SCIRES it
SCIentific RESearch and Information Technology
Ricerca Scientifica e Tecnologie dell'Informazione
Vol 7, Issue 1 (2017), 1-8
e-ISSN 2239-4303, DOI 10.2423/i22394303v7n1p1
© CASPUR-CIBER Publishing, http://caspur-ciberpublishing.it

MAPPING THE ARIADNE CATALOGUE DATA MODEL TO CIDOC CRM: BRIDGING RESOURCE DISCOVERY AND ITEM-LEVEL ACCESS

Nicola Aloia, Franca Debole*, Achille Felicetti**, Ilenia Galluccio**, Maria Theodoridou****

*Istituto di Scienza e Tecnologie dell'Informazione, National Research Council – Pisa, Italy
**PIN Scrl – Prato, Italy
***FORTH – Heraklion, Greece

Abstract

ARIADNE is a European project aiming to integrate existing archaeological research infrastructures, services and distributed datasets, and to develop new technologies and tools to improve archaeological research methodology. The ARIADNE registry contains information about resources available among the various partners of the project and the metadata repository, which contains item level information of these resources. In order to provide an advanced discovery mechanism combining both item level and registry level information we propose a mapping from the ARIADNE Catalogue Data Model, the model of the ARIADNE registry, to the CIDOC CRM, the underlying model of the metadata repository. The paper will present the requirements that led to the choice of different models for the registry and the metadata repository, will elaborate on the mapping, and will propose an integrated interface for information discovery and presentation.

Keywords

Archaeological Infrastructures, Mapping, CIDOC CRM, Registries, Data Models

1.Introduction

Data are one of the greatest resources of modern times, around which many activities are focused both in research and in business. Integration, interconnection, reuse, interoperability are all keywords of the current web-based applications. Governmental and industrial organizations are committed to ensure that this new “raw material” becomes a source of new developments and wealth. The idea that certain data should be freely available (Open Data) is becoming ever more widespread. Behind these thrusts and to help achieve these goals, there are many efforts in the definition and implementation of several tools, ranging from standard ontologies, vocabularies, authority files, languages, triple stores, etc. In recent years, several projects, funded by the European Community, aimed at integrating data and services, available in a variety of contexts. In the area of cultural heritage for example, EUROPEANA[1], is one of the largest efforts with such a goal. ARIADNE[2] (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) is an FP7-INFRASTRUCTURES-2012-1 EU project that aims to integrate the existing archaeological research infrastructures so that researchers can use the various distributed datasets and services together with new and powerful technologies as an integral component of the archaeological research methodology. The ambitious goal raises various challenges at different levels: technological, methodological and organizational. In this paper, we will cover only certain aspects of integration issues at the technological level. There is a large availability of archaeological digital datasets, which differ in structure, aims and functionalities offered, outcome of the research of individuals, teams and institutions; that span different periods, domains and regions; more are continuously created as a result of the increasing use of Information Technology. One of the most important issues concerning the integration of the assets of archaeological data regards their representation through a data model that allows its usability and interoperability. The results so far obtained in the various activities we have mentioned provide us with good points to start

[1] http://www.europeana.eu/
[2] http://www.ariadne-infrastructure.eu/

with. The integration of archaeological datasets and services available among providers is one of the most important activities of the ARIADNE project. The architecture of ARIADNE provides a good example of how archaeology can adapt the FAIR principles of scientific data - Findability, Accessibility, Interoperability, and Re-usability.

CIDOC CRM[3] (Doerr, 2003) is the reference ontology chosen by ARIADNE for the integration and interoperability of datasets. In parallel, ARIADNE implements a registry that provides details about datasets and services available. To collect information about the digital datasets chosen for integration, the ARIADNE Dataset Catalogue Model (ACDM for short) has been defined. ACDM is an extension of the Data Catalogue Vocabulary (DCAT)[4], a recommendation of the W3C Consortium that *"is well-suited to representing government data catalogues such as Data.gov and data.gov.uk*." The reason for adopting the DCAT Vocabulary (apart from reuse) is that DCAT is proposed as a tool for publishing datasets as Open Data. During the project, the architecture of the platform and the role of the registry have been defined. The ARIADNE aggregation infrastructure consists of the ARIADNE registry, which contains information about resources available among the various partners of the project and the metadata repository, which contains item level information of these resources. Advanced query and browse capabilities allow the exploitation of the information available in the metadata repository and the registry.

