Mapping archaeological databases to CIDOC-CRM

Martin Doerr, Maria Theodoridou
Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH)
N. Plastira 100, Vassilika Vouton, GR-700 13, Heraklion, Crete, Greece
E-Mail: {martin, maria}@ics.forth.gr
URL: www.ics.forth.gr/isl

Edeltraud Aspöck, Anja Masur
ÖAW – Österreichische Akademie der Wissenschaften
Fleischmarkt 22, 1010 Vienna, Austria
E-mail: {Edeltraud.Aspoeck, Anja.Masur}@oeaw.ac.at
URL: http://www.orea.oeaw.ac.at/

Abstract

The exponential growth of the Web and the availability of large number of digital datasets revealed the need for integrated access to heterogeneous and autonomous data sources. In the Cultural Heritage domain, ARIADNE supports transparent integration by means of an advanced metadata schema, based on CIDOC CRM, and adapted to the needs of archaeology.

In this paper we present the mappings of two archaeological databases to the ARIADNE Reference Model. The mapping of dFMRÖ, a relational database of Pre-Roman and Roman Imperial period coins, presents an interesting use case of relational database mapping where there is need to address and separate both categorical and factual information. The mapping of Gräberfeld Franzhausen-Kokoron, a database of 400 Late Bronze Age cremation graves, serves as a test case for a common representation of archaeological records and survey data. The importance of conceptually separating past from present archaeological activities and interpretations will be demonstrated.

Keywords: Mapping technology, information access, semantic search, archaeological databases, cultural heritage information, CIDOC CRM.

1. Introduction

The exponential growth of the Web and the availability of large number of digital datasets revealed the need for integrated access to heterogeneous and autonomous data sources. In the Cultural Heritage domain, ARIADNE, an FP7-INFRASTRUCTURES-2012-1 EU project (Grant agreement no: 313193), brings together and integrates existing archaeological research data infrastructures so that researchers can use the various distributed datasets and new and powerful technologies as an integral component of the archaeological research methodology. The primary role of the CIDOC CRM (ISO 21127:2006) is to enable information exchange and integration between heterogeneous sources of cultural heritage information. CIDOC CRM was chosen as a good starting point for defining ARIADNE's Reference Model, a coherent global ontology for deep integration of scientific and cultural-historical evidence and facts. There is an ongoing effort in ARIADNE to systematically provide sample mappings for the most relevant types of archaeological data sets and in this context several mapping activities were initiated trying to convert existing schemata of archaeological data to CIDOC CRM and its extensions.

In this paper we will present the collaborative work between domain experts (OEAW) and IT experts (FORTH) defining mappings of two archaeological databases to CIDOC CRM. The mapping of dFMRÖ (digitale Fundmünzen der Römischen Zeit in Österreich), a relational database of Pre-Roman and Roman Imperial period coins found in Austria and Romania, presents an interesting use case of relational database mapping where there is need to address and separate both categorical and factual information. We introduce two specialized relations in order to support categorical production, similar

to relations used in the FRBR model. The mapping of Gräberfeld Franzhausen-Kokoron, a database of 400 Late Bronze Age cremation graves from Franzhausen-Kokoron, Austria, with about 3800 records, serves as a test case for exploiting the scientific investigation methods for archaeology and the use of CRMarchaeo and CRMsci, the extensions of CIDOC-CRM to support archaeological excavations (Masur et al, 2014) and scientific observation, measurements and processed data.

The paper is organized as follows: Section 2 presents the ARIADNE Reference Model; Section 3 describes briefly the tool and language that we used to define the mappings; Section 4 presents the mapping of the dFMRÖ, a relational database of Pre-Roman and Roman Imperial period coins found in Austria and Romania while Section 5 presents the mapping of Gräberfeld Franzhausen-Kokoron, a database of 400 Late Bronze Age cremation graves from Franzhausen-Kokoron, Austria.

