D14.2: Pilot deployment experiments

Author:
Philipp Gerth, DAI

Ariadne is funded by the European Commission's 7th Framework Programme.

Version: 1.0 (*final*) **11th November 2016**

Author: **Philipp Gerth, DAI**

Contributing Partners **Paola Ronzino, Achille Felicetti, PIN**
Maria Theodoridou, Martin Doerr, Eleni Christaki, FORTH
Edeltraud Aspöck, Anja Masur, OEAW
Wolfgang Schmidle, Sebastian Cuy, DAI
Douglas Tudhope, Ceri Binding, USW

Quality review: **Holly Wright, ADS**

Table of Contents

Executive Summary

The primary goal of the EU project ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) is to bring together and integrate existing archaeological data infrastructures to offer researchers a unified search and discovery on a wide range of distributed datasets, which is very much needed in archaeology. In addition to the aggregation efforts that resulted in a general catalogue of archaeological resources in the ARIADNE portal, several case studies examined a tighter and semantically richer integration using Semantic Web technologies. The purpose of this document (deliverable D14.2) is to describe the activities carried out within Work Package 14 - Addressing Complexity (WP14), which results in the presentation of four datasets.

In general, the integration of various datasets faces some challenges in terms of different languages, terminology and data models used. The integrated datasets have different data models to describe the content according to historical reasons and intents of the project, and where the data was created. After a short introduction (chapter 1), the ARIADNE Reference Model is described in chapter 2. It offers a common schema to describe these different data models. The global ontology used in these scenarios was CIDOC-CRM, which was adapted and extended for the domain-specific use in archaeology by the extension CRM-ARCHAEO.

Chapter 3 includes a classification and description of existing, referenceable systems for controlled vocabulary, thesauri and references that are available online. These contain common terms for time, space and domains that are need to be harmonized and brought together to overcome the differences in language and terminology.

The universal workflow to import the data, transform and export it to a format, which uses the ARIADNE Reference Model and common standards, is described in chapter 4. The tools and software, which help the users map their data, perform data transformation, for data storage and for final presentation, are outlined in chapter 5.

Chapters 6-9 describe the scenarios used to test and confirm the ARIADNE Reference Model with real world data. The scenarios were selected to cover a wide range of different archaeological research perspectives with different recorded classes.

Finally, chapter 10 gives an overview of the data integration activities.

1. Introduction and Objectives

The archaeological domain produces vast amounts of very heterogeneous research data like excavation data, object descriptions, texts, archives, cadastral plans, historical maps coming from various disciplines with different methodologies. This implies a big challenge for organizing, aggregating and publishing the data. The EU project ARIADNE aims at bringing together and integrating existing archaeological data infrastructures. In addition to the aggregation efforts that resulted in a general catalogue of archaeological resources in the ARIADNE portal (http://portal.ARIADNE-infrastructure.eu/), several case studies examined a tighter and semantically richer integration using Semantic Web technologies. To allow an integration of this heterogeneous information without the loss of meaning, a global, extensible schema, defined in D14.1, was used. It uses CIDOC-CRM as a backbone ontology, together with several extensions bound together and described in the ARIADNE Reference Model.

The selected scenarios diverge from the proposed scenario described in the DOW, as data availability was a limiting factor to most of the proposals. The deployed scenarios focussed on different aspects of archaeological research:

Coins: Numismatics is a very traditional science with a lot of experience and early initiatives in standardization. As such, it was chosen as a very good starting point for item-level integration of archaeological datasets. The linkage between repository metadata and item-level data, as well as the integration of natural language processing data, were tested.

Sculptures: This scenario concerns data integration of sources from various disciplines, including sculpture information and its archaeological context. It focuses on the provenance of information according to bibliographic references, leading to advanced literature research by using the knowledge graph for discovery.

Buildings: This scenario is about the application of two extension of the CIDOC CRM, namely CRMba and CRMarchaeo, to model unstructured information collected during the archaeological excavation of the prehistoric Zominthos' Palace in Crete. This is an important achievement as it represents one of the first attempts of integrating archaeological layers and built structure with a knowledge-based model.

Animal Remains: In this scenario different zooarchaeology datasets about bone assemblages found in diverse archaeological sites were integrated and different statistical queries were run on the common repository.

Wood/Dendrochronology case study: An additional case study is described in the ARIADNE deliverable D15.3 - Semantic Annotation and Linking. This case study investigated the semantic integration of archaeological datasets with grey literature in different languages, as a combined effort between ARIADNE WP15 and WP16. The case study is based on archaeological interest in different types of wooden material, samples taken, wooden objects and dating via dendrochronological techniques. Information from Dutch, English, Swedish archaeological reports was extracted by Natural Language Processing pipelines. The work was undertaken by University of South Wales on the technical side, in collaboration with DANS and SND.

Like the case studies described below in D14.2, the wood/dendrochronology case study uses the CIDOC-CRM as a backbone ontology, together with the Getty Art and Architecture Thesaurus for common vocabulary, expressing the semantic integration in RDF. The case study Demonstrator is a SPARQL query builder Web application that seeks to hide the complexity of the underlying ontology. As the user selects from the interface, an underlying SPARQL query is automatically constructed in terms of the semantic framework. Queries concern wooden objects (e.g. samples of beech wood keels), optionally from a given date range, with automatic expansion over hierarchies of wood types.

2. ARIADNE Reference Model

The ARIADNE Reference Model (http://www.ARIADNE-infrastructure.eu/Resources/ARIADNE-Reference-Model) was developed by research activities in WP 14, lead by FORTH, and is extensively documented in deliverable D14.1 – Extended CRM. The ontology CIDOC-CRM (http://www.cidoc-crm.org/) is used as a backbone for the ARIADNE Reference Model. CIDOC CRM is in use for describing information in the context of cultural heritage and museums. In 2006, CIDOC CRM became an ISO standard, and nowadays is widely used in galleries, libraries, museums, research institutes and archives. In the ARIADNE Reference Model it works as a top-level ontology, used to enable semantic interoperability to the more domain specific ontologies, extending CIDOC CRM.

