Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM

George Bruseker, Nicola Carboni, and Anaïs Guillem

Introduction

The problem of data heterogeneity in the cultural heritage sector and its effect on restricting the ability to consolidate, compare, and demonstrate the findings of researchers is well known and forms a field, which has received significant attention in the past decades. While the attraction of digital and digitization projects retains its allure as a fundable and useful epistemic and institutional goal, both the short-term accessibility of the data produced as well as the long-term preservability of such information, remain a problematic question mark underlying such activities. Warnings of a digital dark age by authorities such as Vint Cerf abound, where the failure to resolve the issues of understanding and integrating data structures in a timely manner could mean that whole swaths of data produced under technological and data regimes that were not properly recorded and understood will fade into disuse or, worse, unusability (Ghosh 2015).

Ultimately real solutions will depend on the sustained commitment by specialists and especially by memory institutions to adopt and implement policies and procedures that take up standards and align data structures at some level to widely accepted schemas. And yet, before such a goal can become a reality on the ground, the theoretical underpinnings of mass data integration must not only have been solidly established in themselves but, moreover, have taken on such a theoretical form so as to be accessible not only to computer science specialists but, equally, to domain specialists in Cultural Heritage (CH) and its many constituent disciplines. Only in this way, when those who generate the knowledge at the ground level can participate in building and adding to the digital forms and standards that encode them, will the need for long-term compatibility, maintenance, and commensurability of digitally produced knowledge be met. It is not unfair to say that at this juncture in the study of knowledge integration, this latter problematic forms a foundational issue for the onward development of the field.

This chapter is elaborated within the scope of this problematic. Specifically, we propose to review the approach undertaken in the building

G. Bruseker (✉)
Centre for Cultural Informatics, Institute of Computer Science-FORTH, Heraklion, Crete, Greece
e-mail: bruseker@ics.forth.gr

N. Carboni
UMR 3495 MAP CNRS/MCC, Marseille, France
e-mail: nicola.carboni@map.cnrs.fr

A. Guillem
School of Social Sciences, Humanities and Arts, University of California Merced, Merced, CA, USA
e-mail: aguillem@ucmerced.edu

M.L. Vincent et al. (eds.), *Heritage and Archaeology in the DigitalAge*, Quantitative Methods in the Humanities and Social Sciences, DOI 10.1007/978-3-319-65370-9_6

of CIDOC CRM to manage the integration problem and to outline the directions of research that have been followed in the past years in extending the model to handle knowledge provenance across various disciplines and typical documentation and reasoning activities. To introduce this topic, we will begin by an outline of the data challenge specific to CH and the main approaches towards data integration that can be undertaken to face this challenge. We will then introduce and distinguish knowledge engineering and formal ontology from other information modeling techniques as the necessary approach for tackling the broader domain integration problem. Proceeding from this general background, we will introduce the basic principles of CIDOC CRM, the ISO standard for our domain of interest, and how it addresses some of the main problems and questions raised in knowledge engineering for this domain. With this basis, we will turn to look at the work that has been done both theoretically and in practice over the past five years in developing and implementing CRM as a practical data integration strategy in CH, looking at specific extensions for different types of research and successful implementation projects. Lastly, we will look at the present potentials and challenges for using CIDOC CRM for solving the integration puzzle. The intended audience of this chapter are specialists from all backgrounds within the broader domain of CH with an interest in data integration and CIDOC CRM, in order to give a short account of the meaning and use of this methodology as well as a review of how it is being developed and expanded by different communities presently in order to extend its application.

Cultural Heritage as "Domain," the Nature of Its Data, the Potential for Harmonization

Data coming from the cultural heritage community comes in many shapes and sizes. Born from different disciplines, techniques, traditions, positions, and technologies, the data generated by the many different specializations that fall under this rubric come in an impressive array of forms. Considered together the collective output of this community forms a latent pool of information with the capacity, when integrated, to support potential knowledge generation relative to any period, geographic location, and aspect of human activity in the past even when, characteristically, based on sparse data sets. Despite this potential, the material lack of uniformity in data and in methods means that data integration is generally difficult and is usually brought about manually, meaning that the full of capacities of the possible integrations of different data sets are very hard and expensive to realize and/or repeat.

It could be a natural problem to pose from the beginning: if the data of this community indeed presents itself in such a state of heterogeneity, does it not beg the question if there is truly an identity and unity to cultural heritage data in the first place? It could be argued that *Cultural Heritage,* as a term, offers a fairly useful means to describe the fuzzy and approximate togetherness of a wide array of disciplines and traditions that concern themselves with the human past. The term has a functionality at the least for forming an ideological and perhaps even practical funding umbrella for a wide array of disciplines with analogic interest in a field. Yet, perhaps, when it comes to performing an analytic of the elements of this field, we would discover that, in fact, it is composed of a number of quite separate disciplines such as analytic sciences, humanities, and archaeology. which are essentially incommensurate amongst each other and only at best commensurable at individual levels but certainly not across a wide horizontal plane.

We would not take this position but, rather, argue that not only despite but, indeed, owing to its generality, cultural heritage as a term helps point us to a genuine identity and unity of purpose across the many disciplines it covers (Doerr 2009). The linked nature of these many different disciplines, in turn, points towards the unfulfilled necessity of better data integration. The tie that binds, as it were, the aforementioned disciplines is the common commitment to the scientific analysis and presentation of the human past based on empirical evidence. While

at a high level of generality, this commitment nevertheless binds the related disciplines both to an external standard of rigor and co-implicates their studies with one another. Such a position is coherent with the intentions of international bodies like UNESCO that have long set forth international conventions on the study, promotion, and protection of CH, which adopt such a high level view of the interactive unity of cultural heritage disciplines (UNESCO 1972, 2005). The disciplines of archaeology, conservation, museology, library studies, archives, and so on, should not operate in a vacuum from each other's research results. The outcomes of the one, assuming they all refer to the same objective domain of discourse, have implications on the other which require assimilation and integration into the overall view of affairs, potentially initiating knowledge revisions or new conclusions based on new information revealed by techniques, methods or studies not available in one's own home disciplines.

The conclusion this drives us towards, with regards to the question of identity and unity of the domain of cultural heritage is that it is one with regards to its object, the empirically investigable human past, but several with regards to its approaches (Doerr 2009). Such plurality within CH is not an obstacle to be overcome but a constitutional condition of the "domain." This limiting condition is, in fact, a driving force behind cultural heritage research, in that it does not limit the approaches that could be valid with respect to its object, but remains constitutively open to new sources of data by which to enlighten areas of understanding with regards to our past. The challenge, then, to computer scientists and domain specialists working in tandem is to conceive of commonly understandable and applicable methods whereby data resultant from the multiple sources of cultural heritage knowledge can be expressed by a means that makes them mutually intelligible, at some level, through automated processes. This being said, it is important to stress that there are necessary forms of heterogeneity at the disciplinary and methodological level which neither can nor should be overcome. Rather, it is these very forms which must be modeled and shown in their interrelation. Such issues of difference of approach and methodology are either nonproblems, because working in parallel but non implicated directions or, if there is genuine conflict, are to be sorted out by the data and what it shows, not by any data harmonization process.

Where harmony can be sought at the cross-disciplinary level is through understanding what practices and processes can be inductively abstracted that form common means of approach and conceptualization across the disciplines. Rather than seeing the internal boundaries of the domain as being formed by the traditional disciplinary divisions, it might be the case that we can isolate and abstract new functional unities within the general domain of CH. Here again, though, the idea would be to seek for unique process and structure types of the overall field which allow for a common understanding. The aim of such an exercise would not be to propose some essentialist model of what CH is, but rather to extract how cultural heritage professionals actually work in such a way that we can build common data structures for exchange of information just at the points where we are able to agree.

Sources of Data Heterogeneity: Accidental and Necessary

There are, nevertheless, a number of factors on the practical level that contribute to data heterogeneity that do admit of the possibility of resolution through appropriate strategies for consolidation and harmonization. Such factors seem to lie within the practical aim of a set of strategies for data integration and include: local and disciplinary tradition, technological limits, lacks with regards to standards and funds as well as inappropriate/reductive aggregation strategies. Before proceeding we should consider looking at these conditions and understand the nature of the barriers that can be overcome to achieve data integration.

A large amount of data heterogeneity and sparsity is the result of what globally can be understood as different data recording and

retrieval traditions, which lead both to different forms for the data and different quantities thereof. There is an aspect of "the way it is done" in data systems that has to do with the inertia of institutions and individuals over time. Data structures are adapted to individuals and circumstances, rather than to the objects they model. In other cases, the more general legal environment of data generation may enforce the collection of data in a certain form. In yet other cases, the confines of an academic discipline, the techniques and practices amassed according to a tradition of thought, might dictate and make most obvious certain data formats or expressions over others. Such heterogeneity is sometimes accidental and sometimes necessary. Where accidental, it admits of correction through the adoption of some standard. Data anomalies can be brought into line with standard practices. Where necessary, such as the continuation of an intellectual tradition, it can be viewed as a positive constraint, since while nonreducible, the continuity of the tradition implies a conceptualization that can be generalized as well as a body of evidence from which to understand this conceptualization.

Another prolific cause of data heterogeneity are the technological barriers that arise both from legacy data systems and the proliferation of new data production technologies and techniques.

On the one hand, the cost of investment not only in software and hardware but also in training to run systems at the infrastructural level and to implement them at the protocol level means that certain data structures, especially in cultural heritage, have a long life cycle with no immediate practical likelihood of being taken out of use. Such structures can struggle to keep up-to-date with changing techniques and methodologies of recording, leading to inconsistent documentation. When this occurs, data cleaning and sorting by controlled list and vocabularies and the meticulous documentation of the appropriation of the data structure to new expressions are a necessary practical prerequisite to larger scale data integration. That being said, such systems, insofar as they are consistently used and this use documented, offer a perfect source of data for large integration. Explicit policies make the meaning of the data accessible and therefore translatable to a more general form.

On the other hand, it is just as much the growing number of tools for generating different types of data for cultural heritage purposes, especially with regards to new analytic techniques which raise ever anew the problem of how to align such data. New techniques for describing our objects of interest tend to reveal new features of these objects which entail in turn new data structure needs for which there are no necessary existing standards. The introduction of new technologies and techniques offers us novel views but constant unforeseen challenges to understand the data produced and to align it with existing data sets.

Heterogeneity in our information set exists not only, however, thanks to such positive limits but also as the result of a series of lacks faced in the general cultural heritage field. While there may be a will toward compatible research results, there is often a lack of sufficient resources, in terms of available standards or the understanding thereof, to support the creation of harmonized/able data structures and/or data. The plurality question raised above means that, because of its variety, even when one has the will to apply a standard to some data set, the appropriate standard may not yet exist. The development of such standards, however, demands a commitment that goes beyond the purview of individual projects' and even individual institutions' efforts. The development of a standard requires broad consultation that takes place over a significant span of time and is open to revision (ISO 2016). The investment in time and money is high, and the cementing of the long-term buy-in of a sufficiently broad series of partners very challenging. Such commitments in real world terms make demands on scarce resources. Therefore, even where such standards exist and are applicable, application of them can mean transformation of data structures and transformation of the data itself, all calling again on limited funding resources within a limited funding pot.

Finally, we can reference inadvertent generators of heterogeneity, which, paradoxically, can be the result precisely of efforts at harmonization. Whatever the path to data integration and

harmonization may be, there is no simple and direct one-size-fits-all solution. Some efforts to address this problem have focused on monolithic integrative approaches, attempting to build systems and standards for everyone. Given the necessary data diversity of which we have spoken above, such efforts have counter-intuitively resulted in generating even more data heterogeneity by forcing data into formats that they do not fit. The end result is the loss of knowledge and context through the discarding of the original semantically meaningful structure in which the data was generated (Oldman et al. 2014).

Classical Knowledge Organization: Traditional Solutions to Data Heterogeneity for Data Aggregation

Before proceeding to examine the functionalities of formal ontology in approaching this issue, it is worth taking a step back and understanding our proposed strategy for information integration in its historical context and how this context forms part of the data integration puzzle today. The subject of addressing information heterogeneity through knowledge organization and the derivative challenge to create significantly general intellectual structures to manage this complexity is not a new one. Most of the information system strategies that we deploy today still rely on traditional notions of categorization and information management. Formal ontology is a fundamentally new proposition for how to approach this problem that is often confused with its cousins and ancestors in the field. If we are to understand what formal ontology is and what it can do, then, it is useful to begin by examining what it is not and what it does not attempt to do. We can try to do this by looking at the traditional means of conceptualizing the organization of knowledge and how it is applied in different information management strategies.

The foundations for many of our most commonly used information management strategies have roots that lead back to the foundations of logic in ancient philosophy and the formalization and comprehension of the relation holding between categorical statements, together with the analysis of the manner of defining categories and the properties shared by their instances.

The mental image of the "tree of knowledge" which such strategies implied, led to their common representation in tree-like structures that are still familiar as an organizational structure for knowledge and action in many everyday contexts. The belief that our knowledge of the world can be decomposed into a complete structured tree of branching information (Rosch and Lloyd 1978), is reflected in the tendency to use categorical assertions and to attempt to define knowledge according to a finite set of categories. Famous historical examples of such work include the Porphyrian tree where the categorical system of Aristotle was mapped in a tree-like structure and the titanic work of Diderot and d'Alembert for mapping every subject of their Encyclopedia into a genealogical structure (Weingart 2013). In the latter case, however, we already find the authors beginning to question the viability of a unique unified order of knowledge to which systems of the past subscribed (Le Rond d'Alembert et al. 1995).

Variants of this traditional conception of knowledge organization as consisting of categories where the category delimits a clean set of entities with clear instances in an objective world have a central role to play in Western intellectual history up until the end of the nineteenth century when the foundations of this perspective began to be questioned in the works of thinkers like Peirce and Wittgenstein. Peirce, for example, began to lay out a new perspective that would take into account the relationships of an entity in the understanding of its identity. He introduces a meta-level distinction between the being (Firstness) and the being in relation to something else (Secondness), as well as the mediation that bring multiple entities into a relationship (Thirdness) (Sowa 2000).

Foundational studies like those of Peirce, opened up the complexity of the concept of the identity of a category, as well as the relationships between meaning and sign, which is at the base of the organization of a corpus of information.

This work joined by studies by thinkers such as Husserl, Whitehead, Wittgenstein, Rosch *inter alia* have slowly opened the concept of category to a finer analysis. The problems of potential ambiguities of concepts both in definition and in terms of their set membership and a skepticism towards the possibility of providing complete correspondent information structures to the objective world have becomes central issues of debate and research.

Traditional Knowledge Organization Systems

The formal ontology solution which we will look at for wide-scale data integration is not to be understood in isolation from the traditional information management techniques elaborated over the last century, but rather should be seen as a continuation of this effort, attempting to implement some of the insights arising from over a century of research into categories and knowledge organization and, of course, the massive changes that have occurred in computer science by which we are able to implement such techniques. Most of the data that a formal ontology would seek to integrate would have been elaborated within the context of some classical knowledge organization system. Therefore, in order to better comprehend the methods, achievements and limitations of the application of the classical view of information organization in addressing the data heterogeneity problem, as well as the manner in which such systems can achieve data interoperability, we introduce here a small outline of the main tools used in the information institutions, giving later an extensive account of the issues deriving from their use as well as a possible solution. Specifically, we will look at protocols, controlled vocabularies, taxonomies, thesauri, metadata, and data schemas in order to present the role such tools play in data integration, the methods they employ, their uses, and limits. Table 1 is a summary of these methods, and it lists well-known examples used by memory institutions.

Protocols

Protocols are external to data systems and act as normative devices to indicate how to organize actors with regards to processes and procedures in order to capture the right information at the right time with regards to objects, events, etc. Protocols have the distinct aspect of being prescriptive. They are formalizations made by a body of specialists that articulate a researched and founded ideal set of events that will occur in order to keep track of essential information with regards to some domain of interest. Protocols generally avoid any specific commitment to a particular language or structure, for their use is not in identifying means of expression but rather in identifying what is to be expressed/should be documented.

