. Leboeuf Ed. 2005. *Functional Requirements for Bibliographic Records (FRBR): Hype or Cure-All?*. Haworth Press, Inc.

Gardin, J.-Cl. 1990. The structure of archaeological theories. In *Studies in Modern Archaeology*, vol. 3, 7–25.

GENEREUX, M. AND NICCOLUCCI, F. 2006. Extraction and mapping of CIDOC-CRM encodings from texts and other digital formats. In *The e-volution of Information Communication Technology in Cultural Heritage*, 56–61.

GRUBER, T. R. 1993. Toward principles for the design of ontologies used for knowledge sharing. *Inter. J. Hum.-Comput. Stud. 43*, 907–928.

GUARINO, N. 1998. Formal ontology and information systems. In *Proceedings of the 1st International Conference*. *Formal Ontology in Information Systems.* Trento, Italy. IOS Press, 3–15

GUARINO, N. AND WELTY, C. 2001. Identity and subsumption. LADSEB-CNR Internal Report 01/2001.

HALSALL, P. 1997. Modern history sourcebook: The Yalta conference, http://www.fordham.edu/halsall/mod/1945YALTA.html.

KIM, S., LEWIS, P., AND MARTINEZ, K. 2004. *SCULPTEUR D7.1, Semantic Network of Concepts and their Relationships*. http://www.sculpteurweb.org/html/events/D7.1_Public.zip.

LAGOZE, C., KRAFFT, D. B., PAYETTE, S., AND JESUROGAI, S. 2005. What Is a digital library anymore, anyway? *D-Lib Magazine 11,* 11.

LAKOFF, G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press, Chicago, IL.

LEVESQUE, H. J. 1984. Foundations of a functional approach to knowledge representation. *AI 23,* 2, 155–212.

LEVY, A. Y., RAJARAMAN, A., AND ORDILLE, J. 1996. Querying heterogeneous information sources using source descriptions. In *Proceedings of the 22nd International Conference on Very Large Databases*. Bombay, India, 251–262.

LU, J. J., NERODE, A., AND SUBRAHMANIAN, V. S. 1996. Hybrid knowledge bases. *IEEE Trans. Knowledge and Data Engineering, Volume 8, Issue 5*, 773–785, ISSN:1041-434.

MAGKANARAKI, A., ALEXAKI, S., CHRISTOPHIDES, V., AND PLEXOUSAKIS, D. 2002. Benchmarking RDF schemata for the Semantic Web. In *Proceedings of the 1st International Semantic Web Conference (ISWC'02)*. Sardinia, Italy, Vol. 2342/2002, Springer, Berlin, Germany.

MASOLO, C., BORGO, S., GANGEMI, A., GUARINO, N., AND OLTRAMARI, A. 2001. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf.

NUSSBAUMER, P. AND HASLHOFER, B. 2007. CIDOC CRM in action—Experiences and challenges. In *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Springer, Berlin, Germany, 532–533.

PATEL, M., KOCH, T., DOERR, M., TSINARAKI, C., GIOLDASIS, N., GOLUB, K., AND TUDHOPE, D. 2005. Semantic Interoperability in Digital Library Systems, *DELOS Network of Excellence on Digital Libraries*.

PODNAR, I., LUU, T., RAJMAN, M., KLEMM, F., ABERER, K. 2006. A peer-to-peer architecture for information retrieval across digital library collections. In *Proceedings of 10th European Conference (ECDL'06)*. Alicante, Spain, Springer, Berlin, Germany, 14–25.

RDF SEMANTICS. 2004. W3C Recommendation 2004, version http://www.w3.org/TR/2004/REC-rdf-mt-20040210/. P. Hayes, Ed. http://www.w3.org/TR/rdf-mt/.

REED, P.-A. 1995. CIDOC relational data model, A guide. http://www.willpowerinfo.myby.co.uk/cidoc/model/relational.model/datamodel.pdf.

ROUX, V. AND BLASCO P. 2004. *Logicisme et format SCD: d'une épistémologie pratique à de nouvelles pratiques édito-riales Hermès.* CNRS-éditions.

RODRIGUEZ, M. A., STEINBOCK, D. J., WATKINS, J. H., GERSHENSON, C., BOLLEN, J., GREY, V., AND DEGRAF, B. 2007. Smartocracy: Social networks for collective decision making. In *Proceedings of IEEE Hawaii International Conference on Systems Science (HICSS'07)*, 90.

RYAN, W. AND PITMAN, W. 1998. *Noah's Flood: The New Scientific Discoveries About the Event That Changed History.* Sinon & Schuster.

SCHORLEMMER, M. AND KALFOGLOU, Y. 2005. Progressive ontology alignment for meaning coordination: an information-theoretic foundation. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'05)*. Utrecht, Holland.

SINCLAIR, P., ADDIS, M., CHOI, F., DOERR, M., LEWIS, P., AND MARTINEZ, K. 2006. The use of CRM Core in multimedia annotation. In *Proceedings of the 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM'06)*. Edinburgh, Scotland.

SMITH, B. 2003. Ontology. In *The Blackwell Guide to the Philosophy of Computing and Information*, L. Floridi, Ed. Blackwell, Oxford, UK, 155–166.

SOWA, J. F. 1992. Semantic networks. In *Encyclopedia of Artificial Intelligence, 2nd Ed.* S. C. Shapiro, Ed. John Wiley, NewYork, NY, 1493–1511.

STOREY, V. C. 2005. Comparing relationships in conceptual modelling: Mapping to semantic classifications. *IEEE Trans. Knowl. Data Engin. 17*, 11.

STREITZ, N. A., HAAKE, J. M., HANNEMANN, J., LEMKE, A. C., SCHULER, W., SCHUTT, H. A., AND THURING, M. 1992. SEPIA: A cooperative hypermedia authoring environment. In *Proceedings of the European Conference on Hypertext and Hypermedia (ECHT'92)*, D. Lucarella, J. Nanard, M. Nanard, and P. Paolini, Eds., ACM Press, 11–22.

TOULMIN, S. 1958. *The Uses of Argument.* Cambridge University Press, Cambridge, UK.

UMLS® KNOWLEDGE SOURCES. 2000. February Release 2006AA Documentation. US, National Library of medicine, National Institutes of Health. http://www.nlm.nih.gov/research/umls/archive/2006AA/umlsdoc.html.

VINCENT, K. P. 2005. Text mining methods for event recognition in stories. Knowledge Media Institute, The Open University, Milton Keynes, UK, Tech. Rep. kmi-05-02, http://kmi.open.ac.uk/publications/pdf/kmi-05-2.pdf.

WEIBEL, S., IANNELLA, R., AND CATHRO, W. 1997. The 4th Dublin Core metadata workshop report. D-Lib Magazine.

WIEDERHOLD, G. 1992. Mediators in the architecture of future information systems. *IEEE Computer*.

WIGGINS, D. 2001. *Identity and Substance Renewed.* Cambridge University Press, Cambridge UK.

WITTGENSTEIN, L. 1984. *Tractatus Logico-Philosophicus, Tagebücher 1915–1916.* Philosophische Untersuchungen. Suhrkamp, Frankfurt, Germany.

WORDNET. 2006. WordNet a lexical database for the English language. *http://wordnet.princeton.edu/*.

Received December 2006; revised May 2007; accepted November 2007