2. The ARIADNE Reference Model

ARIADNE aims to bring together and integrate the existing archaeological research data infrastructures so that researchers can use the various distributed datasets and new and powerful technologies as an integral component of the archaeological research methodology. An immense number of archaeological digital datasets and encoded facts are placed on the Web, in digital repositories and other information systems every day. They are the accumulated outcome of the research of individuals, teams and institutions, but form a vast and fragmented corpus and their potential is constrained by difficult access and non-homogenous perspectives.

It is therefore important to build infrastructure and web-services that will allow for exploration, data-mining, semantic integration and experimentation across all these rich resources.

Therefore the ARIADNE Research Infrastructure Project for Archaeology aims at going beyond the current Digital Library paradigm with simplistic findings aids by laying the foundation for the integration of rich, structured information from all heterogeneous sources that may be relevant for answering a research question. The first aim is a common, consistent representation of data that have a potential bearing on questions beyond their local context of creation and use, so that directly and deep-indirectly related facts can be filtered out effectively from the mass in order to support further interpretation by the researcher.

Only Semantic Web technologies and formal ontologies allow for such a common representation and effective management of billions of facts. The respective technology is very rapidly advancing. Therefore the challenge of the day is not to adapt data models to the possibly still limited performance of current platforms, but to develop a global, extensible schema in the form of a formal ontology that allows for integration without loss of meaning, rather than "core fields" and "application profiles". In the end, this appears to be a more demanding task than the development of performant platforms. Also, the creation and maintenance of data in adequate form exceeds the cost of the development of platforms by some order magnitude. Therefore manually restructuring data at each technology step should be replaced by transforming data to comprehensive structures with expected long-term validity, interoperability and extensibility. This is a task of highly interdisciplinary ontology engineering.

In order to address the complexity of archaeological data integration, ARIADNE's main challenge was to develop a global, extensible schema in the form of a formal ontology that will allow for integration without loss of meaning. CIDOC CRM was chosen as the backbone of the ARIADNE Reference Model and a suite of extensions was developed to address the complexity of archaeological data integration. CIDOC CRM (ISO21127) is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It started bottom up, by reengineering and integrating the semantic contents of more and more database schemata and documentation structures from all kinds of museum disciplines, archives and recently libraries as empirical base. CIDOC CRM contains the most basic relationships to describe

what happened in the past at a human scale, i.e. people and things meeting in space-time, parts and wholes, use, influence and reference. More detailed kinds of discourse require extensions.
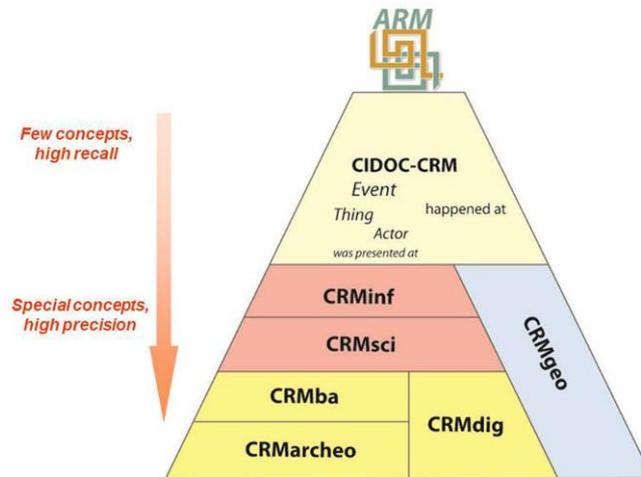
Figure 1: ARIADNE Reference Model (ARIADNE RM)

The ARIADNE RM, presented in Figure 1 comprises the following extensions:

- **CRMinf** (CRMinf, 2015): the Argumentation Model is a formal ontology intended to be used as a global schema for integrating metadata about argumentation and inference making in descriptive and empirical sciences.
- **CRMsci** (CRMsci, 2014): the Scientific Observation Model is a formal ontology intended to be used as a global schema for integrating metadata about scientific observation, measurements and processed data in descriptive and empirical sciences.
- **CRMgeo** (CRMgeo, 2013): a spatiotemporal model that provides an "articulation" (linkage) between the standards of the geospatial and the Cultural Heritage community in particular between GeoSPARQL and CIDOC CRM.
- **CRMdig** (CRMdig, 2014): a model for provenance metadata is an ontology to encode metadata about the steps and methods of production ("provenance") of digitization products and synthetic digital representations such as 2D, 3D or even animated models created by various technologies.
- **CRMba** (Ronzino, 2015): the Buildings Archaeology is an ontology developed for investigating historic and prehistoric buildings, the relations between building components, functional spaces, topological relations and construction phases through time and space.
- **CRMarchaeo** (CRMarchaeo, 2014): the Excavation Model is an ontology to encode metadata about the archaeological excavation process.

3. The mapping process

The process of mapping the schema of an archaeological database to a common coherent global ontology is not trivial and needs support from appropriate tools (Oldman et al., 2014). First we need a sufficient mapping specification to support the transformation of each distinct schema of an archaeological data set (source schema) into the ARIADNE RM (target schema). It is crucial that during the transformation, the information encoded in the source database should not be lost, the initial "meaning" should be preserved as much as possible. In several cases, that we will present in more detail in the following sections, information in the source database is implicit, hidden in forms, user interface fields or in the worst case in the head of a curator, the domain expert. This implies that the transformation process cannot be done solely by technical people. It requires the
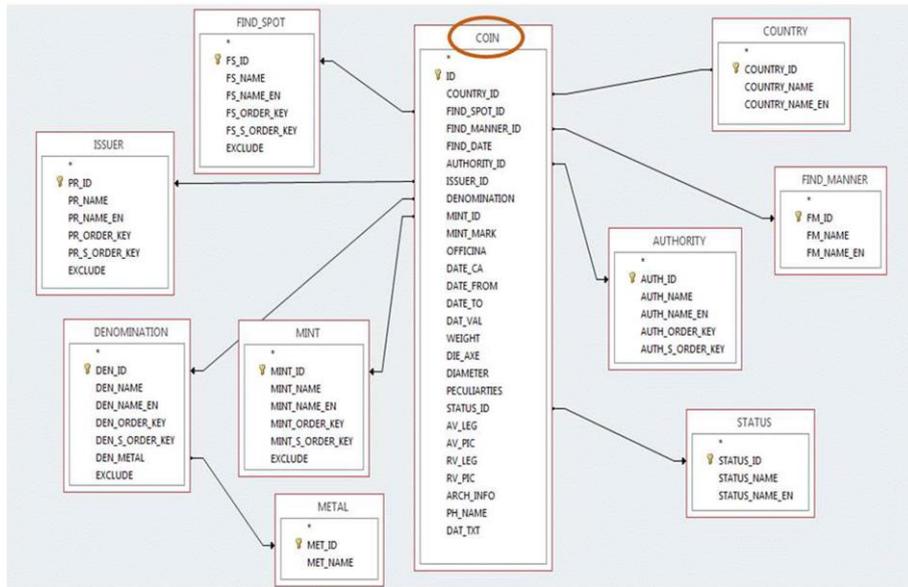
3

close collaboration between the domain experts who possess the domain knowledge and the IT experts who possess the semantic world technical knowledge.

We have implemented 3M (3M, 2015), a Mapping Memory Manager that includes a mapping editor, a mapping specification and a mapping memory to accumulate knowledge and experience. The mapping specification is based on X3ML (X3ML, 2014), an XML based declarative language aiming to support the cognitive process of a mapping definition. It describes schema mappings in both human and machine readable form and supports the close collaboration of domain and IT experts. X3ML separates schema matching – performed mainly by domain experts – from URI generation and terminology mapping – performed mainly by IT experts.