In this paper, we propose an augmented discovery mechanism based on the integration of both item and catalogue level information. Integration is achieved through the aggregation of the registry information into the metadata repository. The ARIADNE Catalogue Data Model is mapped to CIDOC CRM, the underlying model of the metadata repository, and the catalogue records are subsequently transformed into CIDOC CRM compatible records, enhancing the metadata repository with complementary, semantic information. The proposed mapping aims at moving a step further towards a deeper standardization having a single ontology for describing datasets at every level of detail, ranging from general catalogue information to detailed item descriptions.

2. The ARIADNE Catalogue Data Model

As mentioned in the introduction, the ARIADNE project aims to integrate the various data of existing archaeological research, to supply powerful techniques to the scientist for the fruition of data and services that the community makes available. The catalogue of ARIADNE lists and describes what is available from the project partners, and more generally the whole community of archaeologists, to identify, through refined search mechanisms, the resources candidates for integration. In this section, we present the data model of the ARIADNE Catalogue Data Model (ACDM), which describes the available resources among the various partners of the project. The definition of the model and consequently the implementation of the derived tools has gone through several revisions, as a result of new requirements and knowledge gained during the development of the project. For a detailed and updated ACDM please refer to the official documentation on the ARIADNE site.[5] The central notion of the model is the *ArchaeologicalResource* class, that has as instances the main resources categorized in:

- *Services*, representing the services owned by the ARIADNE partners and lent to the project for discovery and access;
- *Language resources*, this is the class of all language resources described in the Catalogue for the purposes of reuse or integration within the ARIADNE community. A language resource is a resource of a linguistic nature, whether in natural language (such as a gazetteer) or in a formal language (such as a vocabulary or a metadata schema). It also includes mappings, understood as associations between expressions of two language resources.
- *Data resources*, representing the various types of data containers owned by the ARIADNE partners and lent to the project for discovery, access and possibly integration. Data resources are categorized in collections, datasets, databases and GIS.

While for the description of the ARIADNE Language and Data Resource class it was possible

[3] http://www.cidoc-crm.org/
[4] https://www.w3.org/TR/vocab-dcat/
[5] http://support.ariadne-infrastructure.eu/

to adopt a standard vocabulary (DCAT), as regards the description of the services to be surveyed for ARIADNE, the situation is somewhat more complicated, due to the fact that there exist several vocabularies, none of which stands out as a *de facto* standard. Deliverable D13.2[6] of the ARIADNE project provides an exhaustive list of the services that will be made available through the ARIADNE infrastructure. Based on the evidence collected in this Deliverable, we classify services in the following categories, reflecting the way a service is accessed:

- *stand-alone services*, tools to be downloaded and installed on one machine;
- *web services*, Web accessible services with an API. The services developed by the ARIADNE project fall in this category;
- *services for humans*, Web accessible services with a GUI only;
- *institutional services*, services offered by some institution and that must be negotiated via a personal interaction with representatives of that institution in order to be accessed.

Correspondingly, we introduce in the model a generic Service class, which is abstract and gathers the properties in common to all services, and four sub-classes of its, one-to-one with the above four categories.

3. CIDOC CRM

The CIDOC CRM ontology has been chosen by ARIADNE to foster datasets integration and interoperability. CIDOC CRM is an international ISO Standard (ISO21127:2014), released by the International Council of Museums (ICOM). Over the last two decades the model has been developed by interdisciplinary working groups, the CIDOC Documentation Standards Working Group and the CIDOC CRM Special Interest Group, chaired by Dr. Martin Doerr. The formal ontology has been designed to represent the complexity and multiformity of the Cultural Heritage world, in order to allow information exchange between heterogeneous sources. Composed of 94 classes and 168 properties, the ontology describes, through a logical paradigm, concepts and relationships inherently enclosed in the CH documentation. Over the last years of research the CIDOC CRM project expanded its vision, becoming a modular model. The CRM family, as of today, includes a collection of models (extensions) specifically developed to encompass the specific needs of various domains. As a natural evolution to the enrichment of the general documentation model, the efforts of the working group aim at the development of specialized thematic models for the requirements of specific projects. The ARIADNE Reference Model (Fig. 1) is based on CIDOC CRM and a suite of extensions suitable to address the complexity of archaeological data integration.

