Mapping, Embedding and Extending: Pathways to Semantic Interoperability The Case of Numismatic Collections

Andrea D'Andrea[1] and Franco Niccolucci[2]

[1] CISA, Università di Napoli „L'Orientale" ,
Napoli, Italy
dandrea@unior.it
[1] STARC, The Cyprus Institute ,
Nicosia, Cyprus
f.niccolucci@cyi.ac.cy

Abstract. This paper illustrates current mappings of Cultural Heritage data structures to CIDOC-CRM. After discussing the features of such mappings and evidencing the problems, it illustrates an example concerning coins collections.

Keywords: Ontologies, mapping, semantic interoperability

1 Introduction

Using ontologies to describe the "fundamentals" of knowledge is paramount to semantic interoperability. It does not guarantee it, though, because different sources of information, even if pertaining to the same domain, may obviously have different reference ontologies. To establish a correspondence among such diverse sets of basic concepts is therefore indispensable. This kinds of exercis is known under different names, such as alignment, harmonization, extension, mapping, and so on [1], which usually refer to different ways of pursuing interoperability. In some cases the result is indeed powerful and enriches the knowledge that can be extracted from the joint result. In others, the interoperability obtained in this way is illusory, because only a poor common organization can be superimposed on the joint information.

This may be the case when digital repositories pre-exist, and there is an effort to join them in a digital library. This operation requires a common basis and should result in preserving – as far as possible – the richness of parent repositories also in the sibling. Sometimes this is achieved; sometimes it is impossible; sometimes this issue is overlooked and the quality of the result may be measured only in terms of "digital objects" made jointly available, regardless of the actual utility of having them in the same repository. Often in the latter case Dublin Core is taken as a panacea, a minimal level of interoperabilty for any data collection and a catch-all alibi against objections of limited usability and usefulness.

More effective harmonization efforts in the Cultural Heritage domain have dealt with disparate archives, pushing the search for commonalities as far as possible. It is indeed plausible that information on similar topics and within the same discipline is organized in a very similar way, despite of apparently different organization and denominations deriving from historic, cultural and administrative reasons. CIDOC-CRM, the well-known ISO standard for Cultural Heritage data [2], may be effectively used as common glue for preserving the richness of the original sources.

In this paper we will survey and summarize such harmonization efforts, trying to classify them, showing problems and, hopefully, how to manage them. Work recently undertaken within a EU project on numismatics will be summarily reported as an example of mapping exercises.

2 Mappings to CIDOC-CRM

The potential of CIDOC-CRM in serving as underlying common reasoning system to Cultural Heritage knowledge has been shown in more than one example.

Many mappings to CIDOC-CRM are already available for data encoded according to Dublin Core [3], EAD [4], AMICO [5], TEI [6] and FRBR [7]. Besides, CIDOC-CRM is aligned to DOLCE, a foundational top-level ontology.

Examples of extensions to other domain/task ontologies have also been developed, for example the one for MPEG7 [8], and more recently one for X3D [9].

National standards, for instance the English monument inventory system (MIDAS [10]), and English [11] or Italian [12] archaeological forms, have already been already mapped onto CIDOC-CRM. An on-going investigation aims at integrating the digital library of Perseus Project with Arachne, the central database for archaeological objects of the German Archaeological Institute (DAI) and the University of Cologne (FA): as both archives are encoded according to different information systems, the mapping on CIDOC-CRM is expected to facilitate mediation and integration of the resources [13].

The above quoted examples show different categories of "mappings", using this term in a rather generic way. Some are qualified as harmonization, i.e. aim at reconciling different approaches. Others in fact embed within CIDOC-CRM alien concepts such as three-dimensionality or multimediality, showing how well they may fit in. Others, finally, transfer pre-existing data structures to the new standard and offer a way to overcome limitations, allowing for example the dialogue between museum information – when organized according to CIDOC-CRM – and on-field archaeological investigations, so far documented in a plethora of different ways that CIDOC-CRM might unify.