A widely known protocol in the museum community is SPECTRUM. It provides a model for setting up collection management procedures that provides normative rules for orienting actors in the world and the actions they should take towards documentation in the practice of collections management. It provides models for how to organize 21 separate procedures for dealing with collections and, with regards to information management, indicates the information that must be collected at particular moments in order to support the long-term understanding and access to the objects in care. Implementation of SPECTRUM is a legal requirement for museum accreditation in the UK. While the possibility of implementing SPECTRUM implies a very specific context, it nevertheless provides a clarified local description of an understanding of a set of activities which stands behind a series of documentation events. It thus stands as an excellent example of the contribution a protocol can have as part of an overall solution to the problem of data heterogeneity by contributing a greater regularity to data and providing part of the solution to sparsity of data by identifying the likely important events (objects, actors, etc.) as necessary variables to document and control.

Table 1 An illustration of well-known examples of different types of knowledge organization systems used in memory institutions

	Library	Museum	Archives
Protocol	ISBD[a]	Spectrum[b]	ISAD[c]
Controlled vocabulary	Library of Congress Name Authority File[d], Authority List for Journal Titles[e]	The Revised Nomenclature for Museum Cataloging, Gazetteer of British Place Names[f],	A Glossary of Archival and Records Terminology[g]
Taxonomy	Dewey Decimal Classification[h]	Traditional Biological Taxonomy	
Thesaurus	LCSH[i]	AAT[j]	UKAT[k]
Metadata and Data Schema	Dublin Core[l], UniMARC[m], METS[n]	Core Data Index to Historic Buildings and Monuments of Architectural Heritage[o], MIDAS Heritage[p], CDWA[q]	EAD[r]

[a]http://www.ifla.org/publications/international-standard-bibliographic-description
[b]http://www.collectionstrust.org.uk/spectrum
[c]http://www.icacds.org.uk/eng/ISAD(G).pdf
[d]http://id.loc.gov/authorities/names.html
[e]http://www-pub.iaea.org/books/IAEABooks/7531/INIS-Authority-List-for-Journal-Titles
[f]http://www.gazetteer.org.uk/
[g]http://www2.archivists.org/glossary
[h]https://www.oclc.org/dewey
[i]https://www.loc.gov/aba/cataloging/subject/
[j]https://www.getty.edu/research/tools/vocabularies/aat/
[k]http://www.ukat.org.uk/
[l]http://dublincore.org/documents/dces/
[m]http://www.ifla.org/publications/unimarc-formats-and-related-documentation
[n]https://www.loc.gov/standards/mets/
[o]http://archives.icom.museum/objectid/heritage/intro3.html
[p]https://historicengland.org.uk/images-books/publications/midas-heritage/
[q]https://www.getty.edu/research/publications/electronic_publications/cdwa/
[r]https://www.loc.gov/ead/

Controlled Vocabulary

A Controlled vocabulary is an "organized arrangement of words and phrases used to index content" (Baca et al. 2006). In its basic version, it is a simple flat terminological list which provides a set of controlled terms that can be used to specify something about an object, its subject for example. Controlled vocabularies can also be more structured, including equivalent terms (context-based synset) and, in case of two or more variants, a preferred term is chosen (e.g.: *USE* Salinity for saltiness) (National Information Standards Organization 2005). Authoritative controls over the vocabulary distinguish it from other forms of free listing of terms, like folksonomy. Vocabulary control is used to standardize naming and improve indexing, browsing, uniformity, and retrieval of the data described (Vállez et al. 2015).

The classical case of vocabulary control happens in libraries, where the bibliographic records are organized based on a process called authority control. In this instance, the form of the name of the authors is closely controlled in order to relate their work to a standardized version of their name. Changes in the form of an actor's name can happen for many reasons, commonly including artistic ends (Prince Rogers Nelson or Prince or Joey Coco or The Artist Formerly Known As Prince) and personal reasons (maiden or marriage name). In any every case, the use of a controlled vocabulary maintains a consistent means of referring to the same entity with the same

name within the bibliographic catalogue, while also accounting for variants which should refer back to the standardized name form.

Taxonomy

A taxonomy is a "*cognitive model of a particular kind [..] built into languages throughout the world*" (Lakoff 1987). It is built up by classical nonoverlapping categories defined by their features. Structurally, a taxonomy relies on a controlled vocabulary and on the use of subsumption relationships for ordering a diverse set of entities. It is usually used to relate an individual to a species, therefore creating a generic/individual type of relationship, or to express the membership of a subset within a superset as in a generic/generic relationship. In the former case, we express a type of predication, for example when we assert that Socrates is a man, while in the generic/generic case we assert a subtype relationship, for example when we declare that a penguin is a bird (Brachman 1983). They enable standardized classification terms.

Taxonomies are used in very controlled information environments. A classical case of the application of taxonomy in the CH domain is related to the natural sciences community. Curators and researchers build and maintain taxonomies of species and particularly track the creation and variant naming of taxa. This evolving structure is related back to specimen evidence and allows curators and researchers to find and re-examine evidence and test conclusions. Taxonomic relationships are also used also for constructing certain classification schemes intended to be used as large taxonomies which rely on a notation language to provide information about their status. An example is the Decimal Dewey Classification, which aims to catalogue the subject matter of any book into one of its categories, assuming that would fit the aboutness of the book in question. Taxonomies resemble ontologies in their strong ontological commitment. They are developed generally on a correspondence model between information structure and world, where the information produced aims to mirror objective reality. Two main differences, which we will explore below, are on the nature of the ontological commitment and the exploration of relations in the world over classification. Being highly structured and regular data, taxonomies are perfect structures for adaptation into information aggregation scenarios.

Thesauri

A thesaurus is a type of controlled vocabulary that relates its terms using taxonomic and semantic relationships, and it is defined as "a controlled vocabulary arranged in a known order and structured so that equivalence, homographic, hierarchical, and associative relationships among terms are displayed clearly and identified by standardized relationship indicators that are employed reciprocally" (National Information Standards Organization 2005). At a functional level it is used for enhancing the retrieval of information from a system (Moreira et al. 2004). Thesauri, too, begin to move towards an information structure that would resemble an ontology. Both subsumption relations (BT/NT[1]) and horizontal relations (RT/UF[2]) can be expressed in thesauri, but they remain an exploration of terminology rather than clearly formalized conceptual entities, moreover there is not a strong focus on the definition of the functions that relate terms, underlining the lack of ontological commitment which would make this type of information structure subject to a number of pitfalls described below.

Thesauri can be developed to deal with the naming of a broader or narrower range of subjects and applied to control data consistency and retrieval. Examples in the domain of cultural heritage might be the targeted thesauri developed by the British Museum organizing terms for describing object names or material. Examples of broader scale initiatives would be the Getty thesauri: Art and Architecture Thesaurus, the Getty Thesaurus of Geographic Names, the Cultural

[1] Broader Term/Narrower Term

[2] Related Term/Use For

Objects Name Authority and the Union List of Artist Names.[3] These thesauri, having a wider range, apply techniques of faceting. A recent European wide example is the work of DARIAH in developing local and backbone thesauri, which attempt to provide both very specialization oriented thesauri linking to a broader back bone of terms. Within the scope of developing common terms for reference to subclasses of objects to particular specialists, and providing homogeneously generated data for further analysis, thesauri execute an important role in the production of standardized data for reuse within aggregation structures.

Metadata and Data Schemas

With the advent of the relational database, and the ability to rapidly create bespoke data structures for data organization, standardized metadata and data schemas have been designed as a means to suggest appropriate models for capturing information in particular domains of interest. The schemas are the result of an interpretation of a domain resulting in an intentional model which delimits the finite set of descriptions that can be assigned within a specific setting (Falkenberg et al. 1998). A schema therefore formalizes, often implicitly, a view of a domain which can have different levels of complexity in relation to the granularity of the initial investigation and its function in the actual world. The complexity in re-applying schemas, their case specific nature, and the usually underanalyzed relation between the data structure and the objective world it describes, strongly limits the possibility of their use in large-scale data integration. The complete replication of such complex schemas from one environment to another is rarely a viable solution even if the purpose of two information systems is the same, given the variable needs and traditions of local contexts. For this reason, another solution suggested in order to capture at least a core of the generic conceptualization of a field and thereby enhance the interoperability between different systems is the metadata schema.

Metadata schema are intended to increase, "the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality" (National Information Standards Organization 2004). A metadata schema consists in a flat formalized set of elements, usually in the form of structured textual information, which standardizes the description of the core elements used to documenting a specific type of information resource (text, video, etc.) or one of its aspect (administration, preservation). Sometimes there are cases where one aspect of a metadata set is considered so important that it is given a unique name, like in the case of paradata. It is important to underline, however, that in these cases, we continue to talk about metadata, under a new name. In the case of paradata the functional aspect of the metadata for tracking provenance of data is emphasized. Using a standard metadata schema allows for the partial preservation of an aspect of the richness of different data schema between diverse databases, thus enabling federated query functionality over this reduced set.

It is important to underline that both the data and the metadata schemas do not have a formal commitment to the explicit representation of their scope.

Well known examples of metadata schemas include Dublin Core,[4] MPEG7[5] and METS.[6] These schemas serve a functional role within specific contexts for the purpose of providing a structure composed of multiple descriptors that allow the documentation and the retrieval of an item. Applied at this level, in conjunction with other Knowledge Information Systems like classification schemas, taxonomies or thesauri, metadata schemas reduce the overall level of heterogeneity within the information space by providing access points towards a small set of standardized information of an object, and allow-

[3] http://www.getty.edu/research/tools/vocabularies/

[4] http://dublincore.org/specifications/

[5] http://mpeg.chiariglione.org/standards/mpeg-7

[6] http://www.loc.gov/standards/mets/

ing an initial analysis of the information coming from systems deploying schema using different conceptualizations.

Limits of Traditional Knowledge Organization Strategies for Data Aggregation

What can be said to be common amongst the above approaches to the resolution to the problem of data heterogeneity is the creation of a pre-established frame that specifies the way and manner of the documentation of the object and aims to provide a unique, correct description of its object by creating language and domain specific constraints, which limit the semantic expressivity of the information we can document in regards to the object. The user of such a system is forced to pick among the available options in order to make statements about their domain of interest. That is to say, within the context of an established field with an agreed viewpoint, such tools are invaluable in rendering data commensurable at a local community level and allowing easy entry of data according to a common world-view. The imposition of a standardized knowledge frame as a means to reduce the data integration problem by identifying distinct areas of investigation can in no way, however, represent a fundamental solution to the overall integration problem over a complex domain such as CH. The solutions cannot scale-up, and the extension of standards across noncompatible data risks confusing users and making data ultimately unusable.

The fundamental reasons why the above classical strategies cannot be used at a broad scale for data integration across heterogeneous data sets sits either with their inapplicability to the problematic or with their commitment to either an untenable exhaustive or minimalist approach to the description of the domain of discourse.

On the one hand, classical hierarchical classification systems such as classificatory schemas, taxonomies and thesauri are inappropriate to the task of large-scale data integration due to the constraints imposed by language itself and the intellectual architecture by which they are expressed. Such systems are stymied in the task of integration by basic linguistic problems, especially the issues of homonymy and polysemy. In the former the words are pronounced alike but they have different meanings, while in the latter they are systematically related. Examples of both are given by Lakoff (1987). He offers for homonym the example of the word "bank," which refers both to the institution and the edge of a river, while he shows the problem of polysemy by reference to the case of "warm," which stands for the temperature and also the type of clothing that allow you to keep such temperature. The inability to differentiate the meanings of the word causes the classical retrieval/description problem, in which producer and users cannot communicate or research the same content because no relation to the entity that the term is supposed to represent is established.

The effectiveness of such systems can be enhanced by the use of hierarchical structure, which would define the words within a particular category, or by the use of textual qualifiers that define its role within the system. The qualifiers could help resolve the issue, but only during the manual browsing of the information structure (Svenonius 2000). In reference to the hierarchical solution, it could help disambiguate some basic terms, but the problem would not be resolved with the vaguer ones. It would be quite challenging for example to force the term "beauty" within a specific category. Moreover, a hierarchical structure is always the product of a context, and therefore the choice of what is to be categorized, the recognition of a gestalt as well as the salience of the word used for constructing the information structure are always context-dependent and they always rely on some modeling-choice, which are usually not clearly stated.

Furthermore, the classical hierarchical categorization systems lack the means to distinguish different types of fundamental relations, often confusing subsumption relations with other kinds of relation, leading to ambiguity or simple incorrectness in description. This occurs particularly with the description of parts and wholes (Gerstl and Pribbenow 1996). The problem arises from treating classes as if they act in the same ways as sets,

therefore conceiving a subclass as a subset, which, *per se,* implies a subsumptive relationship. Guarino and Welty (2002a) illustrates this problem using the example of the relationships between an engine and a car, where the former is sometimes described as subclass of the latter, even if, even with a quick overview, we can easily recognize that they share different properties and their relationships should be described using mereology.

Aside from the problems of ambiguity inherent to the application of classical hierarchical categorization to the possibility of creating large-scale knowledge integrations, there is a general problem with the strategy of traditional categorization that forces it into a closure decision with regards to its scope that is fatal to the possibility of building adaptable integration mechanisms. Because classical categorization holds, at least implicitly, its systematization to be complete and to isomorphically hold in the sense of a one to one correspondence with the world, the attempt to use a classical schema for broad data integration encounters the problem of providing either a maximalist or minimalist solution to data expression.

This problem is most clearly seen in the offer of metadata and data schemes to solve wide domain data integration problems. Such schemas, committed to their correctness and completeness, are committed to fitting the relevant data and to give it unambiguous expression. Faced with the potentially infinite diversity of phenomena that it must cover, such a schema must, therefore, either choose a maximalist set of descriptors that aims to richly cover all possible options or a minimal set to which all data sets produced should conform. While such strategies can have an important role in gaining control over data in a local context by creating a program and a culture of data gathering that is consistent and integrated, the effort to expand it into an open world of discourse is bound for failure.

The data integration problem exists, in fact, at least in part because constitutionally we do not know what new methods, new observations, new data artifacts will be generated that will have a bearing in reference to a particular problem. If we set, a priori, that which can be said, it is as if to say we already know all that there is to be said. Such a position does not have the necessary epistemic flexibility in order to be able to respond to the wide diversity of actors and methods involved in a complex domain such as CH. Knowledge generation is always and necessarily incomplete, meaning that we cannot use a closing of a system in order to perform an integration. Objects of investigation will be taken up by different actors using different techniques, given different names and analyzed at different levels of granularity from different perspectives. If there were only one frame of reference then the job of integration would be simple, but the frames of understanding are in principle unlimited. A maximalist effort to list all possible positions on a domain, will therefore continuously have to undergo *ad hoc* extension in order to cover new approaches and perspectives, with the list of particular practices growing ever more unwieldy.

A minimalist effort, on the other hand, which we have referred to briefly above, involves one in an essentialist truncation. Such an essentialist position attempts to pick out the core data which is relevant to all data across a complex multi-actor domain. The problem is that such an essentialist function runs into a double headed problem. On the one hand, it may take some position on what is the semantically relevant subset of data, and, in so doing, takes a position on what the science with regards to this domain "is." Such a strong epistemic position does not accord with the actual open world of discourse. On the other hand, it may attempt to remain at a thin description of the discourse, giving only fields for identifying data at the broadest level of discourse, in which case it gains universality at the sacrifice of expressibility with regards to the specifics of present science (Oldman et al. 2014).

What we can conclude with regards to predefining a complete classical classificatory schema is that due to the polysemy of language and the seeming impossibility of formal correspondence between the world, or the state of affairs described, and the schema used, such schemas are not appropriate to the task of wide-scale data integration. They lack the flexibility to pick out all the potential objects they are meant to

describe without fundamental ambiguity. They are, furthermore, forced, when extended to a broad domain, into a maximalist or minimalist commitment on data representation, leading to an impossible situation of a complete specification of a constitutionally indefinite domain.

It is important to reiterate, however, that this does not represent a critique of these tools as such. All of the above tools have a strong role to play in gathering and ordering data at a local level with regards to specific problems and, in doing so, they create a body of well formulated data that can be interpreted into a broader integration structure. Such tools play a necessary and on-going role at the point of production of data to ensure that it is well structured and formulated for some local community.