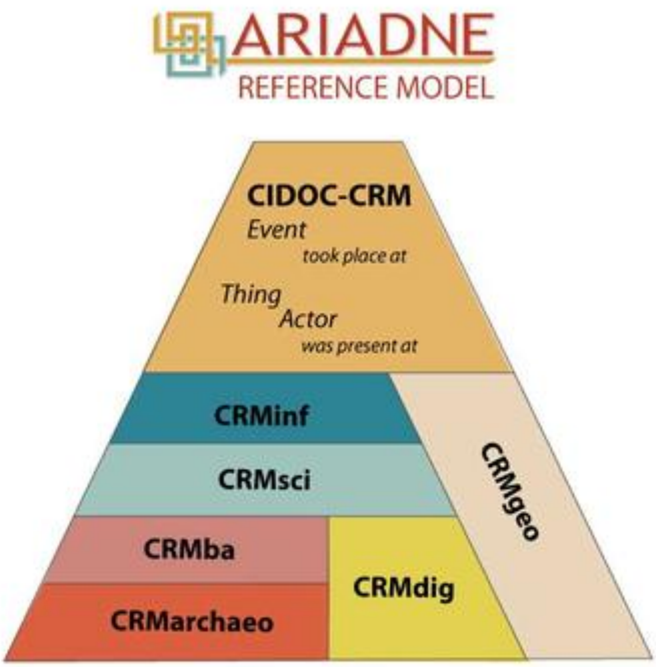


Figure 1 ARIADNE Reference Model

- *CRMinf (CRMinf, 2015):* the Argumentation Model is a formal ontology produced by Stephen Stead, Paveprime Ltd and collaborators, and is intended to be used as a global schema for integrating metadata about argumentation and inference making in descriptive and empirical sciences, such as biodiversity, geography, archaeology, cultural heritage conservation, research IT environments and research data libraries. Its primary purpose is to facilitate the management, integration, mediation, interchange and access to data about reasoning by a description of the semantic relationships between the premises, conclusions and activities of reasoning. Besides application-specific extensions, this model is intended to be complemented by CRMsci, a more detailed model and extension of the CIDOC CRM for metadata about scientific observation, measurements and processed data in descriptive and empirical sciences, also currently available in a first stable version.
  Reference document and RDFS encoding are available in version 0.7.

- CRMsci (CRMsci, 2016): 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, such as biodiversity, geography, archaeology, cultural heritage conservation, research IT environments and research data libraries. Its primary purpose is to facilitate the management, integration, mediation, interchange and access to research data by description of semantic relationships, in particular causal ones. CRMsci is in the process of being validated in the context of the ARIADNE project. The model is not "finished", and all constructs and scope notes are open to further elaboration.
  Reference document and RDFS encoding are available in version 1.2.2

- *CRMgeo (CRMgeo, 2013):* the Spatiotemporal Model provides a linkage between the standards of the geospatial and the Cultural Heritage community, in particular between GeoSPARQL and CIDOC CRM. This models aims at being a comprehensive theory from which all kind of places could be described, such as - this the place of the Varus Battle or is this the place where Lord Nelson died, including geometric specifications.
  Reference document and RDFS encoding are available in version 1.2.

- CRMdig (CRMdig, 2014) is an ontology and RDF Schema 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. Its distinct features are the complete inclusion of the initial physical measurement processes and their parameters. CRMdig has been developed as a compatible extension of CIDOC CRM (ISO21127), which allows for querying the most relevant facts and returning complete descriptions encoded in this model by generic CIDOC CRM terms without the need to refer to its specific properties. The applications so far perfectly confirm the wide applicability and potential of this model for all kinds of scientific data and other digital objects and its superior maturity in terms of coverage, genericity, expressive power and level of detail.
  Reference document and RDFS encoding are available in version 3.2.1.

- *CRMba (CRMba, 2016; Ronzino, 2015):* the Buildings Archaeology model was conceived to support the process of recording the evidences and the discontinuities of matter in archaeological buildings, in order to identify the evolution of the structure throughout the centuries and to record the relationships between each of the building components among them and with the building as a whole. It aims to express the semantic relations of the stratigraphic units of a standing building, taking into account the stratigraphic analysis theory of the standing buildings. CRMba is in the process of being validated in the context of the ARIADNE project. The model is not "finished", all constructs and scope notes are open to further elaboration.
  Reference document and RDFS encoding are available in version 1.4.

- CRMarchaeo (CRMarchaeo, 2016): CRMarchaeo is an ontology and RDF Schema to encode metadata about the archaeological excavation process. The model was created to support this process and all the various entities and activities related to it. The model was created starting with standards and models already in use by national and international cultural

heritage institutions, and has been enriched by continuous collaboration with various communities of archaeologists from different countries and schools. CRMarchaeo is intended to provide all necessary tools to manage and integrate existing documentation in order to formalise knowledge extracted from observations made by archaeologists, recorded in various ways and adopting different standards. Its purpose is to facilitate the semantic encoding, exchange, interoperability and access of existing archaeological documentation. The model documents, in a transparent way, the various aspects of the archaeological excavation process, including the technical details concerning different methods of excavation, the reasons for their application and the observations made by archaeologists during their activities in the field. This approach allows the creation of an objective documentation that can guarantee the scientific validity of the results, making them revisable following further investigations, and reusable in different research contexts, in order to answer further (and potentially different) research questions.
Reference document and RDFS encoding are available in version 1.4.

3. Linked Data Enrichment

The use of an ontology alone does not automatically lead to data integration of different datasets, as the ontology just describes the semantics of the information model. The values and terminologies used will still be diverse in terms of different spellings, errors and languages. For instance, a site name can differ from dataset to dataset. In the example below (Figure 2 Example for Linked Data enrichment for place terms using the iDAI.gazetteer.), two different objects, one sculpture documented in Arachne (Figure 2 Example for Linked Data enrichment for place terms using the iDAI.gazetteer.left) and one sarcophagus documented in the British Museum collection (Figure 2 Example for Linked Data enrichment for place terms using the iDAI.gazetteer.right), were found in the same place. As the value for the same place differs, it could not be identified as the same site, therefore it is ultimately necessary to refer to online resources, i.e. a gazetteer, that act as standardization tools instead of text strings wherever possible.

Different cases, where standards could be applied, were identified and will be explained in further detail.