The process is neither straightforward nor painless, as the examples cited below show. Nevertheless, it is indispensable to proceed in this direction even at the price of slowing down the achievement of ambitious targets – the tenth of millions of digital objects stated in EU policies as a term of reference for digitization initiatives. It is a price, however, that perhaps it is not necessary to pay, thanks to new tools facilitating these mappings (as the AMA tool developed within the EPOCH project [14]) and to the enlarged commitment of the scientific community to develop the mapping first, and then proceed to merge digital repositories.

3 Ontologies and mappings

Different “mapping” exercises may be required to provide interoperability. Therefore, it may be useful to recall the most important concepts about ontologies and the related mapping types.

An *ontology* is a logical theory accounting for the *intended meaning* of a formal vocabulary, i.e. its *ontological commitment* to a particular *conceptualization* of the world.

According to the level of generality, there are different types of ontologies [15]:

- *Top-level ontologies*, describing very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain.
- *Domain ontologies* and *task ontologies*, describing, respectively, the vocabulary related to a generic domain or a generic task or activity, by specializing the terms introduced in the top-level ontology.
- *Application ontologies*, describing concepts depending both on a particular domain and task, which are often specializations of *both* the related ontologies. These concepts often correspond to *roles* played by domain entities while performing a certain activity.

Each domain uses a “local” ontology, producing data according to its own conceptualization. This process produces heterogeneous sources. In order to merge them, it is necessary to do more than a simple labelling mechanism identifying corresponding objects, classes or meanings. In other words, one cannot expect that it is sufficient to recognize that the object *A* in the ontology *X* is the same as the object *B* in the ontology *Y*: were this the case for all concepts, *X* and *Y* would be just the *same* ontology, with objects/classes disguised under different names. Actually it is often the case that concepts do not exactly overlap; that there are multiple correspondences and properties do not correspond; that some concepts of either are more specialized or differentiated than those in the other one. So the process of reconciling inhomogeneous sources specializes in different cases:

- *Extension*: it implies a specialization of the domain ontology, linking some concepts between the two original ontologies. The two conceptual models are in fact complementary, one of them detailing concepts from the other and covering elements or attributes ignored by the second one. This is the case, for instance, of the embedding of the X3D ontology into CIDOC-CRM developed in [9]: in this case it was shown how X3D may be used to extend the concept represented by the CIDOC-CRM element *E36.Visual_item* to detail information about 3D visualization, the scope of X3D but beyond the goal of CIDOC-CRM.
- *Harmonisation*: it implies a semantic equivalence between domain and application ontologies relating to the same ontological commitment. In this case a domain may be regarded as a specialization of the other, which represents a more general or abstract formal level.
- *Alignment*: it implies a generalization of the domain ontology through general axioms and concepts. The two models have (many/a few) general concepts in common.

According to their level of commitment there may be different level of integration (Fig. 1).

- *Extension*: through one or two similar concepts it is possible to set the equivalence between one or two classes in order to specialize the domain ontology with a complementary task-ontology.
- *Harmonization*: two domain-ontologies are harmonised when for a specific commitment it is possible to overlap many concepts or properties mutually.
- *Alignment*: a top-ontology alignments or incorporates a domain ontology through some equivalent and general classes corresponding to the ontological primitiveness, like time, space, and so on.

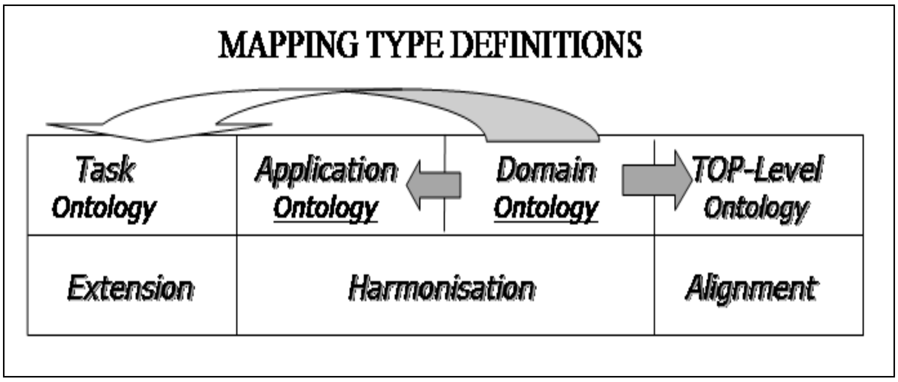

Fig. 1. Mapping types of domain ontology with other types of ontologies.