Figure 2: Mapping Workflow

4. Mapping dFMRÖ (digitale Fundmünzen der Römischen Zeit in Österreich)

Digitale **F**und**m**ünzen der **R**ömischen Zeit in **Ö**sterreich (digital Coin-finds of the Roman Period in Austria) is an online MySQL database of the Numismatic Research Group of the Austrian Academy of Sciences (dFMRÖ 2007). Since the 1990s it documents coin-finds from the Celtic and Roman Period that have been published in various printed volumes of the FMRÖ (**F**und**m**ünzen der **R**ömischen Zeit in **Ö**sterreich / Coin-finds of the Roman Period in Austria) from the 1970s up to 2007. Starting with a Microsoft Access database, it was set up in its current form in 2007 and hosts about 76.000 finds. All coins were found in Austria or, because of a former project cooperation, in Romania. The coins represent an important part of the Austrian cultural heritage.

Figure 3: The relational schema of dFMRÖ

The relational schema of dFMRÖ was chosen as an interesting first mapping use-case since it represents a large class of well-defined traditional databases. Is consists of a main COIN table and nine auxiliary tables linked together with simple joins as it can be seen in Figure 3. The database contains data describing the objects/coins such as measurements, metal, preservation etc. and contextual data e.g. findspot, archaeological context and dating (Figure 4).
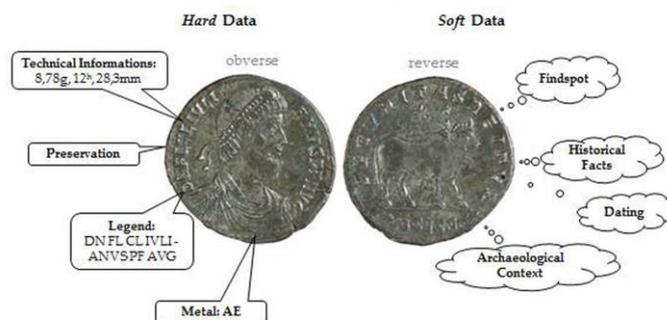

Figure 4: COINS - an archaeological source.

The goal is to interpret the relational schema as a semantic model (*domain – property – range*) and we achieve it through the transformation of the tables and fields of the relational schema of dFMRÖ into equivalent ARIADNE RM paths. Our mapping starts with the specification of the domain mapping which defines that the records of the main dFMRÖ table (COIN) correspond to instances of the entity E22 Man-Made Object of CIDOC CRM and two approaches are possible:

1. Introduce a specialization of E22 Man-Made Object and enhance the ARIADNE RM:
   Exx COIN *subclass of* **E22 Man-Made Object**
2. Define the Type of E22 Man-Made Object:
   E22 Man-Made Object. *P2 has type:* **E55 Type = "Coin"**

The introduction of a new class should comply with the "*Minimality*" modelling principle of CIDOC CRM (CIDOC CRM, 2015):

"*A class is not declared unless it is required as the domain or range of a property not appropriate to its superclass, or it is a key concept in the practical scope.*"

Field names stand for a relationship (*property*) and the kind of contents (*range*) while the field contents stand for entity instances. In the simplest case an equivalent proposition can be defined as a one-to-one mapping like the one in Figure 5. In this example the field ID is the source property mapped to the target property **P1 is identified by** and also the source range mapped to the target range **E42 Identifier**. It was decided that local identifiers in relational database tables will be mapped explicitly only if these identifiers are visible in the user interface and used in other documents as well. Alternatively, the local database identifiers are used only for generating URIs for the record instance, here the coin instance, and the identifier fields are not mapped at all.
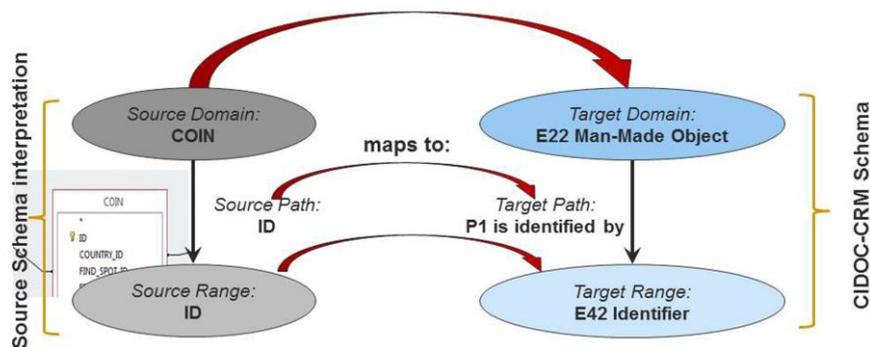
Figure 5: Creating an Equivalent Proposition