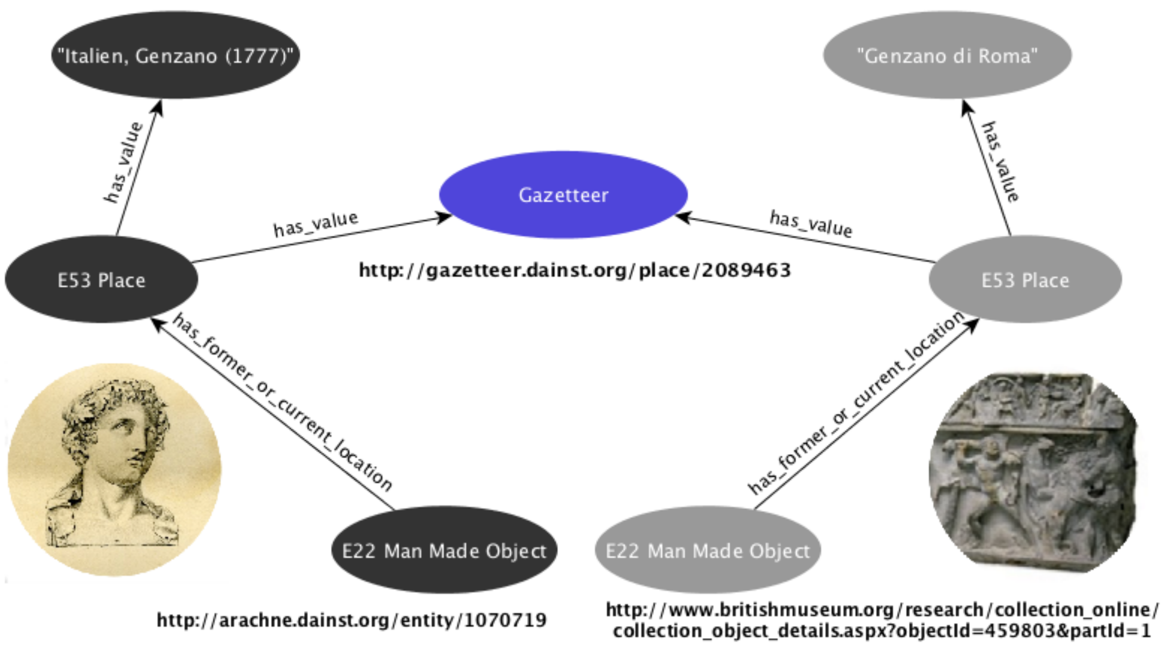


Figure 2 Example for Linked Data enrichment for place terms using the iDAI.gazetteer.

3.1 Places

The location of a find or the different spellings and the position of an archaeological site are fundamental pieces of information used by many systems. Therefore, it is of great importance to provide systems, which offer a referenceable URI to the desired location with different language representations, spellings and location. The concept of Gazetteers as a reference system is quite old and was adopted early on by archaeology (e.g. Clark 1932). It is pretty straightforward to enhance them semantically. In terms of standardization it is the most advanced class and there are some possibilities:

- Geonames (http://www.geonames.org/) is a gazetteer of worldwide place names and offers access to over 10 million place names, which are categorized in 9 different classes. It has a generic approach and offers very dense data of recent populated places. However, archaeological places are quite sparse. For example, Geonames provides only 117 archaeological sites in Greece (http://api.geonames.org/search?featureCode=ANS&maxRows=1000&country=GR).
- Pleiades (http://pleiades.stoa.org/home) was started in 2006 as a digital representation of the Barrington Atlas of the Roman and Greek World, but was soon adopted for a wider scope. With 35.000 ancient sites, it is the largest gazetteer for the ancient world.
- iDAI.gazetteer is geographical database developed at the German Archaeological Institute, which provides more than 1 million entries to describe modern and ancient places that are of interest to archaeologists and also acts as a hub by linking other Gazetteers like Geonames and Pleiades. Its scope also includes geographic data recorded at site or building level, which is particularly useful when connecting different archaeological databases that overlap in geographic coverage.

3.2 Time

Temporal information is as important as spatial information. For the use cases described we have not used a time gazetteer, but it is highly recommended to do so in the future, once the systems are more technically evolved and filled with data. There are a few available time gazetteers.

- PeriodO (http://perio.do/) is a large time gazetteer with a pragmatic data model for the Semantic Web. It is supported by the ARIADNE project.
- iDAI.chronontology (http://chronontology.dainst.org/) is a time gazetteer currently in development at the German Archaeological Institute. It has a more complex data model, which distinguishes between the definition of a period and the dating information based on this definition.

3.3 Literature

The standardization in the context of libraries has a long history, beginning with in-house rules for a specific library and leading to the close cooperation between libraries nowadays (Thacker, 2000). To this end, metadata standards like the MARC21 and international identifiers like the International Standard Book Number (ISBN) were established. ISBN numbers, with their introduction in 1972 as the ISO standard 2108, are commonly accepted and used. Unfortunately older publications don't have ISBN numbers, and as the registration is fee-based, smaller publications and specialised books are often published without ISBN numbers, therefore registration isn't possible.

There are a lot of norm data services for books, for example WorldCat (http://www.worldcat.org/). However, Zenon, with its focus on classical studies and entries for individual articles, was the most obvious choice. Zenon is the bibliographical catalogue of the German Archaeological Institute including the data of all DAI libraries, the DEI Amman and the Winckelmann Institute. The resource was used for a use case since it is one of the biggest databases for classics with more than 1.2 datasets and covers, beside books, e-resources, maps and archival material. The metadata standard MARC21 is used and URIs are provided.

3.4 Archaeological Terminology

In general, there are two different approaches for archaeological terminology. One could classify and define the terminology for the whole archaeology and classics in a very generic way, or one could try to define a terminology only for a very subject-specific subset. There is always a trade off between scope and precision. A combination of both procedures, where the specialised terminology links to the more generic terminology with some overlapping terms, could be a hybrid solution.

Even though there is a lot of specialized vocabulary in use by bigger institutions, they often do not provide their vocabulary or it is not available in an appropriate format. The British Museum's vocabulary is already provided in Semantic Web formats, but unfortunately, as Gruber and Smith (2014, 207) have stated, they are not linked to external vocabularies, which makes the integration with other datasets more complicated, and the datasets and the vocabulary an open but isolated island.