4 CIDOC-CRM as the Semantic Glue for Cultural Heritage

As far as archaeological documentation is concerned, CIDOC-CRM may indeed be considered as the universal glue to merge heterogeneous sources.

It does not define any of the terms appearing typically as data in the diverse data structures, rather it foresees the characteristic relationships for their use. It does not aim at proposing what cultural institutions should document, rather it explains the logic of what they actually currently document, and thereby enables semantic interoperability.

So far there are several attempts to conform and interpret digital content on the conceptual schema of CIDOC-CRM. However, at a more detailed exam of the different solutions, both from a technical and a theoretical point of view, they appear to adopt substantially different mapping mechanisms. Some solutions, based on an extension of the model, come alongside with integrations relying upon harmonization; in some cases the original model has been mapped on CIDOC-CRM classes; in others the model has been made richer creating new sub-classes and sub-properties. In the former case, it is possible to guarantee a higher compatibility with CIDOC-CRM; in the latter, based on specialization, it is possible to enrich the semantic content at the price of a lower compatibility. In any case, it is a standard that should be adopted since the very beginning of the repository design. But, as the AMA project [14] has shown, excellent results can be obtained also from legacy archives. With this tools, not only different archaeological forms were made compatible, but also such archives were merged with unstructured text descriptions offering a unique global documentation system.

5 Mapping to CIDOC-CRM

In the introduction to CIDOC-CRM handbook its authors point out that "since the intended scope of the CRM is a subset of the "real" world and is therefore potentially infinite, the model has been designed to be extensible through the linkage of compatible external type hierarchies."

To explain correctly the method to be followed for such an extension, they precise that "a sufficient condition for the compatibility of an extension with the CRM is that CRM classes subsume all classes of the extension, and all properties of the extension are either subsumed by CRM properties, or are part of a path for which a CRM property is a shortcut."

In this sense "compatibility of extensions with the CRM means that data structured according to an extension must also remain valid as a CRM instance."

The CIDOC-CRM documentation gives a number of procedures that can be followed to allow that coverage of the intended scope is complete:

- Existing high-level classes can be extended, either structurally as subclasses or dynamically using the type hierarchy.
- Existing high-level properties can be extended, either structurally as sub-properties, or in some cases, dynamically, using properties of properties that allow sub-typing.
- Additional information that falls outside the semantics formally defined by the CRM can be recorded as unstructured data using *E1.CRM_Entity. P3.has_note: E62.String*.

Using mechanisms 1 or 2, the CRM concepts subsume and thereby cover the extensions. On the contrary, in mechanism 3, the information is accessible at the appropriate point in the respective knowledge base. CRM authors highlight that "this approach is preferable when detailed, targeted queries are not expected; in general, only those concepts used for formal querying need to be explicitly modelled: in some ontologies entities can have the same name, but different meaning or, vice versa, have different labels but refer to the same concept; in some cases an entity can be complementary to another or partially combine".

Starting and target structures may not correspond exactly: in this case a new (source) model fitting with the CIDOC-CRM hierarchy will be created, by making explicit some concepts that are implicit in the source, in order to parallel the axioms/paths characterizing CIDOC-CRM structure and hierarchy. This is a sort of anisomorphic process modifying the apparent structure of the starting source.

A typical case is the *event*, which is central to the structure of CIDOC-CRM but often is ignored in other models. However, also in these models there is some event determining a change in the state of the object being documented, be it the creation, destruction or any other state change. To map correctly this model to the CRM, such events may be stated explicitly.

There may be many other cases of anisomorphism, for example when a unit (grams, centimetres, etc.) must be specified or a disambiguation is needed. We will show in section 7 other examples of this situation.