In most cases it is not possible to have a one-to-one mapping. Instead, one field of the source maps to a complex path in the target. Figure 6 demonstrates one such case. In the dFMRÖ COIN table, the field WEIGHT that contains the weight value for each coin instance cannot be mapped to a single CIDOC CRM class. In CIDOC CRM, the class E54 Dimension (*which comprises quantifiable properties that can be measured by some calibrated means and can be approximated by values*) has three properties:

- *P90 has value* has as range a literal that is the actual weight value.
- *P91 has unit* has as range an E58 Measurement Unit which in the case of dFMRÖ is grams, information that is not encoded in the relational database but is built into the User Interface of dFMRÖ.
- *P2 has type* has as range an E55 Type that takes the constant value of "weight".
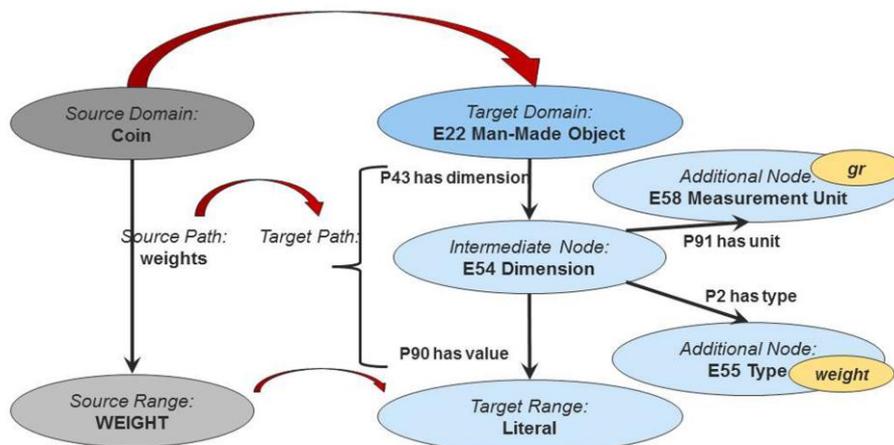
The simple source path "COIN has weight WEIGHT" maps to the complex path

  E22 Man-Made Object. *P43 has dimension*: E54 Dimension. *P90 has value*: Literal
         E54 Dimension. *P91 has unit*: E58 Measurement Unit = "gr"
         E54 Dimension. *P2 has type*: E55 Type = "weight"

E54 Dimension is an **intermediate node** in the X3ML terminology while E58 Measurement Unit and E55 Type are two **constant expressions**. The intermediate node and constant expression constructs of X3ML provide flexibility during the mapping and allow the encoding of implicit, "hidden" information of the source. In the current implementation of X3ML, the source schema is specified as xpaths and dedicated construct allows the specification of relational joins.

dFMRÖ contains factual data describing a specific coin such as the find spot and categorical data describing a category of coins such as historical facts (

Figure 4). In order to distinguish these two types of information we extended our model with the property *PC1 produced things of type* that supports categorical production similar to FRBRoo *R26 produced things of type* (FRBRoo, 2015).

During the mapping process it was identified that the concept "issuing a coin" in the coin database is not yet adequately covered by CIDOC CRM. "Issuer" is an accidental role that does not characterize an actor independently from particular contexts or activity. Therefore the Actor does not have the type "Issuer" but the activity has the type "Issuing":

The join of the source table COIN, field ISSUER_ID to the table ISSUER, field PR_ID (ISSUER_ID == PR_ID) maps to the complex path

  E22 Man-Made Object. *P108i was produced by*: E12 Production
     E12 Production. *P17 was motivated by*: E7 Activity.*P14 carried out by*: E39 Actor
         E7. Activity. *P2 has type*: E55 Type = "Issuing"

We propose the specialization *"gave order"*of *P17 was motivated by* in order to cover the "issuing a coin" concept.