When it comes to expressing such data into a broader community such as the wider CH domain, or even integrating with other specialists working in the same subdomain but deploying other, valid categorical systems, it is necessary to seek a different solution. Such a solution would require a thorough exploration of the conceptualizations expressed in the broad domain of discourse, divorced from linguistic features and accidental structures delimited to some set of objects or tasks, in to understand the general conceptualizations common across these structures at the categorical level. It would require the discarding of the notion of a final classificatory system and, rather the attempt to deploy the new more flexible understanding of categories developed in the past years. Finally, this work on reimagining categorization would have to be expressed in a formal language separated from particular linguistic expression or closed domain expressions. It is to the question of how to achieve this that we turn in the next section.

Knowledge Representation and Knowledge Engineering

The tradition of formalizing propositions in a natural language independent formalism, with the aim of providing a neutral means of presenting conceptualizations and allowing reasoning and description in a certain domain is the typical work of logic and mathematics, but during the second half of the twentieth century, and starting from the 70's (Hoekstra 2009) computer science, and in particular the subfield of AI, begins to adopt these tools in order to try to develop systems able to exploit the definition of formal propositions with the aim of building rich knowledge bases.

The field has come to be known as knowledge representation, which has been defined as "the application of logic and ontology to the task of constructing computable models for some domain" (Sowa 2000). The definition of the ontology, and therefore the specifications of our model is the job of the knowledge engineer (Brachman and Levesque 2004). Before looking particularly at how this movement has been expressed in cultural heritage, it will be useful to give a basic outline of the strategy of knowledge engineering. The field deals with the problem of information integration by bringing a new methodology and conceptual approach to the problem of heterogeneity described above. This approach particularly aims to avoid the problems identified in the classical knowledge organization techniques. That is to say, it attempts to avoid the pitfalls of language ambiguity and to the commitment to a single model of the domain, which forces the maximalist and minimalist approaches described above. The aim is to readdress the problem in a more robust and flexible way, capturing both the complexity of the data produced in large heterogeneous fields while building the conceptual building blocks for creating appropriately generic and reusable data structures and patterns.

The method proposed for building such structures is the generation of a formalization of a conceptual domain. Concretely this means attempting to engage with and describe the fundamental principles, objects and relations appealed to and invoked by a group of users within a wide domain context (Smith 2006). It involves an interdisciplinary dialogue between domain specialists, computer scientists and knowledge engineers (Sure et al. 2009). This forms a fundamental task of understanding and

conceptual design wherein the scope of a domain is investigated as to its meaning and with regards to its typical contents and arguments. The method aims to described the so-called ontological commitment of the user community. As Guarino (1998) puts it, the product of this effort is a formal ontology which is,

> "*logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models*."

By its very manner of construction, a formal ontology attempts to avoid the traps for data integration associated to classical categorization efforts. It does not attempt to provide a universal, one-to-one objective correspondence of its categories, nor present itself as a data surrogate for the world described. The purpose of a formal ontology is functional (Zúñiga 2001). It specifically focuses on finding and describing the particular view of the community of users it aims to help structure data for, and to model this explicitly. It does not present a neutral view, but by making its commitments explicit, it neutralizes the ambiguity and overreach problems reviewed above. The goal is not a perfect representation of knowledge, but one adequate to the aims of the domain users and consistent with reality. It is important to highlight that this kind of approach would differentiate between the ontology, the conceptualization that it is committed to, the language used for its implementation and the objective world that it refers to.

The method deliberately eschews an interest in any particular implementation either with regards to individual projects and even with regards to particular types of encoding (Davis et al. 1993). The work in creating a formalization is an entirely conceptual work undertaken by knowledge engineers in close collaboration with the user community. The examined data comprises the heterogeneous data structures alongside the domain knowledge of the specialists in how this data is formulated and understood, as well as an elaboration of the kinds of questions that domain specialists need to make of their data (Sure et al. 2009). What the process drives towards is a description of the essential points of reference and the relations drawn between these points by the domain users. The effort is to understand the concepts not "in general" but with regards to their functionality within the defined domain of use in question (Davis et al. 1993; Bergamaschi et al. 1998). The problems presented by the maximalist and minimalist approaches to the integration problem are avoided by searching not for a set of terms, fields or data structures adequate to the domain, but rather by searching to isolate the general patterns of argumentation and reference within the domain and to describe these concepts and relations in such a manner that well-formed existing data structures, without any modification to their structure, can find an adequate representation in the generalization produced in the formalism. The formalism therefore becomes an exchange point between data structures which continue to exist in their plurality but which have a possible neutral expression point in order to allow cross structure searches and data exchanges.

The technical means that enable the work of knowledge engineers to develop tractable formalizations from such a process are the expression of domain knowledge in terms of well-defined classes and properties ordered in an isA hierarchy, that will be used as the backbone of a formal ontology. A formal ontology has as its substance a declaration of its scope and a series of classes and relations that result from the generalization work done in the dialogue/research described above.

The scope of the formal ontology describes the domain which is to be taken into account for the construction of the ontological model. It must be explicitly declared in order to limit the intended domain of application of the overall formalization.

A class is a "a category of items that share one or more common traits serving as criteria to identify the items belonging to the class" (Le Boeuf et al. 2016), and serves as a documentation

unit that is described by a scope note, which textually indicates the intension of that class. The intension of a class is a description of the essence of that category such that a human being can read the description and identify instances of it. The clarity of such descriptions is paramount for the effectiveness of an ontology and presently research continues on the best means to ensure clarity of expression (Guarino and Welty 2000a, b).

Properties (also known as relations) are generalizations of kinds of relation that can exist between classes. Their formalization results from research into how users actually do reason over and relate objects in the domain. The discovery of properties is crucial, and even prior in importance to the declaration of classes, as they form the basis for the latter's declaration. It is moreover important for each of them to be given an intentional definition to ensure their proper application. Properties are additionally restricted according to a domain and range of classes (Doerr et al. 2007). That is to say each relation's domain and range scope, that of which it can be said sensibly, is explicitly specified in the formalization, thereby delimiting the types of acceptable propositions that can be made through data encoded in this structure. The specification of these relations is the basis of the possibility of reasoning over the data at later stages.

The central tool for gaining expressive power, however, within the ontology is the application of an IsA hierarchy over the classes and relations. Formal ontologies make use of a function of inheritance provided by the IsA relation in order to be able to order classes from more general to more specific, attributing and restricting along the way the relations that can be used to describe entities at a more general level and those which, when added, create a new functional unity for the class and determine a new level in the IsA hierarchy.

This method of constructing the classes, which can be encoded and reasoned upon, deliver a number of advantages in providing integrative data structures. It allows describing relations that pertain to a broad number of classes at a very generic level just once, and to use these generic relations to model specializing subclasses and relations of any depth. While the ontology will never declare all possible useful classes and relations for a domain, it can be left open to monotonic revision thanks to the powers of the IsA relations. Wherever no specific class exists to capture the semantics of a particular data set, the application of a general class can usually express the data at least at this more generic level, while a process of revision is initiated between the knowledge engineers and domain specialists in order to specifically understand the nature of the new phenomenon and declare an appropriate subclass and/or relations to describe it in the model.

It is this same power of generalization and specialization which makes the method particularly useful for building data structures that enable inclusive and performant queries across data sets that have significant complexity and depth of expression. Through the extensible property of ontologies via specialization it is possible to model both highly specialized data structures while providing facilities to query this data at a more general level. In this way, the formal ontology approach avoids the traps associated with building sophisticated data models which are made unusable by their complexity both for end users and for program and database designers. The generalizations which allow for data integration also allow for inclusive searches where highly specific concepts and relations can be captured by general query patterns (Tzompanaki and Doerr 2012).

Ontologies and Their Encoding in Formal Languages

Having built up an ontology as a conceptual tool, if one wishes to run some automated reasoning processes over some body of collected knowledge encoded according to this ontology, the ontology must be represented in a formal language. Due to possible ambiguity in understanding, it is important to specify that the formalization of an ontology in a particular language results in an information artifact that is a representation of the initial ontology, but is distinct from the latter.

Representing the ontology in some formal language necessarily imposes constraints on modeling practice and inexorably alters the initial statements in order to fit them to the grammar of the chosen language.

That being said, it is through this trade off with pragmatics that functional automated reasoning through ontologies can be achieved. It is therefore of use to telescopically present some common methods for formalizing knowledge. While during the past 40 years several languages have been proposed and studied (KIF, KIR, KL-ONE among others) with the aim to meet this end with the KR community within the knowledge representation (KR) community (Hoekstra 2009), it only is during the last 15 years that, thanks to the practical needs brought forward by the semantic web community, a language of this type reaches a wider and more general public, more specifically with the development of RDF. Below we are going to give a concise account of a select subset of languages used to described web resources. The selection does not in any way mean to suggest a preference for one language over the others, but is based on the relative attention that the CH community has given to them.

RDF is the acronym of "Resource Description Framework," a data model for representing statements about resources in the semantic web. The assertions encoded in RDF take the form of subject, predicate, object <s,p,o>, where the predicate is a relation between the subject and the object, where both resources are available on the web. Such assertions are called triples. A collection of linked triples constitutes a graph, with the subject and the object of the assertions acting as nodes and properties as edges.

In order to keep a stable identity for the assertions created, each object is identified with a stable Web identifier, a unicode string called an IRI (Internationalized Resource Identifier); URL (Uniform Resource Locator) and URN (Uniform Resource Name) are particular types of IRI. The use of an identifier with a global scope is quite important because it helps in resolving the identity problem in the harmonization of different data sources. RDF also provides a machine processable XML-based syntax (RDF/XML) for recording and exchanging the propositions (Allemang and Hendler 2011; Manola et al. 2006).

It is important to underline that RDF itself do not define the meaning of a resource; for this task we should employ an ontology, which can be encoded with RDF syntax using the RDFS (RDF Schema) vocabulary. Even if the vocabulary employed in RDFS is quite small it allows the definition of classes and transitive subclass relations: basic taxonomical relationships. Moreover, it provides the possibility to define property and subproperties, as well as specify their domain and range, providing therefore a basic tool for the encoding of an ontology (Pan 2009).

The syntax and semantics of RDFS, as well as its meta-architecture, were in some cases not considered rich enough, and therefore other proposals for the construction of a KR language for the web have been made. The most successful attempt has been OWL (Ontology Web Language), a product of the Web Ontology Working Group of W3C, built upon RDF and RDFS. OWL is a richer language, and it allows to define features like the local scope of properties, cardinality restrictions, disjointness of classes and special properties (Transitive, Symmetric, etc.). It has three main varieties, OWL Full, OWL DL, and OWL Lite. Some of the main distinctions are the compatibility with RDFS, the restriction in the language and the efficiency in computation. Only OWL Full is fully backward compatible with RDFS (Antoniou and van Harmelen 2009; Allemang and Hendler 2011).

This excursus into some well-known encoding languages for formal ontology aims to underline that given the restrictions entailed by these languages, they should be chosen carefully, with the final application in mind. The use of OWL, for example, instead of RDFS restricts the expressiveness of your statements in exchange for making them more computable. Even the simple use of an XML-based language forces everything into a nested data structure.

It is also salient to highlight that the use of a certain language for expressing a data model does not automatically make the resultant product

an ontology. Having an OWL encoded file does not entail that it or the data therein is an expression of an ontology. It can, for example, simply mean that one has a taxonomy which is encoded in that specific language. Ontologies cannot be identified by a certain encoding, but rather, by whether or not they aim to explicitly represent an ontological commitment in some domain.

CIDOC CRM as Core Ontology for Data Aggregation in CH

In the field of cultural heritage, while there are a number (Mascardi et al. 2007) of widely known upper ontologies that can be brought to bear, the one which has most wide and official acceptance is CIDOC CRM (also referred to as CRM). At present, a great deal of research and implementation is happening around the CRM ontology extending it conceptually, applying it in new scenarios and developing large-scale implementations. For those interested then in the topic of data integration in CH, it seems, therefore, an opportune moment to recap the methodology and outcome of the development of CRM and to understand how this work is presently being extended in the service of CH research and preservation.

CIDOC CRM was initiated in order to solve an engineering problem of knowledge integration across museum databases faced by the International Council of Museums (ICOM) with regards to precisely the heterogeneity problems illustrated above. Following the intuition that there is a generality to the domain of museum information, ICOM had attempted to build a database prototype that would meet the needs of the entire museum community (Reed 1995). The resulting maximalist work was an impressive feat of research work, but resulted in a highly complex relational database model with over 400 tables that was difficult to in practice to put into effect. The problems of a maximalist position to solving the heterogeneity problem were thereby demonstrated empirically and invited a new approach. A turn towards formal ontologies appeared at the right moment to start a new approach towards this problem.

The first years of design effort yielded positive and encouraging results in terms of forming a satisfactory formal domain representation and led to the official creation of the CIDOC CRM Special Interest Group (SIG) in 2000. This group was tasked with the development of CIDOC CRM as an official ontological standard for the museums community. This task was achieved by the SIG by 2006, when CIDOC CRM officially became an ISO standard (ISO 21127:2006). In 2014, the ISO standing was renewed and updated with the development work of the preceding decade.[7] At present, CRM is the only ontology in the CH domain to have this official recognition, which can be read both a result of and also as a cause of its acceptance in the community.

To understand the grounds of the success and applicability of CIDOC CRM, we must review the methodological approach developed by the SIG. The goal will be to understand how it was developed, some key principles of modeling, how it can be applied and how it is being and can be further extended. The main elements of the methodology developed by the committee can be identified as: following an empirical approach; the principles of symmetric construction, context-free interpretation; designing bottom-up; and building modular but harmonized extensions and embedding the development process within communities of users.

The design strategy of the CIDOC CRM SIG was explicitly set as empirical in two basic senses. On the one hand, modeling is done only on the basis of existing information structures and their explanation by expert domain users. Information modeling always proceeds from practical examples and real use cases. Information structures are not built based on a priori theories whose concepts should be linked to the data structures to be modeled, but rather concepts are only derived from the input data structures. If there is no use case, then there is no basis for including a concept within the model, because

[7] http://www.iso.org/iso/catalogue_detail?csnumber=57832

there is no means against which to check the validity of the representation proposed.

This connects to the second and more fundamental sense in which the modeling undertaken by the SIG is explicitly empirical. As a guiding principle, the SIG conceives the data as representations of empirical facts stated in tabular format by cultural heritage specialists. The data is modeled not as an abstraction but always considered as having the same common referent of an objective reality. The actors involved are assumed to be engaged in an "ideal speech situation," describing empirical facts and leaving their propositions open to critical evaluation, against some validity criteria. This stance is taken in order to insist that it is not the researcher's abstractions that are to be modeled. This latter position would not allow for the construction of an integrative ontology but rather result in the modeling of a series of subjective perspectives. Instead, the position is taken that the kind of statement that the researcher is making is about a mutually available objective world which serves as a control to the modeling process. On this basis, we try to model the intent of the statement relative to a known and mutually accessible world.

The surface description and conception of data modeled by the domain specialist is not uncritically modeled, but rather a dialogue is opened to understand the underpinning scientific process and referents represented in the source schemas, testing moreover the conceptualization of the domain specialist against the accumulated experience of integrating hundreds of parallel data designs. In the case of noncoverage or conflict by the existing ontological elements, there is always an external referent to turn to, the world described, in order to seek an adjudication. Rather, than pitting theories against theories, then, and modeling data structures in the air, common understandings are sought by finding the middle objectively acceptable conceptualization. The result of this is therefore critical dialogue between the domain specialist, the knowledge modeler and the collected experience expressed in the converged model (Ciula and Eide 2014).

This empirical approach is supplemented by some specific design principles, which are useful to outline here. The first is symmetry (Doerr and Crofts 1999). The classes are modeled as neutral to a specific point of view within the domain, in order to prevent the description of identical facts as different ones only on the base on the perspective of the documenting actor. The prototypical example is that of E8 Acquisition Event, where the scope note clarifies that every transfer of legal ownership, comprising beginning or end of ownership, can be documented as an acquisition event. In this case, the class itself is constructed in order to avoid modeling the transaction from the perspective of one party or another (acquisition or deaccession), using instead the properties for disambiguate between who surrenders and who acquires the legal ownership of a physical object.