This shows that mapping is neither an easy matter nor a linear process and requires disciplinary competence as well as deep understanding of the implicit assumptions of the source model.

In conclusion, the process of creating the mapping among heterogeneous sources and different schemas may change in case of multiple conditions [KDP06]. Some basic rules have been identified:

- Introducing an intermediate node, in order to precise the whole path for the mapping of the source path. The most common such addition is an event.
- Contracting several classes, e.g. in address (names, coordinates), in one entity.
- Making explicit the causal connection between some related classes in order to allow the interoperability between information from other sources.
- As regards the previous rule, two relations in the source schema has to be mapped with the same intermediate node.
- Some elements may appear in multiple mappings.

The next paragraphs describe how mapping to CIDOC-CRM has been implemented as a data transfer mechanism. Many of the examples come from work made some years ago and rely upon an old CIDOC-CRM version, so they would need updating. Nevertheless they are an attempt to point out the overlapping between various domain-ontologies having in common CIDOC-CRM as an inter-lingua.

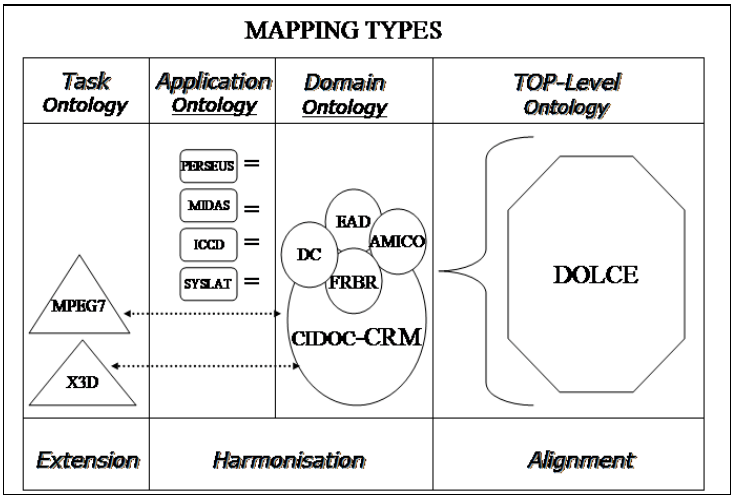

Fig. 2. Existing mappings to CIDOC-CRM.

Unfortunately, such experiments have not received so far sufficient attention by practitioners and have had no enough impact on current practice. They can however be regarded as experimental evidence of the fitness of CIDOC-ROM to its intended goal. Fig. 2 represents the linkages between CIDOC-CRM on one side and other ontologies on the other.

5.1 Harmonising *domain ontologies*: AMICO, DC, EAD, FRBR, TEI and mappings to CIDOC-CRM

AMICO, DC, EAD, FRBR and TEI deal with various domain ontologies reflecting different commitments not always completely covering the Cultural Heritage domain. In same cases the coverage is only partial, while in other cases the model can integrate and extend CIDOC-CRM original scope.

These models aim at:

- Managing standard for museum multimedia (AMICO – the standard is no more supported since 2005);
- Structuring metadata for cross-domain information resource such as video, sound, image, text, and composite media like web pages (DC);
- Designing intellectual or physical parts of archival finding aids (EAD) such as inventories, registers, indexes, and other documents created by archives, libraries, museums, and manuscript repositories (EAD);
- Capturing and representing the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information (FRBR);
- Implementing an interdisciplinary standard in order to enable libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation (TEI).

All these standards describe a structured record using elements, but any kind of relation or property is expressed between attributes or records; moreover discrete entities are not explicitly identified. In order to create a semantic equivalence the existence of common classes with the same meaning has been recognized and highlighted. Besides, according to CIDOC-CRM paths, one or more intermediate nodes need to be added to make the source structure compatible with the target one.

In these cases, beyond the specification of equivalent entities, a structure has been introduced in the mapping process; only adding these paths it has been possible to extract the implicit knowledge embedded in the sources.