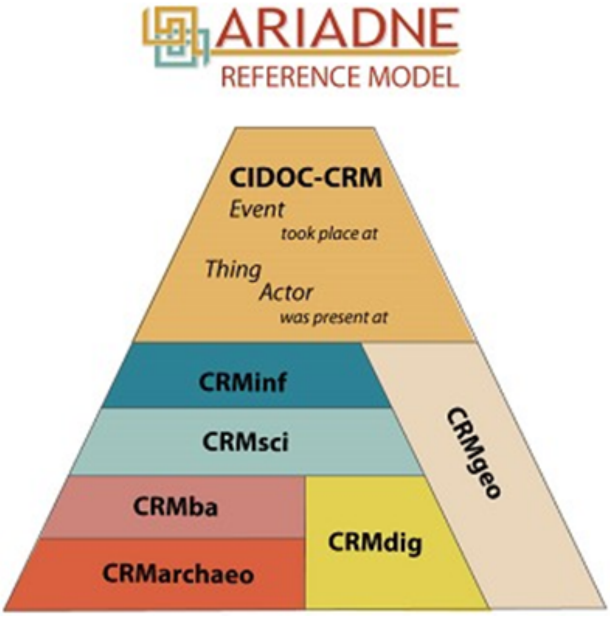

Fig. 1: ARIADNE Reference Model

The extension CRM*archaeo*[7] (Doerr, 2016) has been defined and implemented in the context of ARIADNE. The definition of CRMarchaeo is an important outcome to be highlighted and a remarkable number of institutions are now beginning to use it (e.g. Data Archiving and Networked Services, Deutsches Archäologisches Institut, Archaeology Data Service, Istituto Centrale per il Catalogo e la Documentazione. etc.).

CRM*archaeo* has been created to support the archaeological excavation process and to provide an instrument to manage and integrate existing archaeological documentation. The main goal is the formalization of the heterogeneous knowledge produced by archaeologists, often recorded in different standards. CRM*archaeo* has been further extended within the framework of

[6] http://www.ariadne-infrastructure.eu/Resources

[7] http://new.cidoc-crm.org/crmarchaeo/fm_releases

the ARIADNE project, in which CIDOC CRM encoding of archaeological datasets has played the role of the "Electronic Esperanto" (Doerr, 1999), ensuring a deeper standardization, looking towards the perspective of sustainability.

4. Mapping ACDM to CIDOC CRM

The ACDM harmonization with CIDOC CRM could allow the reuse in the broader context of digital humanities, to support data transformation, merging and to achieve semantic interoperability between different models. The process of mapping the schema of an archaeological database to a common ontology is not trivial and needs support from appropriate tools. First, we need a sufficient mapping specification to support the transformation of the ARIADNE Data model (source schema) into the CIDOC CRM (target schema). It is crucial that during the transformation, the information encoded in the source schema should be maintained; the initial meaning should be preserved as much as possible. In the ARIADNE project the mapping process was assisted by the X3ML mapping framework developed by FORTH (Marketakis, 2016), and it was achieved by the close operation of the CIDOC CRM experts. The X3ML mapping framework includes the X3ML mapping definition language, the 3M Mapping Memory Manager[8], the 3M Editor and the X3ML engine. The X3ML framework takes a completely different approach to other data mapping tools. Designed to separate out many of the technical aspects of creating Linked Data, it allows data experts to play a larger role in Linked Data generation, creating better end results relevant for larger and wider audiences. The 3M Editor is specifically designed to support mapping to richer ontologies and CIDOC Conceptual Reference Model (Doerr, 2015). The 3M Editor provides a simple user interface where the main mapping screen concentrates simply on mapping each element of a source schema to an appropriate sequence (path) of CIDOC CRM relationships and entities. The specification of the URI (addresses) is a separate process which augments the X3ML mapping definition file with instance generation functions. X3ML provides a clear, human readable format for defining schema mappings. Our goal is to interpret the ACDM as a semantic model (*domain-property-range*) and we achieve it through the transformation of the fields of the ACDM schema into equivalent CIDOC CRM paths. Our mapping starts with the specification of the domain, which defines that instances of the ArchaeologicalResource correspond to instances of the E73_*Information_Object* class of CIDOC CRM.

CIDOC CRM is an event-oriented ontology, for this reason, in order to express complex concepts and extract more knowledge, sometimes, the mapping requires articulate paths, the so-called "intermediate node".

In some cases the mapping is an easy one-to-one mapping, as shown in Fig. 2. In this example the *dcterms: title* field is the source property mapped to the target P102_*has_title* property and the source range mapped to the target range E35_*Title* class.