The detailed mapping of dFMRÖ is available at (dFMRÖ , 2015).

5. Mapping Gräberfeld Franzhausen-Kokoron

The Gräberfeld Franzhausen-Kokoron database contains information on 403 cremation graves from the late Bronze Age Urnfield Culture (1050–800BC; Lochner, Hellerschmid 2010). The graves were excavated between 1981-84 and in 1991 in the small market town Franzhausen in eastern Austria. Graves typically contained urns with burnt human remains, un-burnt animal bones and frequently copper alloy objects (pins, fibulas, arm- and neckrings, razors and knives) were found.

A Microsoft Access database was created in 2006 to catalogue and analyse the graves. In 2010 the database was published online with an interactive cemetery plan interface for viewing the records of each grave (Lochner Hellerschmid 2010). The database provides information on the archaeological features (grave pits, position of urns etc.) and finds (e.g. pottery and bronze objects) as well as results from analysis of human remains (age at death, sex) and animal remains.
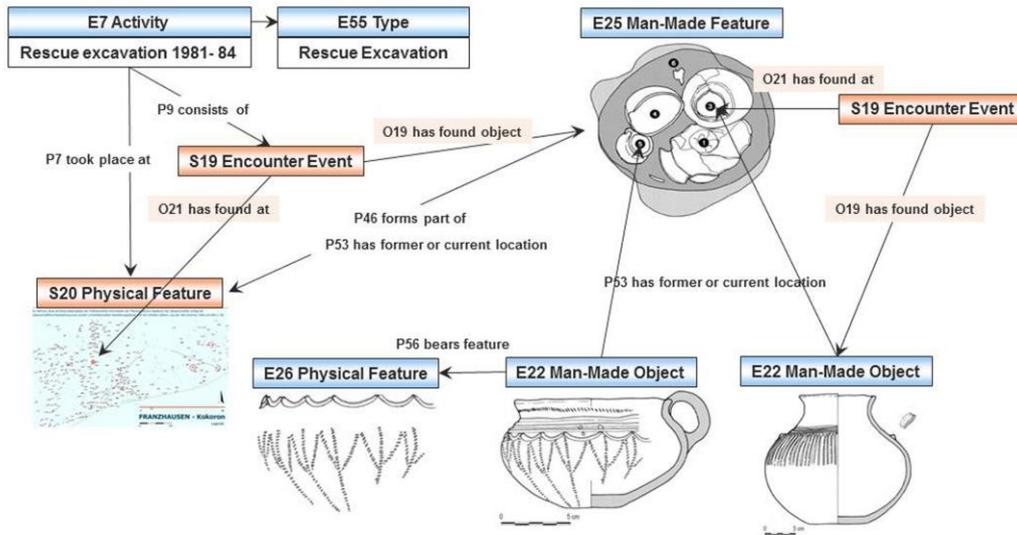
Figure 7: Part of FK Cemetery semantic graph.

Figure 7 presents a part of the semantic graph that is created with the mapping to the ARIADNE RM. The two rescue excavations are mapped to two E7 Activities. Each of these activities consists of several S19 Encounter Events (CRMsci, 2015) which correspond to the finding of each grave (E25 Man-Made Feature) and the finding of each object (e.g. vessels, E22 Man-Made Object) inside the graves. The information about the vessel decoration (E26 Physical Feature) is linked to the vessel through the *P56 bears feature* property. The relational database makes extensive use of compound keys (see Figure 8) and the simple join construct of X3ML is not sufficient. A special construct will be implemented in order to support compound keys.