5.2 Extending/Specializing CIDOC-CRM with complementary *application ontologies* as X3D and MPEG7

X3D and MPEG7 are both task ontologies adopted in order within specific domains: 3D content description the former, and latter multimedia the latter.

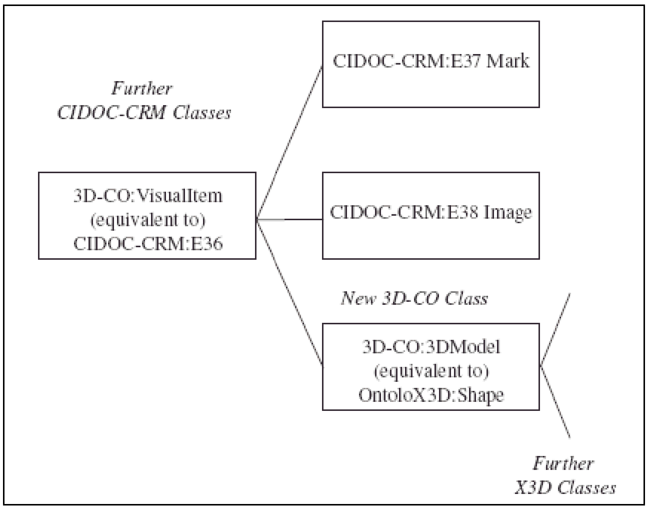

Fig. 3. Diagram of the 3D model extension (from [9] fig.3).

As such specialized aspects are out the scope of CIDOC-CRM, the extension has been realized linking two corresponding classes as shown in figure 3 above. In those cases models remain separated and are joined to each other only through a node, whose task is to specialize a general concept of CIDOC-CRM; in this way it isn't necessary to transform the structure of the source adjusting it to the CIDOC-CRM model. No intermediate node has therefore been added to the original ontologies.

5.2 Mapping *task ontologies* to CIDOC-CRM: MIDAS, English Heritage, ICCD, PERSEUS

Some task ontologies cover fully the CIDOC-CRM scope. They deal with local national standard implemented for managing specific tasks refer to cataloguing, inventorying and documenting archaeological data.

They arose above all from various attempts made at normalizing the production of archaeological documents, mainly coming out from excavations; in order to reduce the subjectivity in the recording process the normalization have lead to the elaboration of a large number of forms. These recommendations or best practices have been translated in tables and relational databases. Several information systems, implemented for the managing of the archaeological forms, are already mapped on CIDOC-CRM.

In such mappings, the original sources, containing data from an autonomous domain or sub-domain within Cultural Heritage, are read and re-interpreted in the light of CIDOC-CRM philosophy, adding only the nodes that are necessary to link entities and properties to events.

Considering some of these mapping procedures, it may be noticed that there are alternative ways of representing the same conceptual archiving practice. While the English Heritage mapping chose to base on the creation of new sub-classes of "IsA" type specializing the original CIDOC-CRM and making it richer, the Italian ICCD mapping preferred to maintain a full compatibility with CIDOC-CRM, fitting the starting source with the destination ontology only through the semantic equivalence between corresponding classes.

Actually, from a theoretical and methodological point of view, both these mechanisms are formally correct, although they may be mutually integrated only at a general level. Perhaps, it would be appropriate to reach a consensus on the preferred way of performing this mapping exercise.

6 Tools for mapping CIDOC-CRM

As already mentioned, the EPOCH project has produced a mapping tool, which is being maintained and improved even after the conclusion of the project. It is denominated AMA (Archaeological Mapping Tool) [14], an acronym that hints to the reconciliation of different sources and, hopefully, to mutual understanding and "love" (*amor*, in Latin). The tool may in fact be used for any kind of data structures, not only for archaeological ones. AMA does not substitute human intelligence in the identification of corresponding concepts and relations. It simplifies the registration of such identification and facilitates the task of transcribing the correspondences. A proposed standardized way of registering the mapping is under way following the indications appeared in [16].

AMA is accessible on-line from the EPOCH web site and is available under a Creative Commons license.