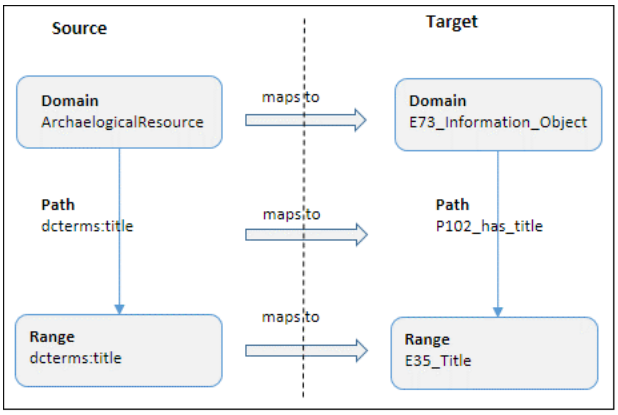

Fig. 2: One to one mapping

In most cases it was impossible to have a one-to-one mapping and Fig. 3 highligths one of such cases, in which one field of the source has been mapped to a complex path in the target. Indeed, in Fig. 3, the *acdm: creator* field of the ArchaeologicalResource entity cannot be mapped to a single CIDOC CRM path.

Within the *acdm: creator* concept, the event of "Creation" is the hidden information. The simple source path "ArchaeologicalResource has a Creator" maps to a complex path:
E73_*Information_Object* → P94i_*was_created by*→ E65_*Creation* → P14_*carried_out_by* → E39_*Actor.*
E65_*Creation* represents the "intermediate node" in the X3ML terminology. This kind of construct provides flexibility during the mapping process and allows the encoding of tacit information in the source.

[8] http://www.ics.forth.gr/isl/3M

Following this kind of methodology, we mapped the ARIADNE registry i.e. each entity of the ACDM schema, to appropriate CIDOC CRM paths. The mapping (ACDM mapping no. 508) is available in the 3M tool.

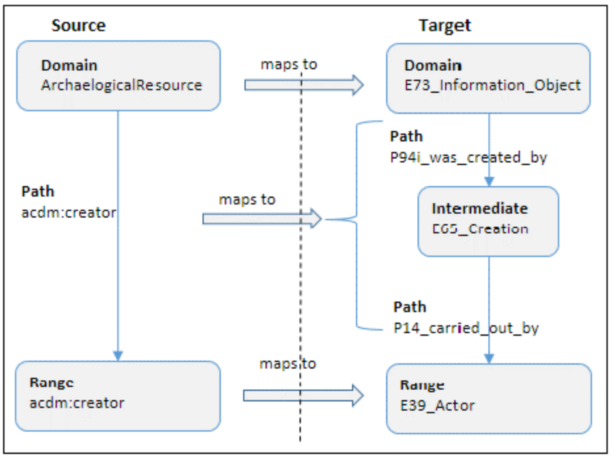

Fig. 3: Complex path

5. The ARIADNE Metadata Repository

Ensuring well-structured data and being compliant with international standards, retrieval information becomes possible. Metadata schemata, rich in semantics, are the best way to expose information about Cultural Heritage collections. There are several metadata standards, chosen by cultural institutions (Caplan, 2003): Dublin Core[9] is the most used, other schemata such LIDO[10], CARARE[11] have been developed within EUROPEANA framework (Ronzino, 2012). One of ARIADNE's goals was to demonstrate that different archaeological databases of specific content can be integrated under a coherent, general schema producing a network of rich, structured and interrelated information. The ARIADNE Metadata Repository is the integrated semantic network, an aggregation of the data produced through the process of mapping and transformation of each provider's source database to the common target ARIADNE Reference Model (CIDOC CRM and extension suite). It stores the uniform, consistent representation of data and supports the query and search mechanisms of the aggregated collection. In order to create the aggregated collection in the metadata repository, we proceeded mapping the schemata of individual datasets to CIDOC CRM. A specific use case of mapping the schemata of several coin databases is described in details (Felicetti, 2015). Each dataset was analyzed by a team of domain and IT experts and an appropriate mapping to CIDOC CRM was defined and materialized with the use of 3M editor and the X3ML mapping definition language. A URI genera- tion policy common to all datasets was decided and the transformation of each dataset to the common representation was performed with the use of the X3ML engine. In parallel terminology mappings were also defined and implemented: Nomisma.org[12] was selected as the common vocabulary across the different coin datasets. The whole process has been repeated several times, for optimal calibration of the mappings and normalization of the results. The result of all the above steps was the generation of a set of RDF statements describing the item level information for each original dataset. These sets were loaded into the ARIADNE integrated semantic metadata repository, implemented in blazegraph[13].