- Specialist Ontologies: the terminology of Nomisma.org (http://nomisma.org/) for describing coins and the evolving terminology of Kerameikos.org (http://kerameikos.org/) for describing Greek ceramics is readily available in Linked Open Data formats. They are also providing links to more generic terminology, which allow an easy integration of the specialized datasets.
- The Getty Arts & Architecture Thesaurus, AAT in short (http://www.getty.edu/research/tools/vocabularies/aat/), is a highly evolved thesaurus for the humanities. It offers a lot of different branches following ISO standards, but as it follows a generic approach, it lacks specialization at some point. Other languages are supported, but rarely exist besides Dutch and Spanish.
- The iDAI.vocab (http://archwort.dainst.org/de/vocab/) is a thesaurus of archaeological terminology with currently 14 different languages. Its aim is to collect and organize the terminology used in the services of the DAI. The German thesaurus operates as the central hub: a German term, for instance, is linked to translations to all the other available languages. The concepts are classified according to the relations specified by the SKOS standard. The words are connected to the equivalent concepts in other reference works, such as the Getty's AAT and Dbpedia.

3.5 Actors

Actors are of importance in archaeology for specific find categories, e.g. the issuer of a coin, generally named on the object, or actors mentioned in inscriptions or iconographic items. Those actors are not necessarily real historical persons, they could also be mythological characters like the deities of Greek mythology.

- The Virtual International Authority File (VIAF) (http://viaf.org/) is the biggest international authority system for persons, which combines several reference datasets and maps them to the same person. Its data mostly derives from the different library systems and thus works well for real historical persons and writers, which can be referenced by URIs, for example http://viaf.org/viaf/60280417. However, the information about mythological actors is very sparse and contains a lot of duplicates.
- Wikidata (https://www.wikidata.org/) is the data backbone of Wikipedia. It contains a lot of mythological characters and currently, together with DBpedia, is the best resource available for mythological characters. It's also 5-star Linked Open Data, as it contains links to other resources, e.g. https://www.wikidata.org/wiki/Q163709.

4. Methodological Approach

In Archaeology, relational database management systems are used almost exclusively for storing information in databases, where it is very common to use desktop applications like FileMaker and Microsoft Access. Older digital research data is often just structured as spreadsheets. Therefore, the data integration approach has to be very flexible to allow an integration of all kinds of datasets, databases and information systems. Various different kinds of structured data could be integrated by using XML as a common minimal standard, which could then be processed further. This applies to desktop databases, which allow exporting to XML, like FileMaker, Online Databases with interfaces for exporting data into XML and even all other kinds of exported CSV datasets, which could then be converted to XML using appropriate tools.

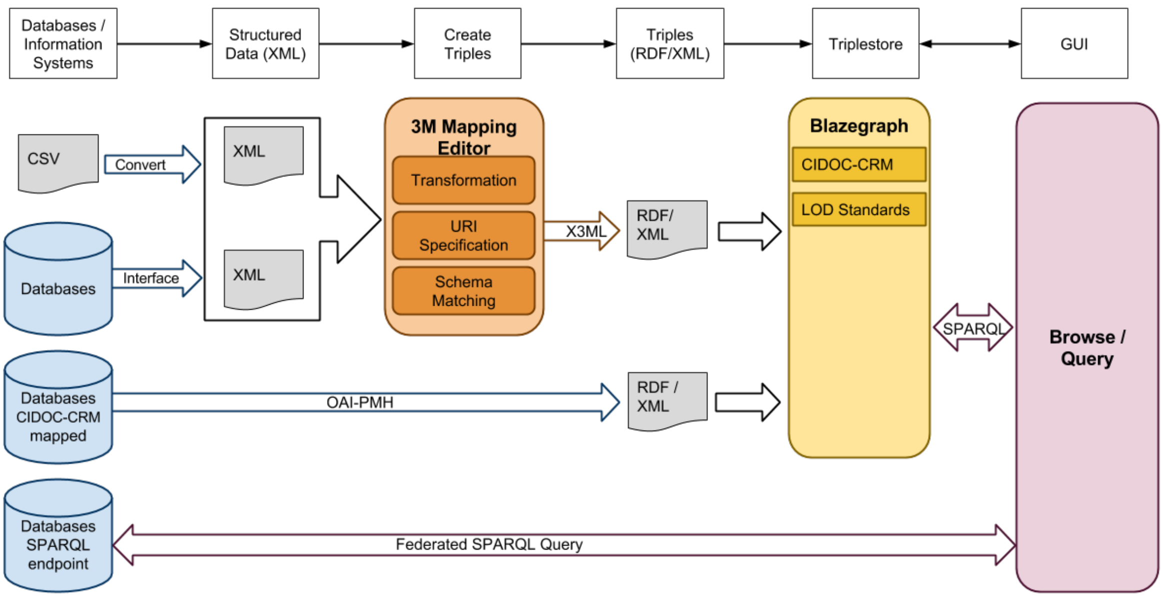


Figure 3 An overview of the data flow for data integration of different datasets

To make this data available for the Semantic Web, one needs to transform the XML data into triples in RDF/XML. Therefore, a sufficient mapping specification is needed to support the transformation of each distinct schema of an archaeological dataset (source schema) into the ARIADNE RM (target schema). It is crucial that during the transformation, the information encoded in the source database is not lost and the initial "meaning" is preserved as much as possible. In several cases, 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 carried out solely by technical people; it requires close collaboration between the domain experts, who possess the domain knowledge, and the IT experts, who possess the semantic world technical knowledge.

This complex process of mapping the schema of an archaeological database to a common coherent global ontology is guided by using a mapping and transformation tool like the 3M Mapping Memory Manager (http://www.ics.forth.gr/isl/3M/) maintained by FORTH. This online open source tool is used to describe and edit schema mappings in a human and machine-readable way. After importing the XML data into 3M, three more steps are necessary:

- Schema mapping: in this first step mappings are produced from attributes in the source schema (dataset) to classes in the target schema (CIDOC-CRM). Therefore, domain knowledge about the explicit and implicit semantics of the source information model, as well as technical understanding of the target ontology and the system, is needed.
- URI specification: the next step requires assigning an appropriate URI to each resource in the dataset. In an ideal world, you would be able to link directly to stable URIs of the integrated data source, if the data is available online. In the case of data available only offline, like the iDAI.field data, it is only possible to create pseudo-URIs.
- Transformation: in this final step the transformation of every source dataset record to a set of appropriate RDF (https://www.w3.org/RDF/) triples is carried out.