7 The COINS mapping: work in progress and lessons learnt

COINS, another EU-funded project [17], is facing the task of reconciling different data organizations in the same domain, in this case numismatics. It was expected that this goal could be obtained with no great difficulty as numismatists have already achieved a high degree of homogeneity and standardization in their descriptions. The case studies are the archives of the numismatic collection at the Cambridge Fitzwilliam Museum, the coins collection of MNIR, the Romanian National Museum in Bucharest, and the huge numismatic collection of the Museum of Palazzo Massimo in Rome of the Soprintendenza Archeologica of Rome (SAR). These digital archives have been chosen as representatives of three different traditions, also geographically, which have nevertheless a long practice of scientific collaboration. Previous work on a similar subject was performed in [18], with a different approach.

Actually the data organization was similar, but differences were obviously present in the three digital repositories.

The coin mapping is a good example for concepts enounced in section 3. They show the need of expanding the data structure to take into account implicit concepts, which are necessary to a correct mapping.

In this case the two relevant events are the "coinage" and the "cataloguing". All the concepts may be related to one of these two events, implementing the event-based approach of CIDOC-CRM. The insertion of the event “coinage” supersedes the slight differences between data structures, which become fully compatible with each other. For example:

Fitzwiliam	CIDOC-CRM	SAR
Maker_mint	Coinage	Zecca
Maker_place		Zecca-Luogo
Maker_State		Zecchiere_nome
Maker		Zecchiere_luogo
Maker_role		Stato
Maker_qualifier		Autorità
		Tipo_autorità

A more complex example refers to the authority minting the coin, usually managed as

AUTHORITY = *value*

for example

AUTHORITY = ATHENAI

meaning that the coin was minted by the town-state of Athens.

To map these concepts on CIDOC-CRM it is necessary to expand this relationship as follows:

P11F.has_participant — E40.Legal_Body (Authority) — P131F.is_identified_by — E82.Actor_Appellation (Athenai)

Even more complex is the apparently simple case of the weight, which is usually recorded as a figure without the indication of the unit, implicit in the source model and possibly, but not regularly, referred in some documentation or a standard one – unlikely to be for coins, which are never weighted in Kg.

P43F.has_dimension — E54.Dimension (Weight) — P90F.has_value — E60.Number — P91F.has_unit — E58.Measurement_Unit (gr)

In this case the field AUTHORITY is the instance of Legal_Body, while the value is the instance of the class E82. So the mapping has to include sometimes not only the value of a specific field but also the same fielding order to create a correct and understandable path.

It would probably be too long to enter here into other examples of further details. Mappings are available as project deliverables on the project web site.

The idea is to abstract from individual mappings – having, as already mentioned, slight differences from each other – to arrive at the construction of a general numismatic reference model, fully compatible with CIDOC-CRM. The latter was chosen as universal glue for the systems, each of which is mapped to it. Such mappings allow semantic interoperability preserving all the richness of original sources. The project has also developed a collection management system enabling semantic searches, aptly named MAD (Management of Archaeological Data).

From the mappings the project team is progressively reconstructing the underlying ontology. Comparisons with other digital repositories and discussing intermediate results with the users' community are planned to extend the validity of this inductive approach.

8 Conclusions

Undertaking a standardization process involving archaeologists and archaeological data may perhaps be considered as a symptom of naivety. Few scientific communities are more individualistic than this, the result being an extreme fragmentation of systems and data models. It is always the case that a minute detail, ignored by the majority, is vital to some individual researcher, two such details never being the same for different investigators. In the end, all these small differences might lead to a paradoxical situation of many different data models that differ very little from each other when considered in couples, but have a rather limited intersection when taken all together.

Mapping to CIDOC-CRM is probably the only way to reconcile such disparate approaches, providing a very large common conceptual area but safeguarding the interests of individual researchers to personalize what they consider relevant information.

The above examples show the complexity and difficulty of the task of reconciling different sources of information and data models. In general, just pairing concepts does not suffice, but further source interpretation, disambiguation and resolution of co-references is required. On this regard, harmonization and alignment result simpler, because their scope is high-level reconciling. Being based on general and abstract concepts, it is usually easier to find correspondences and commonalities. Difficulties come when mapping is undertaken, as it has to cope with reality and practice. Here differences and lack of homogeneity are commonplace, and reconciling different archives may require much more effort and ingenuity.