The mapping exercise identified limitations of the original database that can be resolved or improved in the future with the use of ARIADNE RM. Information about the excavations is not recorded explicitly but semantics are built into the User Interface, the queries or the values of the identifiers. For example *Befundnummer* which is a field that identifies the finding is a three digit number (*nnn*) if the finding belongs to the 1981-84 excavation and a five digit number (*nnn91*) if the finding belongs to the 1991 excavation. The mapping under condition construct of X3ML allows to extract this hidden excavation information from the Franzhausen-Kokoron database and create the appropriate S19 Encounter Events which are parts of the respective rescue excavations (E7 Activity). These events can play the role of hooks for adding information in the future about the excavations, resulting in knowledge revision and augmentation. Current limitations of preservation, migration and extensibility of the database can potentially be solved. A thorough analysis of the goals of excavations (what was the archaeological question?) can maximize the interpretation capability after excavation and allow the comparison of previous excavations on same site or space. Revision of knowledge based on existing data is feasible by adding information about excavations. Statistical analysis of the data is also facilitated and it is possible to identify patterns in the data e.g. age at death related to the location of the graves. Last but not least, the mapping to a common coherent ontology such as the ARIADNE RM will allow future integration with other relevant databases something completely impossible to do in the current state of the database.
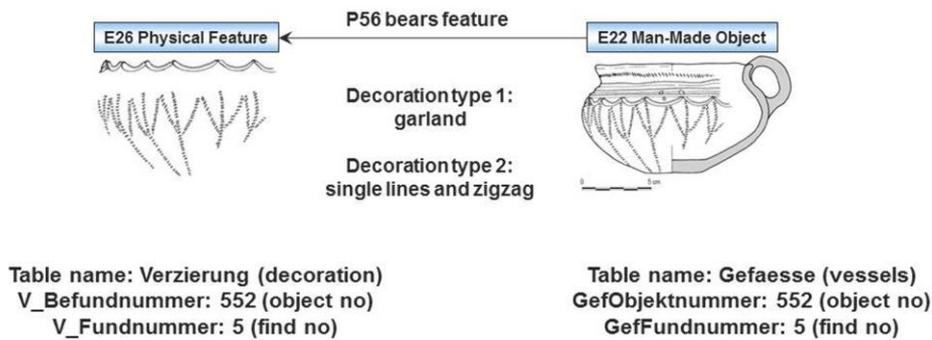
Figure 8: Join with compound keys.

6. Conclusions

In this paper we presented the mappings of two indicative archaeological databases to the ARIADNE Reference Model, a global ontology based on CIDOC CRM and an extension suite that allows deep integration of scientific and cultural-historical evidence and facts. The mapping process was assisted by the 3M editor and the X3ML mapping definition language and it was achieved by the close cooperation of the domain experts with the IT experts.

We have demonstrated that different databases about quite specific archaeological content can successfully be integrated under the much more generic ARIADNE Reference Model without loss of specificity of meaning. We have further demonstrated that the X3ML mapping mechanism is adequate for the domain experts to understand, define and verify the semantics of mapping specifications.

This is a major step forward to realize large-scale, high-quality information integration of scientifically relevant data under a rich schema, as valuable resources for the researcher, in contrast to the core-data approaches of major aggregators, such as Europeana, and some other research infrastructure projects.

Mapping archaeological databases to CIDOC CRM and its extensions can (a) serve as a guide for good-practice data structures; (b) be used to create a Semantic Web of cultural knowledge; (c) compare and/or enhance knowledge from other similar databases; (d) support preservation, migration and extensibility; and (e) enable all kinds of comprehensive statistical studies.

Future work includes the mapping of other relevant databases, the creation of the integrated semantic repository and the compilation of a list of interesting archaeological questions that can be used to validate the repository.