Furthermore, having mapped the ACDM to CIDOC CRM, as described in the previous sections, the Registry records can also be transformed to the common representation and thus the Catalogue information can be integrated into the metadata repository allowing for querying the data at all levels. The steps of the aggregation workflow are presented in Fig. 4.

The ARIADNE integrated semantic metadata repository can support a variety of research questions such as:

- Origin - Where does this coin come from?
- Tracking - How did it arrive here?
- Chronology - First/last appearance
- Practical/symbolic value, incidents - Why is it deposited here?
- Political message - Why was it produced (i.e. "minted")?
- Economic stability, power - Why was it widely used / not used?
- Statistics - Material versus nominal value

[9] http://dublincore.org/
[10] http://www.lido-schema.org/
[11] http://pro.carare.eu/doku.php?id=support:metadata-schema
[12] http://nomisma.org/
[13] https://www.blazegraph.com/

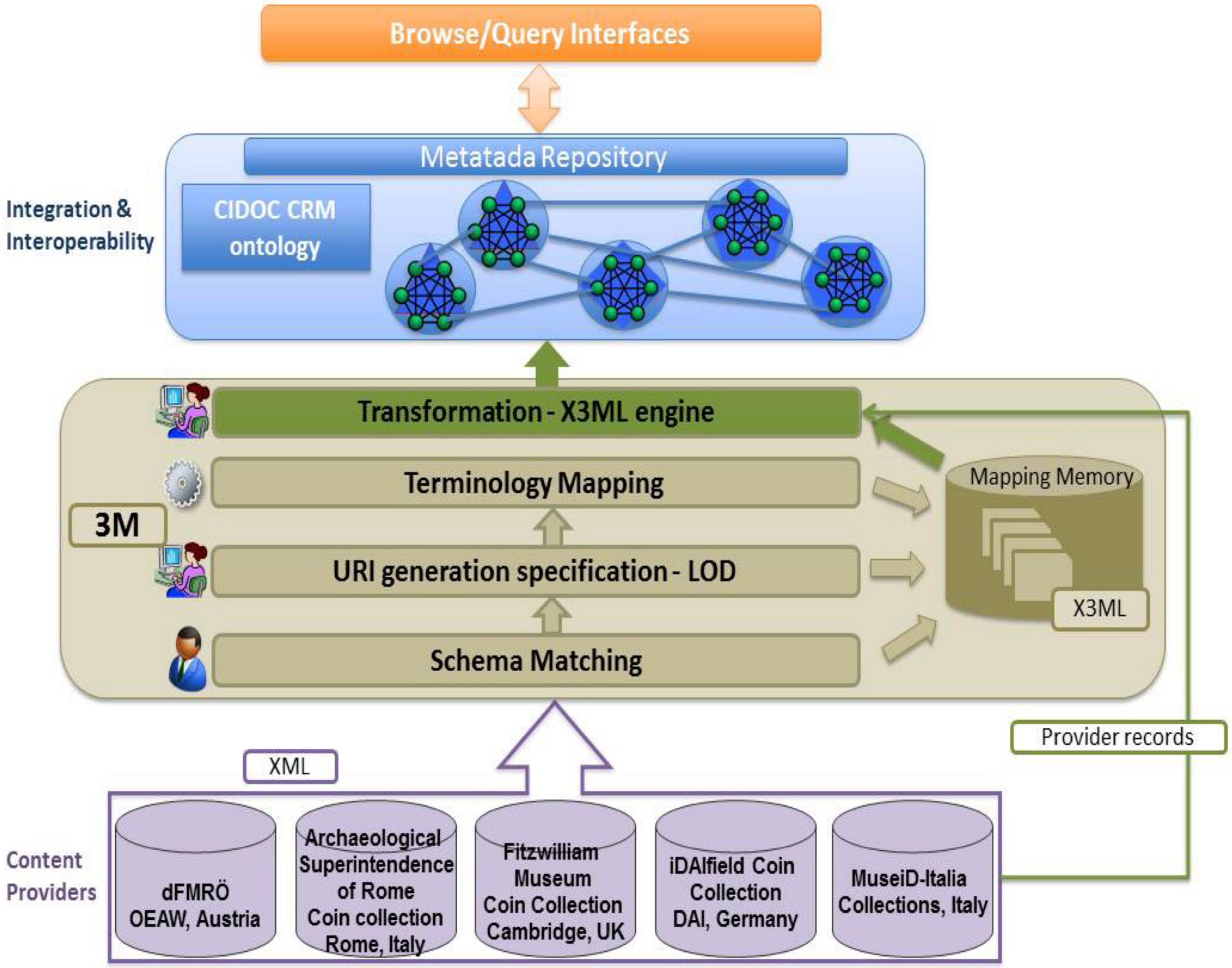
Browse/Query Interfaces
Metatada Repository
Integration & Interoperability
CIDOC CRM ontology
Transformation - X3ML engine
Terminology Mapping
3M
URI generation specification - LOD
Schema Matching
Mapping Memory
X3ML
Provider records
XML
Content Providers
dFMRÖ OEAW, Austria
Archaeological Superintendence of Rome Coin collection Rome, Italy
Fitzwilliam Museum Coin Collection Cambridge, UK
iDAIfield Coin Collection DAI, Germany
MuseiD-Italia Collections, Italy