After these steps are applied, the RDF triples can be imported into a triplestore. We tried different triplestores, choosing Blazegraph (https://www.blazegraph.com/) in the end, as it provides very good performance and usability even with more than several million triples. The data could be imported using console scripts, which should be applied for bigger datasets, and a web based GUI-based import. More elaborate information systems provide an interface (i.e. OAI-PMH) mapped to CIDOC-CRM, like Arachne, that allows exporting the queried data in RDF/XML. These derived files could then also be imported into the triplestore, to unify the datasets into one resource, which could be accessed by a SPARQL endpoint as an open interface.

Once the data is imported into the triplestore, this could be enriched using Linked Open Data web resources, as described in chapter 3. Linked Data Enrichment Technically, the enrichment works by mapping the source terminology to the target terminology, which should be referenced by URIs of web resources. The triples themselves could be changed directly within the triplestore using SPARQL Update (https://www.w3.org/TR/sparql11-update/). Another approach is the manipulation of the triples in RDF/XML before importing by using XSLT, or another programming language or by just using a text editor with REGEX functionality.

In the final user interface, which builds upon the metaphacts platform, the user could access all integrated datasets within the Blazegraph triplestore via the SPARQL endpoint. External data could also be accessed by utilizing Federal SPARQL queries on these distributed triplestores, as long as the ontology and standards used overlap.

Overall, many different datasets with various preconditions were integrated with the procedure described above, showing the flexibility and the usefulness of the utilized tools.

Initial OEAW record (partial)

```xml
<COIN>
      <ID>626</ID>
      <COUNTRY_ID>1</COUNTRY_ID>
      <FIND_SPOT_ID>242</FIND_SPOT_ID>
      <FIND_MANNER_ID>2</FIND_MANNER_ID>
      <FIND_DATE>-</FIND_DATE>
      <WEIGHT>0.43</WEIGHT>
      <ISSUER_ID>243</ISSUER_ID>
      <DENOMINATION>239</DENOMINATION>
      ....
</COIN>
<DENOMINATION>
      <DEN_ID>239</DEN_ID>
      <DEN_NAME>Kls</DEN_NAME>
      <DEN_METAL>2</DEN_METAL>
</DENOMINATION>
<METAL>
      <MET_ID>2</MET_ID>
      <MET_NAME>AR</MET_NAME>
</METAL>
```

Transformed OEAW record (partial)

```xml
<crm:E22_Man-Made_Object
rdf:about="http://www.oeaw.ac.at/COIN/626">
<crm:P52_has_current_owner rdf:resource="http://www.oeaw.ac.at/"/>
<crm:P67i_is_referred_to_by rdf:resourse="http://registry.ariadne-
infrastructure.eu/ACDMdescription/doi%3A10.5878%2F002271"/>
<crm:P43_has_dimension>
   <crm:E54_Dimension rdf:about="urn:uuid:cc4ad6d4-656b-4....">
    <crm:P90_has_value>0.43</crm:P90_has_value>
    <crm:P91_has_unit rdf:resource="…/measurement units/gr"/>
</crm:P91_has_unit>
    <crm:P2_has_type rdf:resource="…/dimensions/weight"/>
<rdfs:label>weight of coin  626</rdfs:label>
   </crm:E54_Dimension>
  </crm:P43_has_dimension>
<crm:P45_consists_of>
  <skos:Concept rdf:about="http://www.oeaw.ac.at/material/2">
      <skos:prefLabel>AR</skos:prefLabel>
      <rdf:type rdf:resource="…/cidoc-crm/E57_Material"/>
   </skos:Concept>
 </crm:P45_consists_of>

</crm:E22_Man-Made_Object >
```

Figure 4 Example of transformed data from source schema (left) to target schema (right)

5. Technical Environment

X3ML mapping framework

The X3ML mapping framework (Marketakis et al., 2016, Minadakis et al., 2015) includes a mapping editor, a mapping specification and a mapping memory to accumulate knowledge and experience. It supports and guides the user during the mapping process. 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.

Blazegraph

Blazegraph is a graph database, which works as an RDF triplestore that natively supports the standardized query language SPARQL 1.1, RDFS and the Web Ontology Language (OWL). It is highly scalable, as it could be run with a standard CPU or, for enhanced speed, with a single GPU and even a GPU cluster. It is used as a backend, to store all the data as triples in a triplestore.

Metaphactory

The visualization of the scenarios are realized by the Metaphactory platform (http://metaphacts.com/). Metaphactory provides a basic semantic data integration by offering a wiki system with widgets, which access the triplestore Blazegraph in the backend. It is an open platform, which can be used for visualization of SPARQL queries, interaction with the knowledge graph and to construct user interfaces.

6. Scenario 1: Coins

Numismatics is a very traditional science with a lot of experience and early initiatives in standardization of the existing data (i.e. Bödefeld, Vacano 1978). In recent years, numismatics excels in terms of Linked Open Data in the Digital Humanities with a high grade of accessible datasets and standardized vocabulary. One major collaborative project is Nomisma.org (2016), supported by many institutions.

Nomisma.org serves as an authoritative resource in the field of numismatics. It collects and provides URIs to common numismatic concepts and terms. Furthermore, a whole ontology (Nomisma Ontology 2016) was created and is used to integrate the open available databases. The ontology provides an easy and understandable way for numismatists to describe their datasets, but as it is just limited to numismatics, it's very domain specific, if compared to the generic approach of CIDOC-CRM. Overall, numismatics provides a very good starting point for the item-level integration of archaeological datasets, as it is highly standardized and data is widely available to demonstrate the usefulness of using ontologies.

6.1. Description of Datasets

The dFMRO archive

Digitale Fundmünzen der Römischen Zeit in Österreich (dFMRO, 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 (Vondrovec 2007). Since the 1990s it documents coin-finds from the Celtic and Roman Period that have been published in various printed volumes of the FMRO (Fundmünzen der Römischen Zeit in Osterreich / 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 in the database were found in Austria and date from the Celtic and Roman period (actually the entire Antiquity); they were properly registered so no illegal finds are included, and most have already been published by the various projects of the FMRO. Since 2007, due to a former project collaboration, the database also lists coins found in Romania. These coins were published in "Colonia Ulpia Traiana Sarmizegetusa", the first volume of "Coins from Roman sites and collections of Roman coins from Romania", and represent an important part of the Austrian cultural heritage. The dFMRO archive was chosen as the first hands-on exercise to map a relational database schema to CIDOC CRM, since it represents a large class of well-defined traditional databases.