At least, it is hoped that the above examples show that apparently easy shortcuts (AKA crosswalks) usually are trivialization of mappings, and often, rather than saving time, lead to an undesired destination. Only concrete practice of real archive mapping will eventually show the superiority of CIDOC-CRM as a global schema for the organization of Cultural Heritage knowledge and information.

References

1 Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18(1), 1-31.

2 Crofts N., Doerr M., Gill T., Stead S., Stiff M.: *CIDOC CRM Special Interest Group*, Version 4.2, June 2005. http://cidoc.ics.forth.gr/docs/cidoc_crm_version _4.2. pdf.
3 Kakali C., Lourdi I., Stasinopoulou T., Bountouri L., Papatheodorou Ch., Doerr M., Gergatsoulis M. In Integrating Dublin Core metadata for cultural heritage collections using ontologies, 2007 Proc. Int'l Conf. on Dublin Core and Metadata Applications.
4 Theodoridou M., Doerr M.: Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM. *Technical Report FORTH-ICS/TR-289*, June 2001. http://www.ics.forth.gr/proj/isst/Publications/paperlink/mappingamicotocrm.pdf.
5 Doerr M.: Mapping of the AMICO data dictionary to the CIDOC CRM. *Technical Report FORTH-ICS/TR-288*, June 2000. http://www.ics.forth.gr/proj/isst/Publications/paperlink/mappingamicotocrm.pdf.
6 Eide O., Ore C.E.: TEI, CIDOC-CRM and a Possible Interface between the Two. http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/ore.pdf.
7 Doerr M., LeBoeuf P., Modelling Intellectual Processes: The FRBR - CRM Harmonization. http://cidoc.ics.forth.gr/docs/doer_le_boeuf.pdf.
8 Hunter J.: Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums, In *Museums on the Web 2002*, Boston, April 2002. http://www.metadata.net/harmony/MW2002_paper.pdf.
9 Niccolucci, F., D'Andrea, A.: An Ontology for 3D Cultural Objects. in Ioannides, M., Arnold, D., Niccolucci, F., Mania F. (Eds), The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST2006, 203-210.
10 Crofts N.: MDA Spectrum CIDOC CRM mapping. http://cidoc.ics.forth.gr/docs/MDA%20Spectrum_CIDOC_CRM_mapping.pdf.
11 Cripps P., Greenhalgh A., Fellows D., May K., Robinson D.: Ontological Modelling of the work of the Centre for Archaeology, September 2004. http://cidoc.ics.forth.gr/docs/Ontological_Modelling_Project_Report_%20Sep2004 .pdf.
12 D'Andrea, A., Marchese G., Zoppi T.: Ontological Modelling for Archaeological Data. in Ioannides, M., Arnold, D., Niccolucci, F., Mania F. (Eds), The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST2006, 211-218.
13 Kummer R.: Integrating data from The Perseus Project and Arachne using the CIDOC CRM. http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/kummer.pdf.
14 Felicetti A.: MAD – Management of Archaeological Data. In M. Ioannides, D. Arnold, F. Niccolucci, K. Mania (eds.) *The e-volution of Information Communication Technology in Cultural Heritage – Project papers*. Budapest, Archaeolingua, 2006, 124 – 131.
15 Guarino N.: Formal Ontology in Information Systems. In *Proceedings of FOIS'98,* Trento, Italy, 6-8 June 1998, Amsterdam, IOS Press, 3-15.
16 Kondylakis H., Doerr M., Plexousakis D.: Mapping Language for Information Integration. *Technical Report 385, ICS-FORTH*, December 2006. http://cidoc.ics.forth.gr/docs/Mapping_TR385_December06.pdf.
17 http://www.coins-project.eu/.

18 Nussbaumer P. Haslhofer B.: CIDOC-CRM in Action – Experiences and Challenges. ECDL 2007: 532-533.