This kind of approach helps establish an ambiguity-free model, and moreover helps to introduce another important feature of CRM, context-free interpretation (Doerr 2003). The principle is to allow a clear interpretation of individual recorded propositions without any other type of contextual data. Thus, for example, saying that "John hasRole Buyer" does not really say anything about the action, and a context is required to understand the proposition. On the contrary if we encode that "John hasParticipedIn Activity" and link the buyer role to the form of participation it has a stronger information value, allowing greater integration of different information sources relative to the buying "of what, from who, when," etc. The assertions we represent with CRM are therefore structured purposefully to achieve this context free status. The advantages of such an approach are clear for the long-term analysis of the data, because they allow an unambiguous representation of the knowledge over the data, encoded in a transparent way, a practical matter which the OAIS (The Consultative Committee for Space Data Systems 2012) strongly advise for the long-term understanding of the preserved data.

The next key design strategy taken by the SIG is to build "bottom-up." This principle is closely related to but distinct from the empirical principle.

The objects from which the models are to be built should always begin at the lowest level. Modeling should begin from particular cases and create abstractions to capture the repeated appearance of such particular cases across different data sets. Generalizations are added to the ontology only once evidence that support the declaration of a general class is found across multiple abstractions. Building generalizations only on the basis of cases clarifies in advance their scope. That the generalization is fit to scope can be tested by making sure that it is logically consistent with the abstractions it generalizes for the purposes of querying and deduction. Generalizations are added or widened in scope by adding use cases. For example, the move to establish a generalized class for E7 Activity which defines an event with intentionality involved, is only made after modeling classes and relations for particular kinds of intentional actions, such as E8 Acquisition Event and E13 Attribute Assignment, that are induced from the form of the modeled sample data structures and allow for generalization to a general notion of intentional activity. The modeled data structures may in fact nowhere directly use a generalized "activity" concept but this concept can be extracted from the modeling of the particular action types. Thereafter, the generalization can serve to support higher level queries and deductions.

Here exactly lies the CIDOC CRM answer to the maximalist–minimalist conundrum described in our discussion of metadata and data schemas. The ontological model must be elaborated to provide specific classes and relations that unify particular data structures and data sets. This gives us the detailed level layer of querying in order to ask specific subject focused questions across specific data sets and stay within scope. But this layer of specific abstractions allows us to begin a generalization process over the specific abstracted classes and relations, whereby we look to discover their common properties and the unthematized implicit conceptual classes and relations that practitioners lean upon to perform analyses and investigate relations. In fact, exploring these generalizations moves us up and out of particular domains as it begins to find common structures of reasoning and thinking that stand at a very high level of generality. These high level concepts are rarely used explicitly, especially in a particular data structure, but they are the implicit conceptualizations that stand behind a wide array of reasoning processes. These become the top level classes and relations that are slowly consolidated and verified over many modeling exercises.

By building generalizations in this fashion, there is a rapid convergence in the initial phases towards higher level abstractions in the model, creating an increasingly more stable upper level model under which specializations can be better understood (Doerr et al. 2007). The long-term outcome of this strategy is a relatively slow moving and unchanging upper part of the ontology. The relative stability and slow moving nature of the resultant model then, principled additionally by the strictures of monotonic reasoning, strive towards the ideal of a formal ontology as an integrative tool, providing long-term broad expressive power for rendering commensurable heterogeneous data sets.

Such a slow moving structure also allows for the creation of modular extensions to the core model, based on a principle of harmonization. By design, CIDOC CRM is open ended, neither a maximalist nor a minimalist model, but a system of basic generalizations open to indefinite specialization according to the needs of the user community. The bottom-up methodology means that there is no end in principle to the specializations that can be made within the standard. The structure remains open to correction relative to the objective domain of discourse, and is practically enriched by the development of extensions which add use cases, supporting the existing structure or providing evidence to improve it or sharpen its distinctions. The core CRM through specialization thus binds ever more specific data sets to broader principles, allowing a wider range of communities to speak with greater specificity while connecting their data to a broader web of resources. To ensure an organized and orderly process of extension, the model is extended in a modular manner, dealing with areas of reasoning or patterns of activity of interest to particular user constituencies. Given the diversity of

approaches in the cultural heritage community, the potential extent of this specialization is virtually unlimited. Nevertheless, practically speaking, the limitations in place are three. First, there must exist a demonstrable use case. Second, there should be a wide enough set of data set exemplars in order to begin the empirical, bottom-up investigation. Lastly, there must be an institutional and community commitment to the development and support of the extension, so that it can not only be developed but maintained and evolved. In other words, the elaboration of an extension implies the same demands as the main ontology, but on a more specialized group of users.

This latter point, brings us to a final broader point with regards to the methodology of development for CIDOC CRM. To build a formal ontology to the level of a standard and maintain this status based on empirical, bottom-up design principles, entails a long-term investment in the intellectual work of building, maintaining and critically evaluating the ontology, in order to monitor the stability of its conceptualization and make adjustments to the constructs in response to the addition of new evidence from the user community. The value of an ontology depends on the willingness of a community to adopt it (Smith 2006). The key here is that there must be responsible parties who organize the ontology and create a feedback loop of use cases and potential critical data and observations from the user community so that the ontology evolves and is correctable. In the case of CIDOC CRM this feedback loop is created by CIDOC CRM SIG and its members. This group which meets several times throughout the year maintains a website to document the evolution of the ontology and its applications, documents its use and is constitutively open to the user community to engage directly with the SIG and/or join it, in order to apply CRM themselves or to critique it.[8] Because of the broad intended scope of such high level ontologies and therefore the learning curve in understanding and applying its concepts, the maintenance of the standard by such a community of experts is absolutely essential to ensure the integrity and applicability of the standard in real scenarios. By maintaining representatives from the major segments of the communities whose work is intended to be covered by its scope, the SIG aims to ensure that the standard is developed in light of a cross disciplinary critique and harmonization process that maintains organic connections horizontally and vertically across subdomains of research and scholarship.

[8] http://www.cidoc-crm.org/

The Basics of the CRM Model

The outcome of twenty years of modeling with the CRM has been the induction of a stable core set of generalizations that form a pattern of relations that can be repeated and specialized in any number of use cases and scenarios with success in the CH field. The ontology, now at version 6.2.1, stands at 92 classes and 153 relations. While there are new developments and monotonic evolutions, there is a stable core to the ontology which can be outlined in a compact manner and can serve as a guiding orientation for understanding how data is modeled in the CRM. In this section, we will briefly outline the top level categories of CRM (Fig. 1) and the discovery of the event oriented character of information understanding, storage and retrieval in CH contexts.

Looking to the CRM hierarchy, the important top level branches can be seen to be: E18 Physical Thing, E28 Conceptual Object, E39 Actor, E53 Place and E2 Temporal Entity. With the addition of entities for documenting E41 Appellation and E55 Type, we have already a powerful set of tools for documenting at the general level, all sorts of CH reasoning. While this picture will simplify a number of details, for the pragmatic purposes of modeling and mapping, this simplification provides a useful conceptual entry point into understanding the basic patterns identified and used repeatedly in CRM modeling.

E18 Physical Thing, E28 Conceptual Object and E39 Actor are all defined under the E77 Persistent Item class in order to indicate their status as endurant entities. Endurants are entities that have a persistent identity through time and can

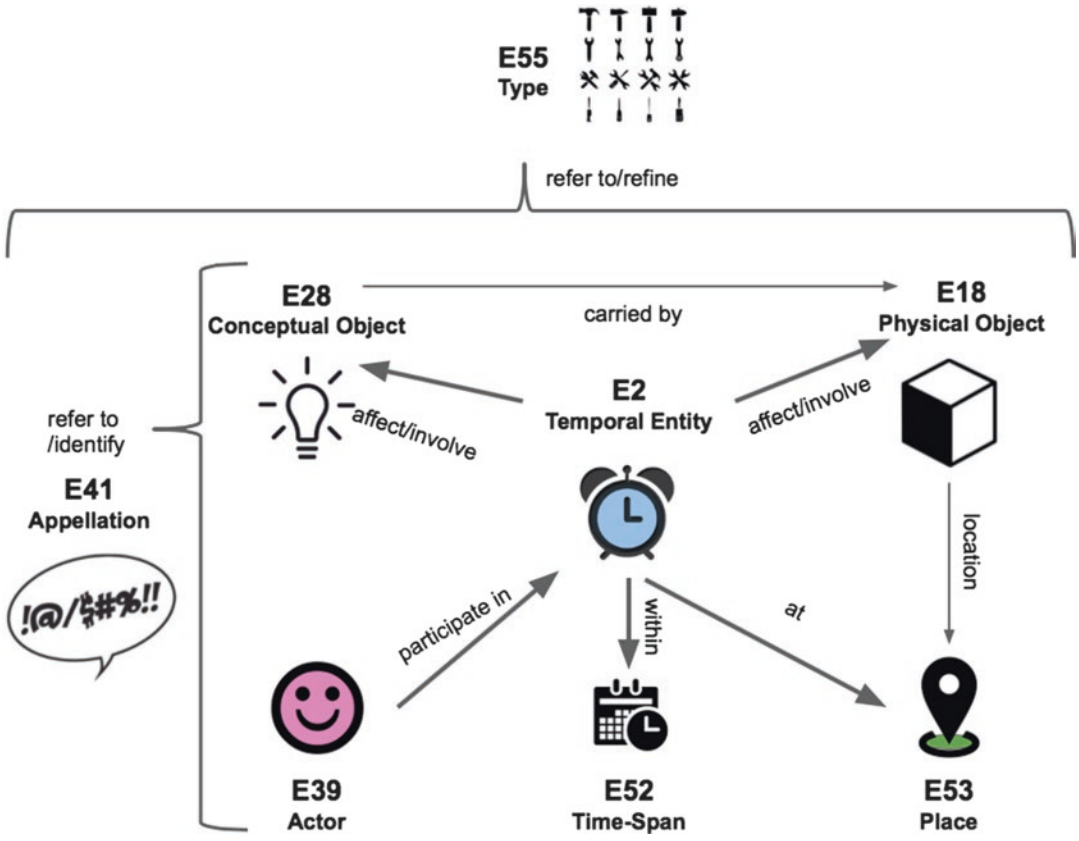

Fig. 1 CIDOC-CRM Top Level Categories

come into and out of relation, can be brought into or taken out of being as a whole or be subject to alterations which only accidentally modify them.

E18 Physical Thing is a class for all objects, man-made or not, that have relative stability of form over time and are physical. Understood quite simply, physical things are the objects of everyday human understanding in their materiality: tables, chairs, cats and dogs. Within the subclasses of this class important differentiations related to parts and wholes and natural vs. man-made are defined.

E28 Conceptual Object falls within the class of man-made and persistent but not physical things. The conceptual object class branch allows the documentation of those ideal objects which have been produced by human thought and ingenuity and that have taken on an identity in discourse such that they can be recognized when encountered in carrier formats. The subject of classification here is the idea or information of which the carriers form a sign: the idea of "Hamlet," the tune of "Waltzing Matilda," the choreography of "Swan Lake" by Julius Reisinger. Here we speak of unique identifiable ideas which can be found expressed in numerous materializations. Within the subclasses of E28 Conceptual Object different functional kinds of conceptual object are elaborated in order to track the specific relations and processes that pertain to such types. It is important to underline that images are also treated as a subclass of conceptual object, E36 Visual Item. We can find the same image repeated in many different carriers, meaning that it is a conceptual representation of something and not a physical object in itself. Concepts and images are not parts of objects themselves but are rather borne by physical objects and are expressions of the thought or interpretation of some actor. These distinctions

are often missed in information systems leading to the inability to link together data through the concepts or images they bear. This is a major differentiation in CRM, that allows a more accurate representation of concepts by distinguishing them from their carriers.

Additionally, within the branch of endurant items, CRM declares the class of E39 Actor. Actors are agents in the world. Actors are distinguished by their ability to perform intentional actions and in turn to be held to account for these actions. Here we speak of the actor purely in the sense of their agency as something endurant through time, divorced from individuals considered in their physical aspect. The E39 Actor class breaks down further into E74 Group, E40 Legal Body and E21 Person classes, representing important distinctions to track with regards to the particular relations that can come to exist and be studied with regards to agency in historical discourse. The last is also declared as a subclass of E20 Biological Object, in order to enable the expression of information regarding an individual human being as a physical entity.

All of the above are investigated as coming in and out of relation in time and space. E2 Temporal Entity has a different identity condition than endurants, having an identity through its coherence over a limited time. For practical purposes, the majority of E2 Temporal Entity instances can be considered as instances of its subclass E4 Period and E5 Event in which we are interested also in space, and the identity is given by the coherence of a physical or social phenomenon over a limited extent of time. Instances of E5 Event allow the documentation of coherent social or cultural phenomena that have specific durations, occur at specific places and form the units of discourse in which CH discourse seeks to understand the historical and causal relations between instances of E18 Physical Thing, E28 Conceptual Object and E39 Actor. E53 Place is declared and defined as a geometric extent.

Adding to these top level classes, two specialized classes exist for attributing names and types to any entity in the model. E41 Appellation can be linked to things, concepts, actors, temporal entities and places. E55 Type provides a mechanism for linking indefinite numbers of classifications to any class in the model. This means effectively that to any entity any number of names or classifications can be given, depending on the agent naming or classifying and the aims they have. This naming and classifying activity in turn can be documented, named, classified and studied.

What arose from the induction and application of these generalizations from the particulars of museum data, was the discovery of an event centric modeling pattern which proved the key for creating an indefinitely repeatable and specializable pattern of information relations. The event centric model is to be distinguished from the common tendency in information systems to focus data modeling on the object being researched and its properties (Doerr 2003; Doerr et al. 2007). The object and its relations, it turns out, are only the outcome of what is of most use and interest to the researcher to understand. Whether we are interested in the historical trajectory of ideas, people or things, what establishes the relations of interest between them is the event, considered as a temporal and spatially restricted coherency volume.

Events are places of the meeting of endurants that cause changes in relations in the world, where some endurants carry on as they are, others are modified and yet others pass away. Starting documentation from the event level allows for a clearer disambiguation between the perspective and aims of the person carrying out the documentation and the entity described, avoiding the construction of a set of classes where the properties reflect only the needs of a particular documentation situation. Coming back to the example of acquisition referred to above, it is a common tendency to document a transaction as a property of the object, when the focus of the documentation is on the object itself. The acquisition, however, is actually a context of understanding in which the object enters into a certain relation with different actors and its status changes as a result. A transaction is not a property of an object, but a relation to an event through which the object passes. By systematically avoiding such elisions of thought and making explicit such hidden enti-

ties—the events that connect things—we create documentation structures that are not limited to a certain perspective but in which we can join in a broad variety of perspectives.

The method of modeling data as always related through events, whether it be the generation processes of the production of a thing, the creation of a concept or the birth of an individual, or equally the dissolution processes of the same, allows for the modeling not of some static set of ossified properties of an object, but of the disposition of states of affairs in time that were capable of bringing about definite historical realities. It is thus that we can model and understand such diverse historical realities as the assassination of Caesar, the birth of a historical figure, the transmission of knowledge of an event such as the victory at the battle of marathon and even such contemporary phenomena such as the process of scanning an object for digital inventorying in a collection management process. While the different elements disposed within the coherency volume of the temporal event change, the basic reasoning pattern remains, to look for the meeting and separation of concepts, things and actors in time, at some place (Fig. 2).