7. Acknowledgements

8. References

CIDOC CRM. (2015) *Current Official Version of the CIDOC CRM. Definition of the CIDOC Conceptual Reference Model. The version 6.1 of the reference document.* [Online] Available from: http://www.cidoc-crm.org/docs/cidoc_crm_version_6.1.pdf [Accessed: 23th June 2015]

CRMarchaeo. (2014) *CRMarchaeo: the Excavation Model, version 1.2.1.* [Online] Available from: http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo1.2.1.pdf.
RDFS encoding available from: http://www.ics.forth.gr/isl/CRMext/CRMarchaeo.rdfs [Accessed: 23th June 2015]

CRMdig. (2014) *CRMdig: an Extension of CIDOC‐CRM to support provenance metadata, version 3.2.* [Online] Available from: http://www.ics.forth.gr/isl/CRMext/CRMdig/docs/CRMdig3.2.pdf.
RDFS encoding available from: http://www.ics.forth.gr/isl/CRMext/CRMdig.rdfs [Accessed: 23th June 2015]

CRMinf. (2015) *CRMinf: the Argumentation Model, version 0.7.* [Online] Available from: http://www.ics.forth.gr/isl/CRMext/CRMinf/docs/CRMinf-0.7.pdf [Accessed: 23th June 2015]

CRMsci. (2014) *CRMsci: the Scientific Observation Model, version 1.2.2.* [Online] Available from: http://www.ics.forth.gr/isl/CRMext/CRMsci/docs/CRMsci1.2.2.pdf.
RDFS encoding available from: http://www.ics.forth.gr/isl/CRMext/CRMsci.rdfs [Accessed: 23th June 2015]

dFMRÖ (2007) dFMRÖ – digitale Fundmünzen der Römischen Zeit in Österreich [Online] Available from: http://www.oeaw.ac.at/antike/index.php?id=358 [Accessed: 25th June 2015]

dFMRÖ (2015) dFMRÖ mapping to CIDOC CRM [Online] Available from: http://139.91.183.3/3MEditor/Index?type=Mapping&id=Mapping153&lang=en [Accessed: 25th June 2015]

DOERR, M., Hiebel, G. (2013) CRMgeo: Linking the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement. TECHNICAL REPORT: ICS-FORTH/TR-435, April 2013. [Online] Available from: http://www.ics.forth.gr/isl/CRMext/CRMgeo/docs/TR435-CRMgeo.pdf
RDFS encoding available from: http://www.ics.forth.gr/isl/CRMext/CRMgeo.rdfs [Accessed: 23th June 2015]

FRBRoo. (2015) *FRBR object-oriented definition and mapping from FRBR$_{ER}$, FRAD and FRSAD (version 2.1)* [Online] Available from:
http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V2.1_2015February.docx
[Accessed: 23th June 2015]

RONZINO, P. (2015) *CIDOC CRMBA. A CRM extension for buildings archaeology information modeling, Unpublished PhD Thesis*

LOCHNER, M., HELLERSCHMID, I. (2010) *Ein Gräberfeld der jüngeren Urnenfelderkultur aus Franzhausen-Kokoron (UFK).* Katalog und Abbildungen, Version 02/epub, doi:10.1553/KatalogUFK [Online] Available from: epub.oeaw.ac.at/franzhausen-kokoron [Accessed: 25th June 2015]

Masur, A., May, K., Hiebel, G., Aspöck, E. (2014) *Comparing and mapping archaeological excavation data from different recording systems for integration using ontologies.* CHNT Proceedings, Stadtarchäologie, Museen der Stadt Wien (Hrsg.), CHNT Proceedings 2013. [Online] Available from: http://www.chnt.at/wp-content/uploads/Masur_etal_2014.pdf [Accessed: 25th June 2015]

OLDMAN, D., DOERR, M., DE JONG, G., NORTON, B., WIKMAN, T. (2014) *Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning & Aggregation Services for the Digital Humanities.* D-Lib Magazine 20(7/8) (July/August), doi:10.1045/july2014-oldman. [Online] Available from: http://www.dlib.org/dlib/july14/oldman/07oldman.html [Accessed: 23th June 2015]

X3ML (2014) *The X3ML engine*. [Online] Available from: https://github.com/delving/x3ml [Accessed: 23th June 2015]

3M (2015) *The Mapping Memory Manager*. [Online] Available from: http://www.ics.forth.gr/isl/3M [Accessed: 23th June 2015]