Fig. 4: Aggregation Workflow

There exist several queries that are trivial to be answered by each dataset separately; however, they become important if they can be answered by the aggregated repository:

- Find coins minted in the same place/area or by the same authority
- Find coins produced in the same period or time span (typically the same century or half/quarter century)
- Find coins having common shape/iconography/inscriptions
- Find coins made by a specific material.

The ARIADNE Catalogue supports querying according to subject, space, time, actors and keywords. The ARIADNE portal provides a user interface to query the catalogue by using maps, timelines and faceted browsing.

The two types of searching/querying can be further enhanced through the integration of the catalogue records into the metadata repository.

More complex queries involving catalogue information and item semantic information can be specified:

- Find all the European datasets containing information on roman coins in Britain, freely accessible according with an open access policy
- Find all the coordinates of archaeological sites containing bronze coins, falling within the data range “2600BC-700BC” and show them on a map
- Build a timeline of chamber tombs construction events in Europe during the Bronze Age in order to track a trend of diffusion of this construction technique.

6. Conclusions

In this paper we presented the mapping of ACDM to CIDOC CRM and the integration of the catalogue information with the item level data information in the context of the ARIADNE project. We presented the repository and service architecture of ARIADNE as well as the two models used for the catalogue and the metadata repository. The mapping of ACDM to CIDOC CRM was achieved by the close cooperation of the domain experts with the IT experts, it was assisted by the X3ML mapping framework and it is presented with a few representative examples.

We also described the aggregation of data from several providers to the common ARIADNE integrated metadata repository. Finally, we presented the integration of the metadata repository records with the catalogue records and we described potential queries that will be supported.

References

Caplan., P. (2003). *Metadata Fundamentals for All Librarians*. Chicago, IL: American Library Association.

Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24(3), 75-92.

Doerr, M., & Crofts, N. (1999). Electronic esperanto: The role of the object oriented cidoc reference model. *Archives*, 157–173. Retrieved from http://cidoc-crm.org/docs/doerr_crofts_ichim99_new.pdf

Doerr, M., Theodoridou, M., Aspöck, E., & Masur, A. (2015). Mapping archaeological databases to CIDOC CRM. In *CAA 2015. Keep the revolution going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology* (pp. 443-452). Oxford, UK: Archaeopress.

Doerr, M., Felicetti, A., & Hermon, S. (2016). *Definition of the CRMarchaeo, an extension of CIDOC CRM to support the archaeological excavation process*. Retrieved from http://new.cidoc-crm.org/crmarchaeo/sites/default/files/CRMarchaeo_v1.4.1.pdf

Felicetti, A., Gerth, P., Meghini, C., & Theodoridou, M. (2015). Integrating Heterogeneous Coin Datasets in the Context of Archaeological Research. In *Proceedings Workshop EMF-CRM2015, Poznań* (pp. 13-27).

Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., & Doerr, M. (2016). X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, 18(4), 301-319.

Ronzino, P., Hermon, S., & Niccolucci, F. (2012). A metadata schema for Cultural Heritage documentation. In V. Capellini (Ed.), *Electronic Imaging & the Visual Arts: EVA 2012 Florence* (pp. 36-41). Florence, IT: Firenze University Press.