CRM: Extensions and New Directions

Because of the manner of its production and maintenance (building generalizations over existing well understood classes and ensuring their relevance, moving generic properties up the class hierarchy, and harmonizing proposed new classes along the way) CRM is open to the extension of its practical scope to expand more or less indefinitely within the functionality of supporting data related to investigation of the human past. The solid top level categories function as a common ontology under which vertical extensions can

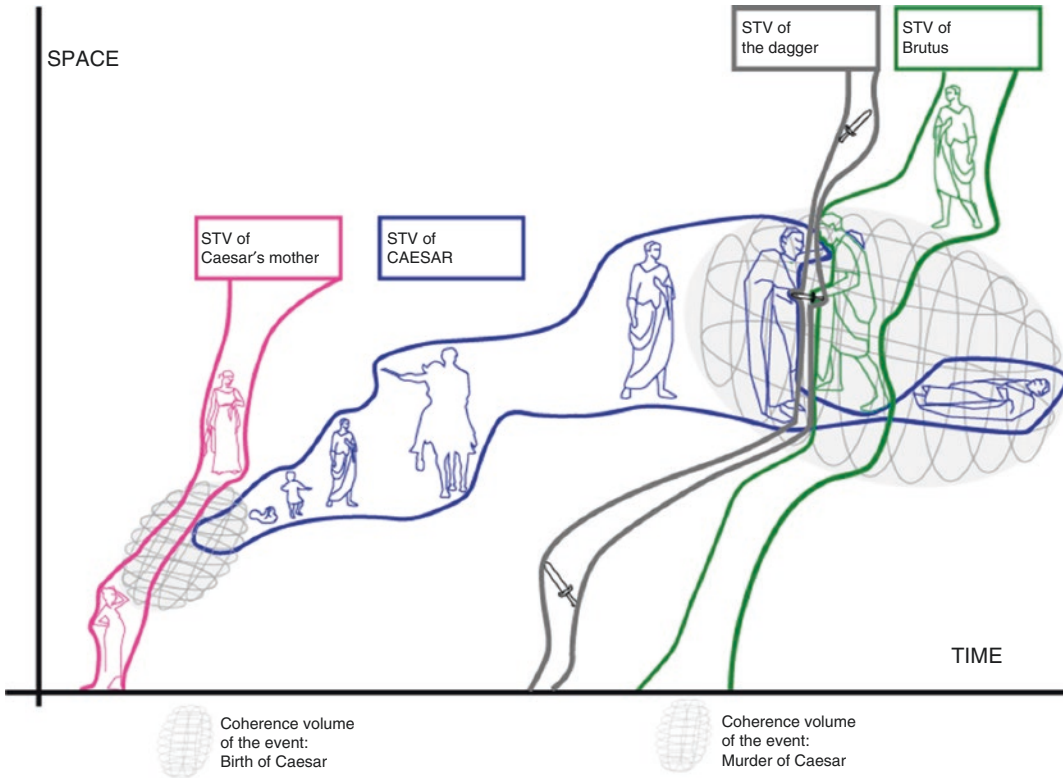

Fig. 2 Spacetime volumes theory as used in CRM

comfortably fit, extending and specializing the generic patterns seen at the top level. Thus, the standard, while retaining among its central constituency, museums, is being applied and extended not just within the scope of memory institutions in general, e.g. libraries and archives, but also within the scope of the analytic sciences and processes related to the research and discovery of the human past more broadly. This has meant that the potential for modular development of the CRM standard for the creation of integrating structures for specific subdomains of interest and practice within the CRM community has been taken up by an increasingly wide set of specialists within the CH domain.

The result is impressive because it manages to combine an integrative approach that enables the construction of a level of compatibility in two different directions. On the one hand, the proposed extensions tend to offer harmonizations of existing standards to allow a neutral expression amongst competing standards for some set of practices. At the same time, elaborated as CRM compatible extensions, these harmonized models allow integration to the broader scope of CRM expressed data. This has extremely high potential for creating novel connections between the knowledge ecosystems of different disciplinary groups who work on related data but do not normally share them. Thus, following in the tradition of enterprise systems, but at a broader level, the CRM is building the tools for a knowledge integration platform at a cross-disciplinary level for research on the human past.

In this section, we will provide a quick review of some of the most notable extensions that have been developed in the past years, highlighting notable features while leaving it to the reader the possibility to explore the details of the depth of the extensions of CRM. The main thrust of the direction of research in the past few years, while touching on many domains and ideas, could be said to be the question of provenance on one hand and how to connect different knowledge ecosystems on the other. Research into digital provenance (CRMdig) led to an examination of provenance in the sciences in general (CRMsci). From here these general provenance ideas were tested in particular domains, particular archaeology (CRMarchaeo and CRMba) and geo-spatial sciences (CRMgeo).

CRMdig

CRMdig is a model proposed for integrating data generated by digitization processes. At the time of writing it is in version 3.2 (Doerr and Theodoridou 2014) and has an expression in RDFS.[9] It is founded on the processes proposed by the Open Archival Information System (OAIS), customized and improved for covering the workflow for the creation of models. The general guiding motivation behind this extension was the sheer volume of work and funds being dedicated in the CH sector to the digitization of objects for their analysis, promotion and its geometric documentation. While the digitization sector is obviously a highly busy market with many competing products and methods, the needs in the CH community with regards to digitization information have certain particularities. These are driven by the fact that the object or series of objects are unique in some way. The CH professional's use of the digital models therefore goes beyond the need for a "pretty picture" of the heritage item. There is crucial information, which can be gathered from the digitization process. First, it is highly useful as a means to produce and preserve multiple measurements of the object in support of a better understanding of it. Second, it is interesting to trace the process itself not only to maintain data with regards to parameters going into the digitization process (and therefore scientifically evaluate the outcome and drawback of its analysis), but also to understand this digitization as part of the history of the objects itself. These facts lead to a modeling of this process which is uniquely concerned with provenance from the moment of transition between the physical and digital world to the many transformations that occur to digital objects once stored in some

[9] Available at: http://www.ics.forth.gr/isl/index_main.php?l=e&c=656 as of 14/4/2016.

digital environment (Doerr et al. 2010). Such modeling can support eventual reasoning over properties propagated through digital transformation (Tzompanaki et al. 2013).

On the object side, the important new classes have to do with integration of digital objects. Here the distinction held between the ideational content of intellectual creation and the particular physical carrier, which we discussed above, are found equally salient to the digital domain. Digital objects are modeled with a new class D1 Digital Object defined as a subtype of the CRM class E73 Information Object, pointing to its substance as encoded information whose identity is in the information held and its particular encoding, not the particular carrier (e.g. file). Likewise, a distinct class is proposed for documenting the particular carrier(s) on which digital objects are stored, D13 Digital Information Carrier.

The strong innovation of the model, however, is in representing the relevant events which lead to the creation of such objects. Events leading to the creation and modification of instances of D1 Digital Object are modeled under a D7 Digital Machine Event class which is purposively modeled as either the immediate or delayed result of a human action. This is to emphasize the ultimate causal origin of digital events in decisions and actions of human actors, who are the responsible agents to whom we can return for questions of data provenance. Instances of D7 Digital Machine Event are documented according to their relations to digital inputs and outputs—other instances of digital objects—and the effective parameters. A special subclass of D7 is modeled also as a CRM E16 Measurement class. The reasoning for declaring this class, D11 Digital Measurement Event, is that at the moment of digitization certain information is captured that holds measurement data of use in understanding of the object, but only when the circumstances of its production can be controlled for. We need to know the conditions under which digitization took place also as a physical event in order to evaluate the end product. Finally, processes that take place purely within the digital realm such as derivation and transfer activities are modeled with the D12 Data Transfer Event in order to be able to trace the results of transforms on data, the features that are preserved or lost from the original digitizations. A representative example of the use of CRMdig is provided in Fig. 3.

Originally developed in the EU funded project 3D-Coform, the model has been successfully deployed in the Greek national project 3D-SYSTEK,[10] in an NSF-funded project for RTI tools lead by Cultural Heritage Imaging, San Francisco,[11] in the ARIADNE project for scientific data in archaeology and in InGeoClouds for geological observational data.

CRMsci and CRMinf

CRMsci initiates a broader investigation of provenance relative to empirical science methodology. The extension is in version 1.2.3 (Doerr et al. 2015) at time of writing and has an RDFS expression.[12] It was built after an investigation of a number of unharmonized models related to different subdomains of empirical science practice. Specifically, the following models were considered: INSPIRE—earth science oriented, OBOE—life science oriented, SEEK—ecology oriented and Darwin Core—biodiversity. The aim of the model is to provide a neutral common ontology for integrating empirical science results which, in turn, creates an interface to CRM and thus the broad network of general CH information.

Aside from moving to a higher level of generalization for integration, there are two key differences between CRMsci and the models it integrates and generalizes over. First, thanks to the event-centric reasoning form, it more clearly formulates the distinction between the event of observation and its results, as well as the different modes of relation to the object under analysis which include a variety of acts including sampling and various forms of argumentation. Second, the identities of the observed object and the sampled object are more clearly defined.

[10] http://www.ics.forth.gr/isl/3D-SYSTEK/

[11] http://culturalheritageimaging.org/

[12] Accessible at: http://www.ics.forth.gr/isl/index_main.php?l=e&c=663 as of 21/3/16.

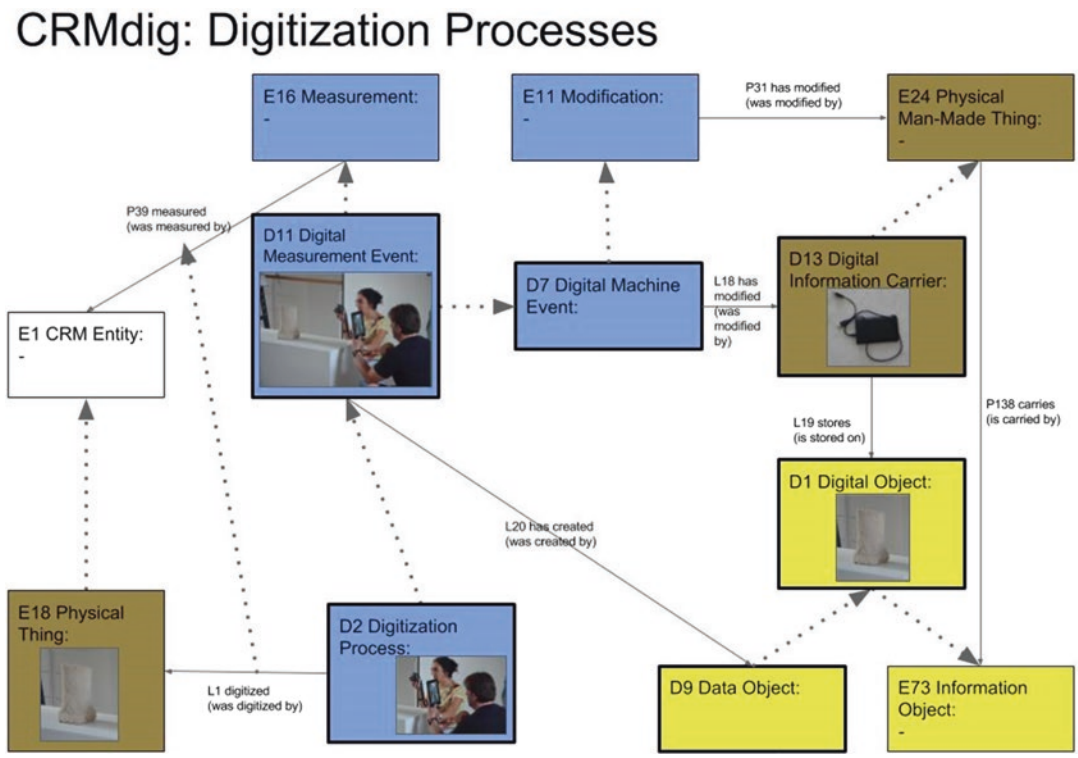

Fig. 3 Description of a digitization process using CRMdig

At the core of the proposed model is the S4 Observation class. Observation is modeled as an act limited in space and time which may or may not use devices and leads to an increase in scientific knowledge about states of physical reality. For this reason, S4 Observation is considered to be a subclass of CRM E13 Attribute Assignment. The latter class is used to model the activity of assignment of new attributes to existing things. Observation is distinguished from S2 Sample Taking which is defined as a case of matter removal. This step is potentially confused in other models. When sample taking explicitly entails a measurement, it is then modeled as S3 Measurement by Sampling and is declared as a subclass of S4 Observation. Other important new classes modeled include S19 Observation Event, which brings in the notion of "encounter" of particular use in archaeology, which allows the documentation of the moment of observation of a thing that is relevant to the research being undertaken and is considered as new to the community undertaking the research. The model also recognizes that the provenance of knowledge begins with observations but is built up through argumentation, for which it provides a number of classes for documenting the kinds of argumentation undertaken in empirical scientific discourse, namely: S8 Categorical Hypothesis Building, S7 Simulation or Prediction, and S6 Data Evaluation.

Indeed, because of the usefulness of the documentation of argumentation across all forms of scholarship, it was decided to extract the classes dealing with argumentation over factual states of affairs and develop a separate extension which can be implemented as a focused extension of CRMsci, borrowing its observation classes but allowing for the complete documentation of inferential argument and belief adoption. This extension is called CRMinf and is presently in version 0.7 (Paveprime Ltd. 2015). One of its particular innovations was to propose classes for the documentation of states of belief. This class, I2 Belief, allows for the documentation not of the

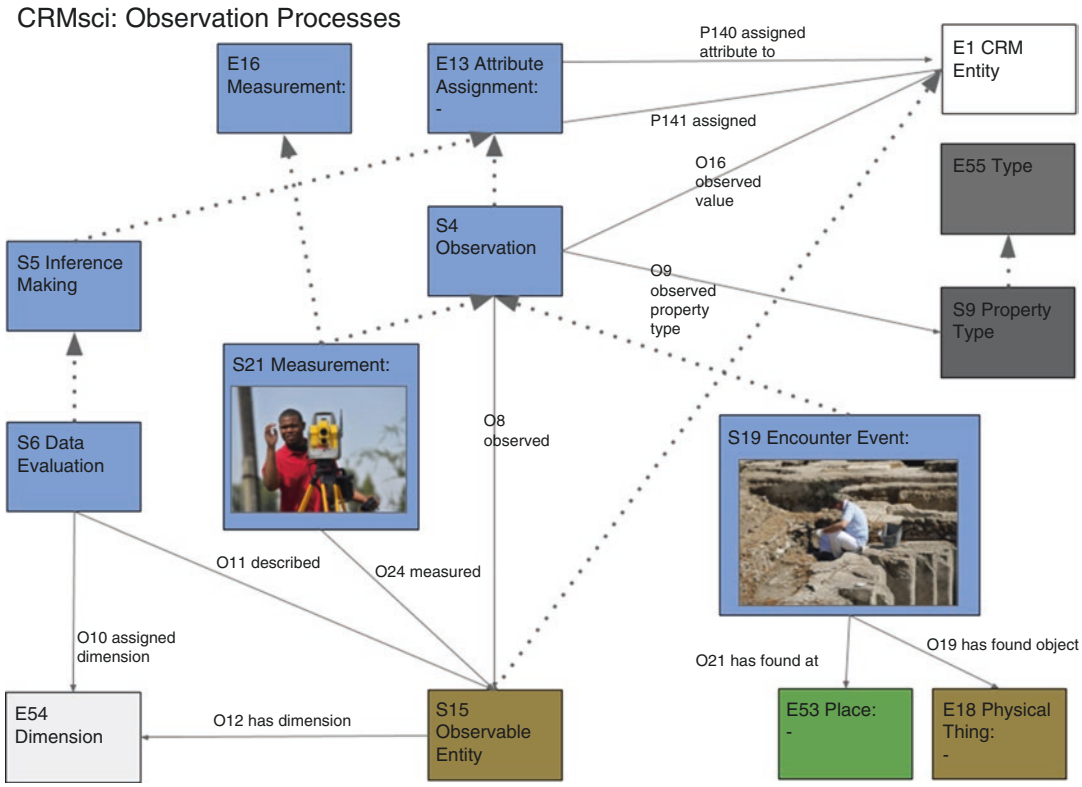

Fig. 4 Exemplar mapping of an observation process in CRMsci

truth or falsity of propositions, which would imply an absolute and completed process, but rather the documentation of the states of belief in the truth or falsity of propositions held by certain actors for certain periods of time based on certain logics—I3 Inference Logic—and certain sets of evidence—I4 Proposition Set. This extension was elaborated by Stephen Stead, also referencing work undertaken on a more expansive integrated argumentation model (Doerr et al. 2011). A representative example of CRMsci is provided in Fig. 4.