The CIDOC-CRM Mapping is available online in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping209

Numismatic archives from the COINS project

Another source of information we have taken into account comes from two numismatic archives already used within the COINS project. They include a set of 1670 numismatic records coming from the Cambridge Fitzwilliam Museum archive (FWM) and a set of 630 records coming from the database of the Soprintendenza Archaeologica di Roma (SAR).

The COINS project (Combat On-line Illegal Numismatic Sales) aimed at providing a substantial contribution to the fight against illegal trade and theft of coins by using state-of-art Information Technology. The project developed standardized inventories by integrating legacy archives encoded in different formats and using different languages. The creation of a reference collection of Roman and Greek coins was also one of the most relevant outcomes of the project.

The FWM archive: The FWM subset comes from the Department of Coins and Medals of the Fitzwilliam Museum Database, which records information on medals and coins of different types and age, discovered during excavations or coming from various acquisitions or donations. It is currently kept by the FW museum. Relevant fields used by FWM archive include: coin maker, production location, mint, coin type, category, coin name, inscription, dimensions, production technique, and references to images. Databases also include notes concerning record creation and modification, date and time, museum acquisition information.

The CIDOC-CRM Mapping is online available in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping484

The SAR archive: The SAR database (originally a Microsoft Access DB) was created to catalogue the archaeological finds of a monetary type managed by the Archaeological Superintendence of Rome, coming from public and private collections, and from archaeological excavations made in the city of Rome and its immediate surroundings. The main purpose of the archive is to record and provide the date, the accurate descriptions (by indicating the precise origin or place of issue) and physical characteristics of the various coins. Further, it also shows the conditions of discovery (excavation, auction, seizure, donation, etc.), the state of preservation and the current location (museum, superintendencies, collections and so on).

The SAR database, in addition to the FWM fields reported above, also provides information concerning the physical features and conditions of the coins, the region in which a specific coin was minted (apart the exact location), specific information on the chronology (i.e. the age, century or period during which coin minting took place), obverse/reverse inscriptions of iconography and the current location of the specific exemplar the record refers to.

The CIDOC-CRM Mapping is online available in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping168

Arachne

Arachne16 is the central object database of the German Archaeological Institute (DAI). Currently it contains more than 3,700,000 digital objects with corresponding metadata and over 300,000 highly structured descriptions of artefacts of archaeological interest. Arachne also allows research projects to store, manage and publish their data in the available online catalogues. There are currently 485 coins from the digitized museum inventory and research project data that present varying levels of metadata quality. Some are of excellent quality, such as the 107 coins with figures related to harbours found in the DFG funded ”SPP-Häfen”. Besides a detailed description, these provide extensive information about the bibliographic references and dating opinions of different authors. Arachne’s data is mapped to CIDOC-CRM and is provided as RDF/XML by an OAI PMH interface (http://arachne.uni-koeln.de/OAI-PMH/oai-pmh.xml?verb=Identify), which was used to harvest the above-mentioned data.

iDAI.field - Pergamon Project

Since its first usage in 2005, the iDAI.field database has been adopted by around 35 archaeological projects. Its main focus is to provide a field research database for archaeological research activities. The modular system also contains a find module with specific attributes for coins found during excavations or surveys. In a first integration test, 517 coins of the Pergamon project were used with detailed information about the archaeological context.

The CIDOC-CRM Mapping is online available in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping9

MuseiD-Italia collections

We also investigated the possibility to integrate the collections of MuseiDItalia, the digital library integrated in CulturaItalia. The data are in CIDOC CRM form and can be extracted via the OAI-PMH of the repository. MuseiDItalia includes several collections of coins from Italian museums such as:

- Museo archeologico nazionale di Venezia
    - Il medagliere: serie romana - imitazioni o falsificazioni moderne, 86 coins
    - Il medagliere: serie greca e bizantina, 758 coins
    - Il medagliere: serie romana e barbarica, 2307 coins
- Museo archeologico nazionale di Crotone, Reperti archeologici e Numismatica, 31 coins
- Collezione Museo Archeologico Nazionale - Reggio di Calabria, 136 coins
- Collezione numismatica Museo Archeologico Nazionale di Altamura, 99 coins
- 3008 coins from Regione Umbria

The major problem with these data is that they don’t contain appropriate labels in order to be displayed. However, if they undergo a pre-processing stage, they can be integrated perfectly well with other datasets.

Natural Language Processing data

As a further experiment in the coin integration use case, we investigated the integration of natural language processing data (described in D16.4). 37 records of NLP data from Heslington East Excavation Archive were integrated with the rest of the coin datasets.

ACDM data

The ARIADNE Registry is the main resource of the project, where all the available collections and/or datasets are described. So, in order to have a complete description of the coin datasets, we included the information contained in the registry.

6.2. Knowledge Graph

The datasets that we adopted in this use case have several differences concerning the origin, language, purpose of creation and use. Having been created by various institutions and for different purposes, they have quite different data structures, despite the similarity in content. The databases of the SAR and the dFMRO, for instance, were created with the purpose of documenting archaeological discoveries, which occurred during excavations or surveys, and contain many fields reporting information on provenance and discovery conditions. FWM, on the other side, is a museum database whose sole purpose is to catalogue acquisition and inventory data of objects owned and stored by the museum itself, regardless of the archaeological provenance conditions. The mapping of the coin datasets started with the dFMRO archive, which was chosen as the first hands-on exercise to map a relational database schema to CIDOC CRM, since it represents a large class of well-defined traditional databases. In close cooperation with the domain experts we tried to identify information that was implicit, hidden in forms, in user interface fields or was known only by them. A detailed description of the mapping of the dFMRO archive to CIDOC CRM was presented during the CAA2015 conference (Doerr et Al. 2015). The dFMRO mapping was used as a guide for the mapping of the SAR and FWM datasets. The records of the FWM archive contain fields with condensed information that needs to be pre-processed and normalized before it can be mapped to CIDOC CRM. For example, all the information concerning the dimensions of a coin (height, width, weight) is encoded in one field:

<Dimension> image(height), 22, mmimage(width), 20, mmweight, 3.74 </Dimension>

and needs to be normalized before the actual transformation takes place. It is worth mentioning that the mapping of a schema to CIDOC CRM is not necessarily unique. There may be different ways of approaching the problem, all correct. However, what is individually correct may turn out to be problematic if considered in the context of a larger process. For example, in the dFMRO and SAR mappings, the coin denomination was mapped as an E55 Type, while in the iDAIfield it was mapped as an E54 Dimension. Conceptually both approaches are correct, but their coexistence in the same process is clearly problematic. Rather than imposing a unique style, we have chosen to reconcile such differences at the query level. The dates are also a crucial point in the integration of the datasets. Different formats and approaches may have been used to encode temporal information in the source databases. To mention just a simple issue, the value zero is used as a date in some of the datasets, possibly with different meanings: for instance, such a value might indicate the year in which Jesus was born, or the fact that the year is unknown, or not recorded. This poses several problems. First, zero is not a valid date in RDF (or in the underlying XML type system), so the value has to be transformed into a valid date. But in order to carry out this transformation, it is important to clarify the semantics of the zero value in each dataset.

6.3. Query Possibilities

The ultimate goal of the integration of the diverse coin datasets is to create an environment where users will be able to specify queries that will be evaluated on the common aggregated repository and will be able to combine results coming from the different datasets. We have identified the following research questions:

- 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

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.

Combinations of the above queries can be found useful by the researchers of the numismatic community and our first experiments with such queries on the aggregated repository are quite promising. Our experimental aggregated repository contains 72 records (all Roman coins) of the dFMRO archive, 627 records (all Roman coins) of the SAR archive, 517 records (12 Roman coins 1 empty record) of the Pergamon archive (iDAIfield) and 1 record from MuseiD-Italia. The results of some simple queries can be seen in the following table:

Query	Total	dFMRO	SAR	Pergamon	MuseiD
Find all coins	1216	72	627	516	1
Find Roman coins	711	72	627	12	
Find bronze assarius	82	29	52	1	
Find bronze coins	676	50	270	355	1
Find bronze sextans	47		46		
Find coins produced in the year -32	22	4	18		

Table 1 Results from example queries in the Coins scenario

The query possibilities described above use the item specific information that is contained in the datasets. However there is also the catalogue (registry) information, which is of significant importance.

The actual query power lies in combining item level with catalogue info:

- Find all bronze Antoninianus coins (item level info, retrieves datasets from multiple providers)
- Find the publishers of all collections that contain coins (catalogue info)
- Find the publishers of all collections that contain bronze Antoninianus (item level and catalogue info)

The coin experimental demonstrator proves that the integration of item level information with NLP data and catalogue information under a common conceptual schema provides a powerful tool that can answer complex queries on heterogeneous data at different levels of detail.

7. Scenario 2: Sculptures

The research on sculptures is very traditional and highly embedded in the context of museums. Even though the museum community under the hood of ICOM's "International Committee for Documentation"(http://network.icom.museum/cidoc/) has developed a lot of different metadata standards like LIDO or CIDOC-CRM for data exchange, there aren't many subject-specific standards established for the description of sculpture objects. In the following, the selected datasets are briefly described. Only databases with permissive licenses and where the data could be extracted with reasonable effort were taken into account.

7.1. Description of Datasets

Athenian Agora

The American School of Classical Studies at Athens started in 1931 systematically excavating the Athenian Agora. The database of the project is a single integrated online accessible archive (http://agora.ascsa.net/research?v=default) to all the different resource materials: the current digital-born excavation documentation, finds, architectural plans, drawings, reports and publication. Prior to that the documentation relied on four different analogue notebooks, which got digitized and annotated. All the objects are documented and mapped to the Dublin Core standard and allow a search and open access to over 355.000 objects created in 85 years of research activity.

The Solr (http://lucene.apache.org/solr/) powered online database offers an XML export, which was used to extract more than 1000 sculpture objects. These objects are highly contextualised with all other available sources, from which 1700 archaeological contexts and 56.000 bibliographic references to more than 240 documents were selected.

The CIDOC-CRM Mapping of the Agora Database are online available in the 3M Mapping Memory Manager:

- Sculptures: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping487
- Publications: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping511

iDAI.objects Arachne

iDAI.objects Arachne (https://arachne.dainst.org/), as the central object database of the German Archaeological Institute, provides researchers a free searching tool available online. Originally started in 1995 as a Filemaker database, it was intended as an image database for sculptures of the research archive for antique sculptures (Forschungsarchiv für Antike Plastik) of the University of Cologne (Scheding 2014, 38). In 2004 the database migrated to MySQL and extended its focus to all kinds of archaeological objects. Currently in its fourth version, the database holds more than 3.6

million objects with associated metadata. All 83.085 sculptures in Arachne are from Roman and Greek times, obtained mainly through research project data, digitization of archives and museum collections, like the complete Catalogue of Sculptures in the Antiquities Collection of the Berlin State Museums (Antikensammlung der Staatlichen Museen zu Berlin). The majority is of excellent quality, as the sculptures contained a detailed description and are contextualised by more than 1000 archaeological types and 88.000 bibliographic references with links to the literature catalogue Zenon (http://zenon.dainst.org/). Arachne's data is mapped to CIDOC-CRM and provided by an OAI PMH interface, which was used to harvest the above mentioned data in RDF/XML format: http://arachne.uni-koeln.de/OAI-PMH/oai-pmh.xml?verb=Identify

British Museum Collection Online

The British Museum Collection Online database currently provides access to around 2.28 million records. Digitization of the collection is still on going and will cover every object in the Museum collection with corresponding scientific and conservation metadata. The data is also exposed for the Semantic Web as a SPARQL service: http://collection.britishmuseum.org/, providing access to all objects in the Collection Online, with all object information mapped to CIDOC-CRM and an online published thesaurus. For this scenario, a total number of 52.000 openly available sculptures were used.

iDAI.field project "Chimtou"

Since 2005, the field research documentation system iDAI.field (Schäfer, 2011), based on the commercial software Filemaker, was developed and maintained by the German Archaeological Institute (DAI). It provides a very complex information model and supports modules for documentation of excavations, surveys, building studies, iconographic studies, material studies, restoration work and scientific sampling, which can be adapted for project-specific needs. Currently iDAI.field is used by around 35 field research projects from different universities and the DAI for a total of over 1 million datasets. For the current approach we included the dataset of the Chimtou project's database (https://www.dainst.org/projekt/-/project-display/33904), with around 500 stone objects, including sculptures, which were found in roughly 120 archaeological contexts. The data was exported by using FileMaker's export functionality, which supports the XML format.