Equally as important to the proposed event classes in CRMsci are the changes which the model suggests with regards to top level classes in the CRM itself. In the CRM one of the basic divisions as explained above lies between the E77 Persistent Item and E2 Temporal Entities. Taking into account, however, the practices of empirical sciences, it is clear that that which the scientist or scholar observes can as much be a persistent item or an event. Furthermore, what the scholar or scientist attempts to observe can be as much conceptual as it can be physical. In fact, that which we attempt to observe tends to be some complex mix of things, their states, behaviors or their interactions and events. The observed entity can as much be observed by the senses as by digital devices that augment the sense, while concepts or events can be observed through their characteristic traces. Therefore, CRMsci declares a superclass that stands above the main division of perdurants E2 and E77 endurants called S15 Observable Entity. This allows the documentation of all possible observation acts. A representative example of CRMinf is provided in Fig. 5.

Presently this model informs and is being tested in the projects InGeoCloud,[13] Ariadne,[14]

[13] https://www.ingeoclouds.eu/

[14] http://www.ariadne-infrastructure.eu/

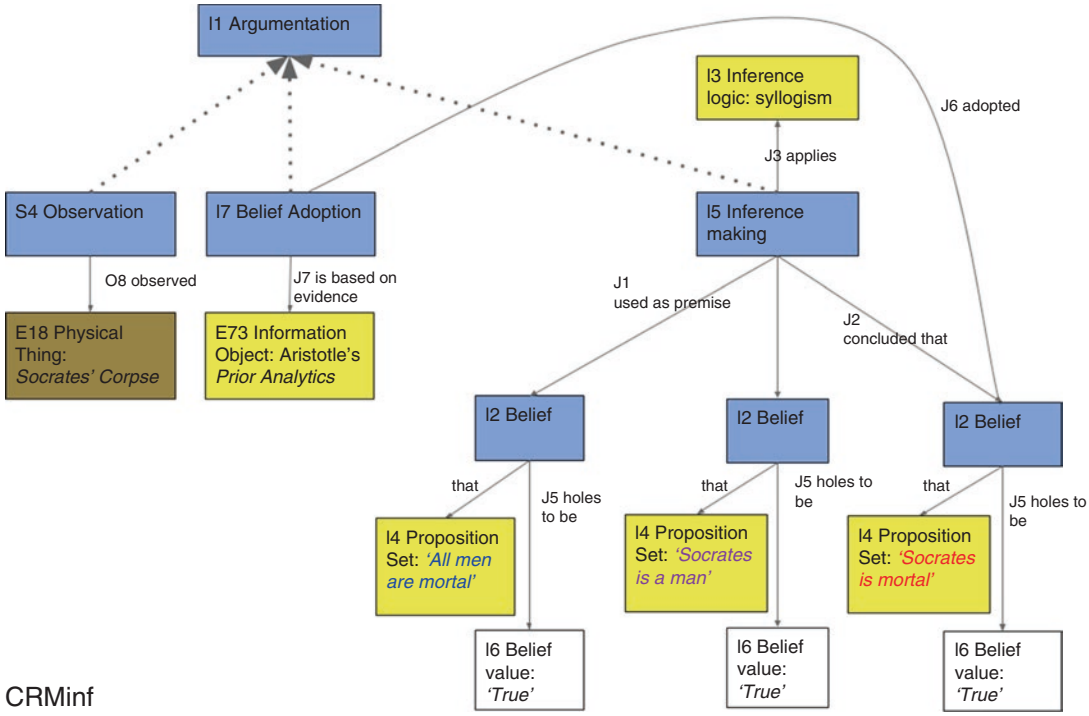

Fig. 5 The argumentation and inferences behind an observation mapped using CRMinf

iMarine[15] and as a base for a further extension in Monumentum (Messaoudi et al. 2015).

CRMArcheo and CRMba

Archaeologists represent an important segment of users within the CRM community. With regards to the issue of provenance, they face a very acute and particular problem specifically during the collection and retrieval of commensurable data about the excavation process. For this reason, a group formed to create a particular provenance model for excavation data, called CRMarchaeo. The extension, at time of writing, has reached version 1.4 (Doerr et al. 2016) and an encoding in RDFS is also available.[16] The rationale behind the construction of the model was to maximize the interpretive capability and reassessment of the data created during an excavation. In particular, because of the destructive nature of the archaeological process, the accurate and explicit recording of the actions of the excavation into a document are key to the validity and usability of the action. And yet, despite the universal recognition of this fact within the discipline, a standardized model, both for providing an intellectual guide to the creation of archaeological recording systems and/or for allowing the comparison of the stored data, is not available. CRMarchaeo was devised collaboratively across seven participating institutions, analyzing the data structures and protocols from across Europe. The resulting model supports knowledge provenance and comparison across archaeological datasets.

Excavation archaeology provides a powerful, closed knowledge paradigm for modeling because it relies on common reasoning, detection of events through depositional sequences, and a

[15] http://www.i-marine.eu/Pages/Home.aspx

[16] Availabe at: http://www.ics.forth.gr/isl/index_main.php?l=e&c=711 as of 14/4/2016.

commitment to systematic observation techniques. Therefore, the task was to model, on the event side, the typical events of the excavation activity and, on the object side, the identity of that which is excavated and the relation between strata. The chief class on the object side is the A8 Stratigraphic Unit which is seen to be the result of an A4 Stratigraphic Genesis Event. The physical ordering of stratigraphic units as being on, above, below, cutting each other, etc., aids in arguing for the chronological order of events and the construction of relative chronologies. The notion of an object as an A7 Embedding in an A8 Stratigraphic Unit documents the object as it is understood by the archaeologist as a record of a present state that may shed light on a state of affairs in the past as well as enabling the object embedded to be qualified separately as an instance of E18 Physical Thing. The chief unit of documentation for capturing the event of excavation is the A1 Excavation Process Unit. It is modeled as a subclass of CRMsci's S4 Observation class, because it is considered as a specialized form of observation. In particular, a number of relations are expressed in order to capture the precise changes that the excavation activity brings about in the physical remains, especially stratigraphy, under study, in order to be able to reconstruct this process. A representative example of CRMarcheo is provided in Fig. 6.

The work on CRMarchaeo was followed up by the thesis of Ronzino proposing to further elaborate the latter to include the methods and practices of building archaeology (Ronzino et al. 2016; Ronzino 2015), which also uses the notion of a stratigraphic unit in order to reason over the order of production, modification and destruction of a building. At time of writing, CRMba is currently a standalone extension but work is presently being done to test and harmonize it with CRMArchaeo, especially in order to tackle the difficulties of representing the spaces of buildings. By introducing the concept of B4 Empty Morphological Building Section alongside B3

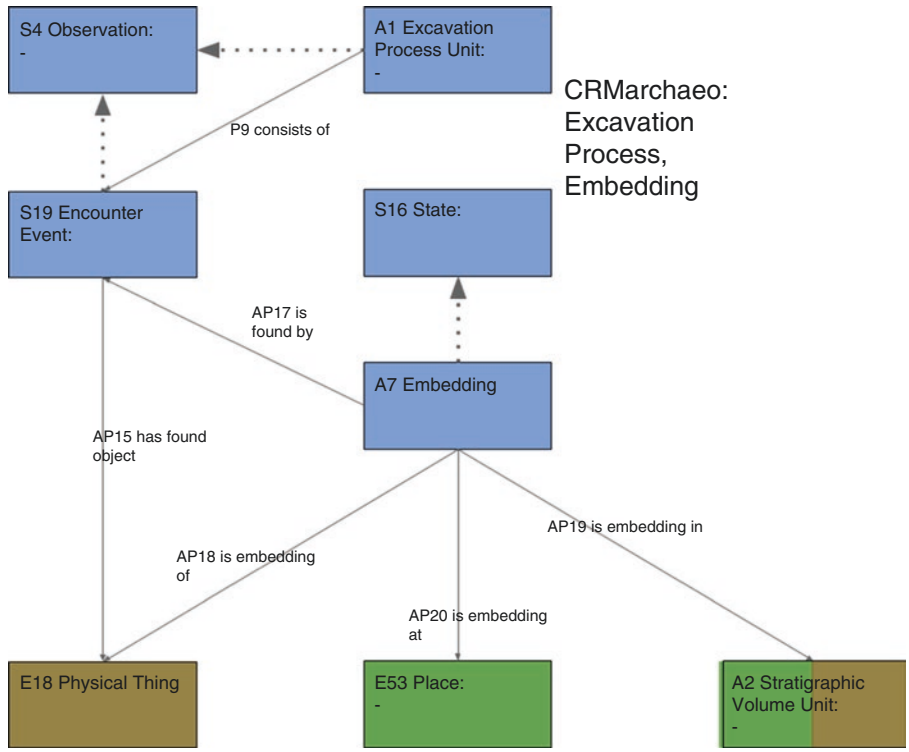

Fig. 6 An archeological excavation process mapped with CRMarcheo

Filled Morphological Building Section, where the former can be bound and filled by the latter and the latter is made up of instances of B5 Stratigraphic Building Unit, the model provides a comprehensive way to bring together data of relevance to building archaeology and to document the physical presence and absence of the architectural space.

CRMarchaeo is presently used in the Ariadne project where its implementation has been considered in many contexts (Masur et al. 2013; Hiebel et al. 2014; Aspöck and Masur 2015).

CRMgeo

Finally, in order to support reasoning over space in the geophysical sense, an effort was undertaken to integrate the OGC/ISO standards for geographic information and the CRM (Doerr and Hiebel 2013). The proposed model, in version 1.2 at time or writing, has an RDFS encoding.[17] This move was motivated by the heavy interest of tying CH data to georeferenced data sets, in order to increase potential understanding and analysis. The analysis of how to bring about this join was particularly productive with regards to modeling geographic reasoning more precisely. Whereas the core CRM model makes reference only to E53 Place as a geometric abstraction, the needs for understanding the reasoning process in making geophysical arguments led to an extended investigation that would posit several new ideas, some of which would retrospectively be brought into the CRM core.

Particularly, modeling geophysical reasoning opened up the distinction between the phenomenal place and the declarative place. That which we want to define geometrically is actually outside of our ultimate measurement, because it is an object always in becoming, always beyond final fixing. From a physics point of view, we are interested in defining an SP1 Phenomenal Spacetime Volume. But in doing so we need to distinguish clearly the SP7 Declarative Spacetime Volume, that is the declarative spacetime by which we try to capture the phenomenon, from the phenomenon itself. In particular, in order to begin the process of approximating some real space time volume, whether we consider it completely or in some spatial or temporal projection, we must declare a space time volume which we think approximates it. Such expressions however only make sense relative to some system of projection. Here a class is declared for documenting such projections, SP4 Spatial Coordinate Reference Systems. The system of projection in turn only makes sense in regards to some fixed points in a physical world that hold for some period of time which can also be documented and correctly related to these events of approximation.

This issue, therefore, pushes us back to the general question of provenance. In fact, georeferenced data provides approximations of real things or activities that occurred which we can trace by looking for typical forms of evidence depending on the target phenomena. But the knowledge generated is not absolute, even if the research is highly successful but is bound to particular forms of projection related to typical physical reference features that do change, no matter how slowly, over time.

Aside from the creating a powerful interface by which to join OGC generated data to CRM compatible data, the major achievement of CRMgeo was to introduce the concept of SpaceTime volumes into CRM core. The high level entity, E92 Spacetime Volume, has officially been added into CRM core and enters the hierarchy as a superclass of E2 Temporal Entity and E77 Persistent Item. That which we observe, be it a perdurant or endurant is something which we can potentially reason over and track either with regard to its entire path through time or to understand where it had been and in contact with what, when. In fact, this returns us to the coherence volume reasoning of the original CRM (Fig. 2) but now provide tools for documenting and tracing these relations in a mathematically more precise way. The introduction of this class enabled the work of Papadakis (2014) to model more accurate time relation operators than the

[17] Available at http://www.ics.forth.gr/isl/index_main.php?l=e&c=661 as of: 22/3/2016.

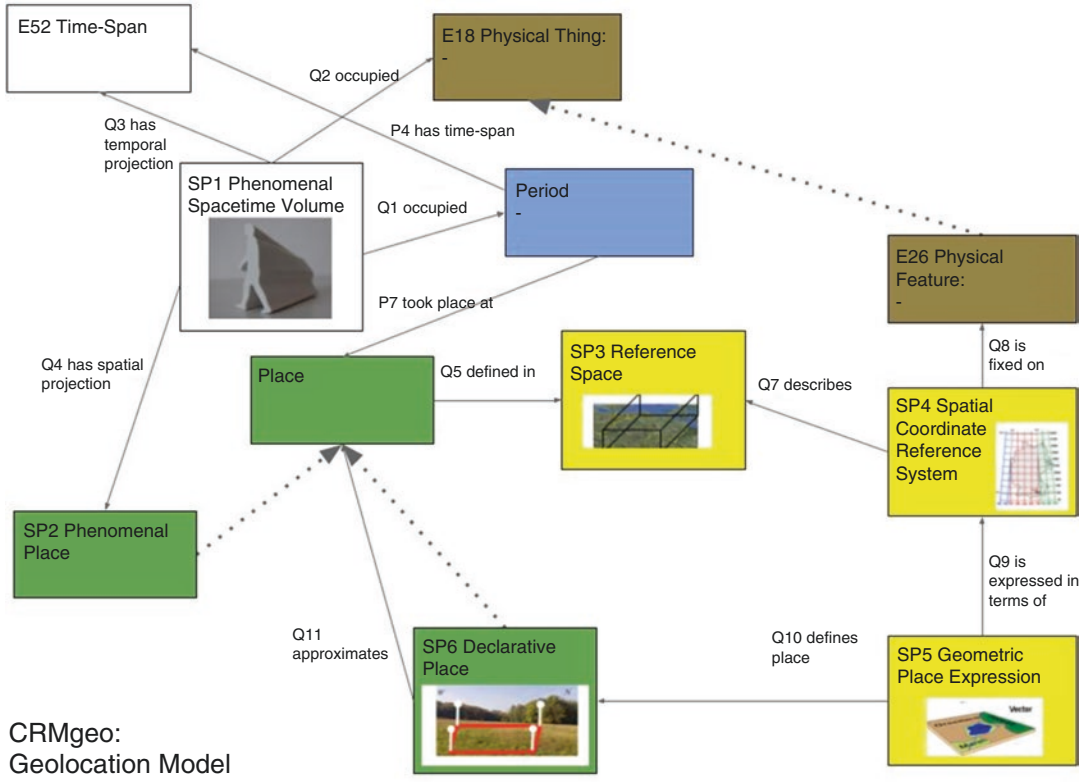

Fig. 7 Example of use of CRMgeo

Allen operators, by allowing for fuzzy volume reasoning on time relations based on positive and negative evidence indicators. A representative example of CRMgeo is provided in Fig. 7.

CRM in Implementation

With the wide acceptance of the core CRM model and the growth of specific extensions for different tasks, technical development of implementations that take advantage of the model are spreading. Common aspects of these implementations tend to be a commitment to the production of an open source platform which is extensible by the user community. They tend to have arisen from the effort to solve either an intra-institutional or inter-institutional data aggregation and sharing problem, but they have been developed with an eye to wider-scale adoption. Here we will briefly introduce four such projects which are running in United Kingdom, Germany, United States and Italy, their goals and the means that they set out to achieve them.

Research Space Project

Research Space is a project supported by the Mellon foundation that, amongst others, has built the integrative data infrastructure for the web presence of the collections of the British Museum.[18] As an institution with a great depth of historical information and breadth of departments creating data, it presents a particular challenge to the goal of data aggregation of heterogeneous sources. Research Space took on the challenge of how to bring the various data

[18] http://www.researchspace.org/

from different departments and curatorial traditions into a unified platform for public use with the aim not only of not losing information in the aggregation process, but in rendering more explicit the connections amongst data and their origin to the public audience. While developed in relation to the data of the British Museum, the software being developed will be released as open source on maturity.

The tool produced allows for the ingestion of mapped data sets from various sources, from humanities to natural sciences, to be represented in a unified environment enabled with a semantic search tool. This tool takes advantage of the notion of "fundamental categories" (Tzompanaki and Doerr 2012) in order to create a simplified expression of the CRM model with regards to search, so as to be able to empower normal users to make sophisticated semantic searches in an intuitive manner. The main search interface allows the user to choose to perform contextual search over Things, People, Places, Period, Time Spans and Concepts. The search tool intuitively allows and guides users to create searches that explore the relations between these entities. These semantic searches are made possible by the ingestion of CRM mapped, distributed data sources.

The software platform is distributed in a prepackaged Docker format which provides a preconfigured operating environment in which Research Space runs. It implements the Blazegraph graph database system which stores and manages the Linked Data produced, MetaPhacts which provides a Semantic Wiki environment, and the Research Space integrated environment. The platform runs on TomCat server, implements the Apache Solr indexing system and is written in the Java programming language. The system is web based and allows users to access through a browser.

The software will be available as open source and would be installable locally, at an institution or inter-institutional level. The provisioning model to date suggests that operating the software as a shared service may be the most efficient way of implementing it since it eliminates the need for local institutional setup and maintenance of the system and shared service providers could provide support in terms of expertise in cultural heritage information, data modeling and management, application customization, digital preservation, and access to a growing repository of data.

WissKI Project

WissKI is a German acronym for "**Wiss**enschaftliche **K**ommunikations**I**nfrastruktur," which can be translated as "Scientific Communication Infrastructure". It is a joint venture project supported by three partners from different institutions and scientific domains: the Digital Humanities Research Group of the Department of Computer Science at the Friedrich-Alexander-University of Erlangen-Nuremberg (FAU), the Department of Museum Informatics at the Germanisches Nationalmuseum (GNM) in Nuremberg, the Biodiversity Informatics Group at the Zoologisches Forschungsmuseum Alexander Koenig (ZFMK) in Bonn.[19]

The WissKI project has the goal of providing scholars and scientists with the technical means to model and then express their data in a CRM enabled system (Scholz et al. 2012; Scholz 2013). Particularly, they want to support researchers to move beyond local files and databases to an online, CRM integrated virtual research environment. WissKI provides a system in which CRM compliant semantic data can be created either in a Wiki style format or through a familiar tabular data entry format. The WissKI system enables the loading of the Erlangen CRM/OWL encoded implementation of CIDOC CRM[20] which it uses as its recommended standard. Additionally, a system ontology which extends the CRM for application purposes, is loaded into the system. The Wiki environment allows manual marking of entities in texts as well as the configuration of tools for named entity recognition of people, places and periods among others. A purpose built pathbuilder system, allows administrators to

[19] http://wiss-ki.eu/

[20] http://erlangen-crm.org/

build form interfaces that express CRM paths but allow intuitive user data entry. The scholar is then enabled through common forms to produce data that are CIDOC CRM compliant. Data standardization is further enabled by the systems support for the use of controlled vocabularies either defined locally or pulled from linked open data sources.

In order to do this, they have built an extension to the popular CMS Drupal, which extends the system to make use of a triple store such as ARC2 (Mysql-based), or Sesame. Thus, the end user and administrator can access the system through a browser in a relatively low demand computing environment. The system, moreover, has been developed as an open source software. It is therefore available to any scholar to download, install, customize and begin producing data that will be CRM compatible.

From an institutional point of view, for the participating members of the consortium, this creates a long-term cross searchable pool of knowledge, with a single update trajectory. From an epistemic point of view, the project enables the coexistence of humanities and natural science research within the same system, allowing cross-disciplinary searches that would not otherwise be possible. From a practical point of view, the system enables multiuser, internet based access to a common registry system that can be used, at the same time, as the public data delivery system, for making research results available to the public audience.

ARCHES Project

Arches[21] is another project which proposes an open source software that implements CIDOC CRM at its core. The software was developed jointly by the Getty Conservation Institute and World Monuments Fund (Carlisle et al. 2014).

A strong feature of the original focus of the project was on the management of built cultural heritage and environments. It therefore has powerful built in support for GIS data management especially using OGC/ISO standards. The functionalities which the Arches aims to support include, "identification and inventory, research and analysis, monitoring and risk mapping, planning for investigation and research, conservation and management, and raising awareness among the public, governmental authorities, and decision makers." (Arches Factsheet 2015)

The logic behind Arches returns somewhat to the original efforts of ICOM to build a single system for cultural heritage management. Recognizing the similarity of the intellectual and practical challenges faced by CH institutes, it offers an advanced data management system specifically designed for use by CH institutions for free. That being said, the Arches project foresees the open ended expansion of the database and its functionality by adopting CIDOC CRM for modeling the data documented therein.

With regards to the semantic representation of data, the system is released with CIDOC CRM in Gephi graph format pre-encoded for the basic entities in the database. As the user expands the database system, they can extend this modeling, describing the semantic relations of the data stored in the relational database with CRM. This ensures the long-term interoperability of data generated through Arches independent of the project itself.

The system runs on PostgreSQL, PostGIS, and uses Python and GEOS. It is available as open source software and is not envisioned as a data aggregation tool but as an institutional data management and repository system. End users access the system through a browser. It allows multiuser access and data entry and management with different user roles and rights.

ONTOP

The use of CIDOC CRM as conceptual layer to query Relational Database was lately investigated, mostly using the open-source Ontology-Based Data Access (OBDA) framework Ontop[22]. The latter, developed by the Free University of

[21] http://archesproject.org/

[22] http://ontop.inf.unibz.it/

Bozen-Bolzano, acts as a translator between an ontology, a previously given mapping, and the set of data. The mapping can be given in R2RML and allow to construct a declarative specification that relates the elements of the conceptual layer and the data layer/s used. Thanks to the mapping Ontop is able to generate a virtual RDF graph, which can be queried using SPARQL. The use of SPARQL engine Quest allows an on-the-fly re-writing of the SPARQL queries into complex SQL queries. Ontop can be used as a Protégé plugin, a Java library or a SPARQL end-point through Sesame's Workbench. Ontop it is not integrated at a database level, therefore it does not alter any previous schema, but it is quite useful to re-use the SQL based resources, or as a federated database (Bagosi et al. 2014; Kontchakov et al. 2014; Rodriguez-Muro and Rezk 2015).

This solution was recently used in a few projects (Le Goff et al. 2014; Mosca et al. 2015; Calvanese et al. 2015) dealing with cultural heritage, mostly for re-using existing resources, and for merging different types of data coming from diverse source. Ontop is a mid-level solution between a full implementation of a RDF store and the current state of RDM, and it can be quite useful in an intermediate phase where an institution has already a big amount of data stored in few databases across its projects and do not want or do not have the resources for facing a complete migration towards a triplestore.

CRM Looking Forward: Expansion, Application and Education

So, as it enters its 20th year of research and 10th year as a formal ontology standard, CIDOC CRM presents both a viable tool for integration of CH data in the here and now but also an active area of research in itself to seek ever better ways to structure systematic research data. It is useful at this phase to round up the present challenges and new research directions that present themselves as topics for expansion in the coming years. To summarize this challenge, we could argue that CRM is at crucial juncture with regards to reaping the benefits of the conceptualization work by intensifying implementation cases and this, in turn, entails a greater popularization of the methods and techniques of CRM modeling from computer science specialists to domain experts. This latter transfer of knowledge, which is already underway, does and will continue to allow specialist communities to seek to elaborate their own extensions in order to formulate general models for specific objects or kinds of research that will nevertheless be able to benefit from compatibility with a universe of provenanced data made available by other researchers through a network of knowledge.

With the core standard at a strong state of maturity with very few and slow changes to the high level conceptualizations being necessary, even while integrating a strong suite of extensions, the robustness of the ontology has shown itself over time. It is able to perform integration over its originally intended data sets, plus over data from memory institutions and CH heritage research considered more broadly. In many cases, it is able to perform this integration with the help of extensions in order to be able to support the specific reasoning processes of subdirections of research within CH communities. What is beginning to be built now through projects such as the Research Space, WissKI, Arches and Ontop among others are the kind of mappings of broad and extensive datasets that will scale up the CIDOC CRM offer by providing a wide array of sample data, providing a practical demonstration of its effectiveness as a tool, moreover providing extensive examples of data mapping from different types of research areas, useful for experts to refer to in thinking through a mapping of their own. This growth of CRM expressed data can be the kernel of as an ever expanding network from which to work from and respond to (building a virtuous circle of data implementations) as part of normal CH practice.

While, as demonstrated above, many projects, small and large, have either mapped their own data to CIDOC CRM or extended it on the base of their own requirements, in order for this work to benefit from a larger mass effect that supports day to day CH work and research, it is necessary

for many more datasets to be mapped to CRM and to have a home for their integration.

Towards this end, a number of tools for mapping data from a source data schema into a target schema are available. OpenRefine and KARMA are two such tools which allow an easy mapping. The former (Verborgh and De Wilde 2013), previously known as Google Refine, can, in association with the RDF Extension, transform and map manually or semi-automatically tabular, JSON, XML data into RDF file, on the base of an ontology of choice. KARMA (Szekely et al. 2013) works in a similar way but no external modules are necessary for this operation, and, moreover, it provides an easy-to-use visual interface. While these are powerful and useful tools, one feature which they do not provide is a community memory of mapping solutions. X3ML (Marketakis et al. 2016) is another tool in this field that offers both a mapping manager that allows the systematic mapping of datasets along with mapping scholia and the eventual transformation of datasets into RDF instances along with the ability to store previous mappings in a library of mappings in order to support future work. What this allows for is a body of repeatable knowledge with regards to how to map different dataset types to CRM (or any other schema for that matter), a body of knowledge to support the community of researchers and, practically, to allow tactical exports of data from data entry systems into CRM format for integration into aggregation infrastructure.

Such work, however, reaches a natural bottleneck depending on the general spread of knowledge of how to use and apply formal ontologies and particularly CIDOC CRM. In order to build such a virtuous cycle, datasets that are already produced by researchers and professionals on a daily basis must be mapped to the standard. Here, however, it is neither possible technically nor practically for the load of the work to fall to a cadre of computer scientists to implement mappings from CH data sets into CIDOC CRM. As described above, the entire method of empirical ontology development is interdisciplinary. While most researchers will likely not display a direct interest in developing or expanding on an ontology as such, insofar as they want to express their data in such a common system, it requires an understanding of the ontology because the data producer is best placed to produce the most representative translation of their data into the common expression. It is the domain specialist who has the knowledge of what their data means and what questions they want to be able to ask of it. Aside from avoiding obvious errors of syntax and misunderstanding of terms, there is no "correct way" to map to CIDOC CRM or any standard. There is no one size fits all solution for a dataset especially if the data schema is a purpose built data schema. There are patterns of mapping that can and should be elaborated, but at the end of the day, knowledge is not in the machine but in the researcher. For this reason, one of the main challenges in the coming years with regards to CRM is to build up training materials and tools which can communicate its use to the level of specificity that a domain specialist, in the first instance, might want it. That is to say, the domain specialist has interest in the CRM not as an end in itself, but as a new means of expression of their data which both make it more accessible but also more connected. The domain specialist wants to add to a collection of information and to take back from that collection of information in order to achieve some task. Mapping should become a natural part of this process, not as an end in itself but as a means to facilitate this goal on a broader, more automated and efficient level.

What this requires is not primarily a question of computer science but is instead a question of how to achieve social embeddedness of these techniques of knowledge sharing and propagation in a manner that makes such procedures an integral part of scientific practice. The problem with kick starting and spreading formulas for data sharing and aggregation lies in the lack of institutional and social frameworks that truly value and have the pragmatic business interest to support these activities in the long term. So long as the effort to bring about such ends is viewed as something extra to or even competing with everyday needs in CH management and research, such efforts will be faltering and executed on a case-by-case basis. With the maturity of CRM as a

standard however, there is presently the opportunity to build data sharing and standards informed by knowledge engineering principles into broader curricula in the CH sector, to inform the daily practices of specialists. The latter have the opportunity to spearhead implementations which will build a critical understanding of the methodology in order to obtain the desired goals of rendering research resources more transparent, accessible and findable as well as having access to broader data resources in return. This move would be able to draw from the experiences in enterprise of building enterprise resource planning and strategic planning through data integration tools. Again, such efforts aren't really goals in themselves but would actually form part of a more general strategy of taking control of and understanding data at a broader intra-disciplinary and cross-disciplinary level.

It is this last move which promises some of the most interesting problems to research at a general level with regards to the development of formal ontologies and CIDOC CRM in the coming years. As more data is modeled and expressed in top level compatible models and the questions that have begun to be opened in terms of knowledge provenance and acts of knowledge creation are explicitly encoded, we face great challenges in terms of understanding and modeling the processes of knowledge production within specific communities or knowledge ecosystems, in terms of who generates knowledge, with what and for whom. Then we face the additional question of what these users of produced knowledge, in turn, do with that knowledge. Modeling such knowledge ecosystems individually also opens the challenge of building information systems that are able to represent the relations of the generated data across disciplines so that new, broader cross-disciplinary exchanges and even programs can be supported and engaged in. Understanding how and to what detail argumentation and experiment can be modeled in detail in a tabular format in order to support the repeatability and testability of produced information for the generation of new knowledge is a large open challenge connected to this problem. Likewise, in a related issue, building trust in large-scale data pools by ensuring authenticity of data and being able to attribute data to responsible persons and institutions forms an important domain of research.

Conclusion

In this chapter, we aimed to look at the problem of data heterogeneity and aggregation and the potentials of formal ontology especially CIDOC CRM to address this challenge in the CH field. We laid out a view of the nature of cultural heritage data as a complex but unified phenomenon whose identity is fixed by the common interest in the scientific investigation of the human past. We proceeded to an analysis of its necessary and accidental sources of heterogeneity. In order to understand the proposition of formal ontology as a solution to data heterogeneity in large-scale aggregation within its historic and technologic context, we looked at the traditional understanding of categorization and how it informs and is used in systems for data management such as classification schemas, taxonomies, thesauri and protocols, looking at the use and limits of such systems. We then introduced formal ontology in general and the approach proposed for CH in the CIDOC CRM standard. The latter half of the chapter introduced the innovations in the CIDOC CRM standard in terms of the development of modular extensions to deal both with discipline specific problems and general problems of knowledge provenance. Finally, we introduced a number of paradigmatic implementation projects offering examples of possible implementation strategies, using this as a means to introduce the question of the possible future directions of implementation and research.

References

Allemang, Dean, and James A. Hendler. 2011. *Semantic Web for the Working Ontologist – Effective Modeling in RDFS and OWL*. Second ed. San Francisco: Morgan Kaufmann.

Antoniou, Grigoris, and Frank van Harmelen. 2009. Web Ontology Language: OWL. In *Handbook on Ontologies*.

Arches Factsheet. 2015. Getty Conservation Institute.

Aspöck, Edeltraud, and Anja Masur. 2015. Digitizing early farming cultures customizing the arches heritage inventory & management system. In *2015 Digital Heritage*, 2: 463–464. IEEE.

Baca, Murtha, Patricia Harpring, Elisa Lanzi, Linda McRae, and Ann Whiteside. 2006. *Cataloging cultural objects*. A Guide to Describing Cultural Works and Their Images. American Library Association.

Bagosi, Timea, Diego Calvanese, Josef Hardi, Sarah Komla-Ebri, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, Mindaugas Slusnys, and Guohui Xiao. 2014. The ontop framework for ontology based data access. In *The Semantic Web and Web Science*, 480: 67–77. Berlin, Heidelberg: Springer.

Bergamaschi, Sonia, Silvana Castano, S. De Capitani Di Vimercati, S. Montanari, and Maurizio Vincini. 1998. An intelligent approach to information integration. In *Formal Ontology in Information Systems*, 253–267.

Bouchou, Béatrice, and Cheikh Niang. 2014. Semantic mediator querying. In *IDEAS*, 29–38. BytePress. doi:10.1145/2628194.2628218.

Brachman, R.J. 1983. What IS-A is and Isn't: An analysis of taxonomic links in semantic networks. *Computer* 16: 30–36. doi:10.1109/MC.1983.1654194.

Brachman, Ronald J., and Hector J. Levesque. 2004. *Knowledge representation and reasoning*. Amsterdam: Elsevier.

Calvanese, D., A. Mosca, J. Remesal, M. Rezk, and G. Rull. 2015. A "Historical Case" of ontology-based data access. In *Proceedings of the 2015 Digital Heritage International Congress*, 2: 291–298. IEEE. doi:10.1109/DigitalHeritage.2015.7419510.

Carlisle, P.K., I. Avramides, A. Dalgity, and D. Myers. 2014. The Arches Heritage Inventory and Management System: a standards-based approach to the management of cultural heritage information. In *CIDOC Conference: Access and Understanding – Networking in the Digital Era*. Dresden. Germany.

Ciula, Arianna, and Øyvind Eide. 2014. Reflections on cultural heritage and digital humanities: modelling in practice and theory. In *Proceedings of the first international conference on digital access to textual cultural heritage*, 35–41. DATeCH '14. New York, NY, USA: ACM. doi:10.1145/2595188.2595207.

Davis, Randall, Howard E. Shrobe, and Peter Szolovits. 1993. What is a knowledge representation? *AI Magazine* 14: 17–33.

Doan, AnHai, and Alon Y. Halevy. 2005. Semantic integration research in the database community: a brief survey. *AI Magazine* 26: 83–94.

Doerr, Martin. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine* 24: 75.

———. 2009. Ontologies for cultural heritage. In *Handbook on Ontologies*, 463–486

Doerr, Martin, and N. Crofts. 1998. Electronic communication on diverse data-the role of an object-oriented CIDOC reference model. In *18th General Conference of the International Council of Museums and CIDOC '98*. Melbourne.

Doerr, Martin, and N Crofts. 1999. Electronic esperanto: the role of the object oriented CIDOC reference model. In *Selected papers from ichim99: the International Cultural Heritage Informatics Meeting*. Washington DC.

Doerr, Martin, and Gerald Hiebel. 2013. CRMgeo: Linking the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement. ICS-FORTH.

Doerr, Martin, and Maria Theodoridou. 2014. *CRMdig an extension of CIDOC-CRM to support provenance metadata*. Technical Report 3.2. Heraklion: ICS-FORTH.

Doerr, Martin, Christian-Emil Ore, and Stephen Stead. 2007. The CIDOC conceptual reference model: a new standard for knowledge sharing. In *Tutorials, posters, panels and industrial contributions at the 26th international conference on Conceptual modeling-Volume 83*, 51–56. Australian Computer Society, Inc.

Doerr, Martin, Katerina Tzompanaki, Maria Theodoridou, Christos Georgis, Anastasia Axaridou, and Sven Havemann. 2010. A repository for 3D model production and interpretation in culture and beyond. In *VAST*, 2010:11th.

Doerr, Martin, Athina Kritsotaki, and Katerina Boutsika. 2011. Factual argumentation—a core model for assertions making. *Journal on Computing and Cultural Heritage* 3: 8:1–8:34. doi:10.1145/1921614.1921615.

Doerr, Martin, Ioannis Chrysakis, Anastasia Axaridou, Maria Theodoridou, Christos Georgis, and Emmanuel Maravelakis. 2014. A framework for maintaining provenance information of cultural heritage 3D-models. In *EVA*.

Doerr, Martin, Athina Kritsotaki, Yanis Rousakis, Gerald Hiebel, and Maria Theodoridou. 2015. *Definition of the CRMsci an extension of CIDOC-CRM to support scientific observation*. Technical Report 1.2.3. Heraklion: ICS-FORTH.

Doerr, Martin, Achille Felicetti, Sorin Hermon, Gerald Hiebel, Athina Kritsotaki, Anja Masur, Keith May, et al. 2016. *Definition of the CRMarchaeo: An Extension of CIDOC CRM to support the archaeological excavation process*. Technical Report 1.4. Prato, Italy: PIN S.c.R.L.

Dougherty, J.W.D. 1978. Salience and relativity in classification. *American Ethnologist* 5: 66–80. doi:10.1525/ae.1978.5.1.02a00060.

Falkenberg, Eckhard D., Wolfgang Hesse, Paul Lindgreen, Bjorn E. Nilsson, J.L. Han Oei, Colette Rolland, Ronald K. Stamper, Frans J.M. Van Assche, Alexander A. Verrijn-Stuart, and Klaus Voss. 1998. *A framework of information system concepts*. The FRISCO Report. International Federation for Information Processing.

Gerstl, Peter, and Simone Pribbenow. 1996. A conceptual theory of part-whole relations and its applications. *Data & Knowledge Engineering* 20. Modeling Parts and Wholes: 305–322. doi:10.1016/S0169-023X(96)00014-6.

Ghosh, Pallab. 2015. Google's Vint Cerf warns of "digital Dark Age." BBC News.

Gilchrist, Alan. 2003. Thesauri, taxonomies and ontologies – an etymological note. *Journal of Documentation* 59: 7–18. doi:10.1108/00220410310457984.

Giunchiglia, Fausto, and Pavel Shvaiko. 2003. Semantic matching. *The Knowledge Engineering Review* 18: 265–280. doi:10.1017/S0269888904000074.

Guarino, Nicola. 1995. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies* 43: 625–640. doi:10.1006/ijhc.1995.1066.

———. 1997a. Semantic matching: formal ontological distinctions for information organization, extraction, and integration. In *Information extraction a multidisciplinary approach to an emerging information technology*, 1299:139–170. Berlin, Heidelberg: Springer.

———. 1997b. Understanding and building, using ontologies. *International Journal of Human-Computer Studies* 46: 293–310. doi:10.1006/ijhc.1996.0091.

———. 1998. Formal ontology in information systems. In *Formal ontology in information systems. Proceedings of the First International Conference (FOIS'98)*. Trento, Italy: IOS Press.

Guarino, Nicola, and Christopher Welty. 2000a. Identity, unity, and individuality: towards a formal toolkit for ontological analysis. In *ECAI*, 2000:219–223. Citeseer.

Guarino, Nicola, and Christopher A. Welty. 2000b. A formal ontology of properties. In *Knowledge engineering and knowledge management methods, models, and tools*, 97–112. Juan-les-Pins: Springer. doi:10.1007/3-540-39967-4_8.

Guarino, Nicola, and Christopher Welty. 2002a. Identity and subsumption. In *The Semantics of Relationships*, 111–126. Springer: Netherlands.

Guarino, Nicola, and Christopher A. Welty. 2002b. Evaluating ontological decisions with OntoClean. *Communications of the ACM* 45: 61–65. doi:10.1145/503124.503150.

Guarino, Nicola, Massimiliano Carrara, and Pierdaniele Giaretta. 1994. Formalizing ontological commitment. In *Proceedings of the National Conference on Artificial Intelligence*, 560–567. Morgan Kaufmann.

Hernández, F., L. Rodrigo, J. Contreras, and Francesco Carbone. 2008. Building a cultural heritage ontology for Cantabria. In *Annual conference of the International Documentation Committee of the International Council of Museums (CIDOC) 2008*. Athens, Greece.

Hiebel, Gerald, Martin Doerr, Klaus Hanke, and Anja Masur. 2014. How to put archaeological geometric data into context? Representing mining history research with CIDOC CRM and extensions. *International Journal of Heritage in the Digital Era* 3: 557–577.

Hitzler, Pascal, Markus Krötzsch, Bijan Parsia, Peter F. Patel, and Sebastian Rudolph. 2012. *OWL 2 Primer*.

Hoekstra, Rinke. 2009. Ontology Representation – Design Patterns and Ontologies that Make Sense. 10.3233/978-1-60750-013-1-i, IOS Press 2009.

ISO. 2016. ISO Standards Website. ISO. http://www.iso.org/iso/home/faqs/faqs_standards.htm. Accessed April 15.

ISO 21127:2014 – Information and documentation – a reference ontology for the interchange of cultural heritage information. 2016. http://www.iso.org/iso/catalogue_detail?csnumber=57832. Accessed April 14.

Kontchakov, Roman, Martin Rezk, Mariano Rodriguez-Muro, Guohui Xiao, and Michael Zakharyaschev. 2014. Answering SPARQL Queries over Databases under OWL 2 QL Entailment Regime. In *The Semantic Web – ISWC 2014 13th International Semantic Web Conference*, 552–567. Riva del Garda, Italy: Springer. doi:10.1007/978-3-319-11964-9_35.

Lakoff, George. 1987. *Women, fire, and dangerous things*. University of Chicago Press.

Le Boeuf, Patrick, Doerr, Martin, Ore, Christian Emil, Stead, Stephen. 2016. Definition of the CIDOC Conceptual Reference Model. Technical Report 6.2

Le Goff, Emeline, Olivier Marlet, Xavier Rodier, Stéphane Curet, and Philippe Husi. 2014. Interoperability of the ArSol (Archives du Sol) database based on the CIDOC-CRM ontology. In *CAA2014. 21st century archaeology. concepts, methods and tools. Proceedings of the 42nd annual conference on computer applications and quantitative methods in archaeology*, 179–186. Archaeopress.

Le Rond d'Alembert, Jean, Richard N. Schwab, and Walter E. Rex. 1995. *Preliminary Discourse to the Encyclopedia of Diderot*. University of Chicago Press.

Manola, Frank, Eric Miller, and Brian McBride. 2006. *RDF Primer*.

Y. Marketakis, N. Minadakis, H. Kondylakis, K. Konsolaki, G. Samaritakis, M. Theodoridou, G. Flouris, and M. Doerr. 2016. X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, June 2016. doi: 10.1007/s00799-016-0179-1.

Markhoff, Béatrice Bouchou, Sophie Caratini, Francesco Coreale, Mohamed Lamine Diakité, and Adel Ghamnia. 2015. Semantic Web for BIBLIMOS (position paper). In *Proceedings of the First International Workshop Semantic Web for Scientic Heritage at the 12th ESWC 2015 Conference*.

Mascardi, Viviana, Valentina Cordì, and Paolo Rosso. 2007. A comparison of upper ontologies. In *Dagli Oggetti agli Agenti Agenti e Industria: Applicazioni tecnologiche degli agenti software*, 55–64. Genova, Italy: Seneca Edizioni.

Masur, Anja, Keith May, Gerald Hiebel, Edeltraud Aspöck, and others. 2013. Comparing and mapping archaeological excavation data from different recording systems for integration using ontologies.

Messaoudi, T, Livio De Luca, and P Véron. 2015. Towards an ontology for annotating degradation phe-

nomena. In *Proceedings of the 2015 Digital Heritage International Congress*, 2:379–382. doi:10.1109/DigitalHeritage.2015.7419528.

Moreira, A., L. Alvarenga, and A. de Paiva Oliveira. 2004. Thesaurus and ontology: a study of the definitions found in the computer and information science literature, by means of an analytical-synthetic method. *Knowledge Organization* 31.

Mosca, Alessandro, Jose Remesal, Martin Rezk, and Guillem Rull. 2015. Knowledge Representation in EPNet. In *New Trends in Databases and Information Systems*, 427–437. doi:10.1007/978-3-319-23201-0_43.

National Information Standards Organization. 2004. *Understanding metadata*. National Information Standards Organization.

———. 2005. *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. National Information Standards Organization.

Noy, Natalya Fridman. 2004. Semantic integration: a survey of ontology-based approaches. *Special Interest Group on Management of Data* 33: 65–70. doi:10.1145/1041410.1041421.

Oldman, Dominic, Martin de Doerr, Gerald de Jong, Barry Norton, and Thomas Wikman. 2014. Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning and Aggregation Services for the Digital Humanities (A Position Paper) System. *D-Lib Magazine* 20. doi:10.1045/july2014-oldman.

Pan, Jeff Z. 2009. Resource description framework. In *Handbook on Ontologies*, 71–90.

Papadakis, Manos, Martin Doerr, and Dimitris Plexousakis. 2014. Fuzzy times on space-time volumes. In *eChallenges e-2014 Conference: Belfast, United Kingdom, 29–30 October 2014*. Belfast.

Paveprime Ltd. 2015. *CRMinf: the argumentation model an extension of CIDOC-CRM to support argumentation*. 0.7. Heraklion: FORTH.

Reed, Patricia Ann. 1995. CIDOC relational data model a guide. http://icom.museum/resources/publications-database/publication/cidoc-relational-data-model-a-guide/. Accessed April 14 2017.

Rodriguez-Muro, Mariano, and Martin Rezk. 2015. Efficient SPARQL-to-SQL with R2RML mappings. *Web Semantics Science, Services and Agents on the World Wide Web* 33: 141–169. doi:10.1016/j.websem.2015.03.001.

Ronzino, Paola. 2015. *CIDOC CRMba: a CRM extension for building archaeology information modeling*. Nicosia, Cyprus: The Cyprus Institute.

Ronzino, Paola, Franco Niccolucci, Achille Felicetti, and Martin Doerr. 2016. CRMba a CRM extension for the documentation of standing buildings. *International Journal on Digital Libraries* 17: 71–78.

Rosch, Eleanor, and Barbara B. Lloyd. 1978. *Cognition and categorization*. Lawrence Elbaum Associates.

Scholz, Martin. 2013. A mapping of CIDOC CRM events to German Wordnet for event detection in texts. In *17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013)*. Valetta, Malta: Vladimir Alexiev, Vladimir Ivanov, Maurice Grinberg.

Scholz, Martin, Günther Görz, Günther Görz, and Günther Görz. 2012. WissKI: A Virtual Research Environment for Cultural Heritage.

Smith, Barry. 2006. Against idiosyncrasy in ontology development. *Frontiers in Artificial Intelligence and Applications* 150: 15.

Sowa, John F. 2000. *Knowledge representation: logical, philosophical, and computational foundations*. Brooks/Cole.

Staab, Steffen, and Rudi Studer, eds. 2009. *Handbook on ontologies*. Berlin, Heidelberg: Springer.

Sure, York, Steffen Staab, and Rudi Studer. 2009. Ontology engineering methodology. In *Handbook on ontologies*, 135–152. Springer.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization*. MIT Press.

Szekely, Pedro A., Craig A. Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E. Fink, Rachel Allen, and Georgina Goodlander. 2013. Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In *The Semantic Web: Semantics and Big Data: 10th international conference, ESWC 2013*, 593–607. Montpellier, France: Springer. doi:10.1007/978-3-642-38288-8_40.

The Consultative Committee for Space Data Systems. 2012. *Reference Model for an Open Archival Information System (OAIS)*. Consultative Committee for Space Data Systems Secretariat.

Tzompanaki, Katerina, and Martin Doerr. 2012. A new framework for querying semantic networks. In *Proceedings of Museums and the Web 2012: the international conference for culture and heritage on-line*.

Tzompanaki, Katerina, Martin Doerr, Maria Theodoridou, and Irini Fundulaki. 2013. Reasoning based on property propagation on CIDOC-CRM and CRMdig based repositories. *CRMEX 2013 Practical Experiences with CIDOC CRM and its Extensions*: 37.

UNESCO. 1972. Convention Concerning the Protection of the World Cultural and Natural Heritage. UNESCO.

———. 2005. Convention on the Protection and Promotion of Diversity of Cultural Expressions. UNESCO.

Uschold, M., and R. Jasper. 1999. A framework for understanding and classifying ontology applications. In *KAW*, unknown.

Vállez, Mari, Rafael Pedraza-Jiménez, Lluís Codina, Saúl Blanco, and Cristòfol Rovira. 2015. Updating controlled vocabularies by analysing query logs. *Online Information Review* 39: 870–884. doi:10.1108/OIR-06-2015-0180.

Verborgh, Ruben, and Max De Wilde. 2013. *Using OpenRefine*. Packt Publishing Ltd.

Weingart, S.B. 2013. From trees to Webs: Uprooting knowledge through visualization. In *Classification and visualization: interfaces to knowledge: proceedings of the International UDC Seminar*. The Hague, The Netherlands: Ergon Publishing House.

Welty, Christopher A., and Nicola Guarino. 2001. Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering* 39: 51–74. doi:10.1016/S0169-023X(01)00030-1.

Zúñiga, Gloria L. 2001. Ontology: its transformation from philosophy to information systems. In *Proceedings of the international conference on Formal Ontology in Information Systems–Volume 2001*, 187–197. ACM.