The CIDOC-CRM Mapping of the iDAI.field is online available in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping303

Gazetteer of Stone Quarries in the Roman World

For our second use case we added a list of marble quarries to our data. Since we also extracted additional information, this data is a hybrid between the norm data sources described in chapter 3. Linked Data Enrichment and the sources of "pure" content highlighted in this chapter.

The study published by Ben Russell in 2013 is currently the most comprehensive collection of ancient marbles, and is available either online (http://www.romaneconomy.ox.ac.uk/databases/stone_quarries_database/ or as a PDF:

http://oxrep.classics.ox.ac.uk/docs/Stone_Quarries_Database.pdf). It is operated and hosted by the Oxford Roman Economy Project. All 794 quarries present in the datasets were included via the transformation of the PDF Document to structured data in the form of CSV, which was in turn transformed to XML. The information submitted consists of site, coordinates, location, country, and former affiliation to Roman Imperial provinces, a short declaration of the material and its characteristics, and the most important literary sources. This overview is suitable for an inventory and common basis if we also point out some minor problems in the terminology, e.g occasionally modern terms are favoured instead of missing ancient terms, as they are probably more well-known.

The CIDOC-CRM Mapping is online available in the 3M Mapping Memory Manager:

http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping647

7.2. Knowledge Graph

Once all the necessary standards (chapter 3. Linked Data Enrichment) are applied and the data is imported into the triplestore, the full knowledge graph becomes accessible for further investigations on the data.

CIDOC-CRM was chosen as a common ontology, as it is a widely accepted and used ontology in the realm of museums and is gaining more importance in digital humanities as well. There are also extensions, beside the CIDOC-CRM Core, which can be used if it is required by the data. We have used CRMsci and CRMarchaeo to describe scientific data acquisition and archaeological excavation processes. Furthermore, we used the Functional Requirements for Bibliographic Records (FRBR) ontology to describe the bibliographical records, and the W3C Basic Geo vocabulary (https://www.w3.org/2003/01/geo/) for simple geometry description.

Before applying a unified query over all integrated datasets, the mappings needed to be harmonized. Smaller differing mappings, like the assignment of measured object dimensions, needed to be harmonized. For the linked open data enrichment (chapter 3. Linked Data Enrichment) we chose Wikidata for actors, Getty AAT and Wikidata for archaeological terminology, Zenon for literature, and iDAI.gazetteer for places.

Figure 5 Graphical overview of the CIDOC-CRM mapped sculpture datasets.

7.3. Query Possibilities

Object centric query for similar objects

The object-centric query is about a fragmentary head of a satyr that was found during excavation at Chimtou in November 2010 (Scheding, 2013). First the object was registered in idai.field and afterwards in idai.objects Arachne (http://arachne.dainst.org/entity/2295540). The extent of the head, the designation as a satyr, the stylistic and iconographic creation and the material named as white marble are facts that could be useful in finding comparable objects. The aim is to learn more about the exemplar at hand, to get information about other exemplars and to get an overview of quarries where the same kind of marble (white marble) was produced. If, for example, our research question is to search for comparable satyr statues and the “type” that is shown by the fragmentary head, the parameters for a query can be defined as:

- Extent: under 1.15 m height
- Material: white marble
- Term: satyr head

For the technical fulfilment of the query it was necessary to enrich the triples within the triplestore with Linked Open Data standards using SPARQL Update. All terms for the material were linked accordingly to appropriate Getty AAT terms by creating new triples, while the actors represented on the sculpture were normalized by using Wikidata as a common standard. For example:

<http://arachne.uni-koeln.de/entity/1092353> crm:P45_consists_of <aat:300011599>

<http://arachne.uni-koeln.de/entity/1092353> crm:P62_depicts <wd:Q163709>

metaphacts sparql Account

Object: Head of a Satyr

Properties

Filter Results

Predicate	Object
P102_has_title	Kopf eines Satyrs Typus Ludovisi
P102_has_title	title
P108i_was_produced_by	t1465624
P2_has_type	objekt
P2_has_type	objektplastik
P2_has_type	1242717
P2_has_type	300047090
P2_has_type	t1464930
P2_has_type	t1464949
P2_has_type	t1465289

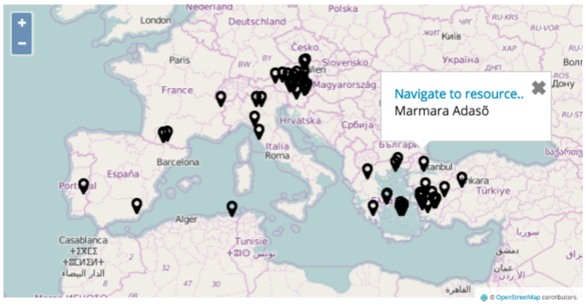
« 1 2 3 4 5 »

Show Possible Quarries

- Only Quarries, were White Marble is produced.

```
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX aat: <http://vocab.getty.edu/aat/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>

SELECT ?lat ?lng ?description ?link
WHERE {
        ?site crm:P45_consists_of ?material;
          rdf:type crm:E25_Man-Made_Feature;
          crm:P53_has_former_or_current_location ?location .
        ?location geo:long ?lng;
              geo:lat ?lat;
              crm:P1_is_identified_by ?description .
        ?material crm:P127_has_broader_term aat:300011571.
    FILTER NOT EXISTS { ?description rdf:type crm:E47_Spatial_Coordinates }
}
```



Show Similar Objects

- Material Height below 115cm
- Depicts a Satyr
- Is made out of white marmor including childs

```
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX aat: <http://vocab.getty.edu/aat/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX wd: <http://www.wikidata.org/entity/>

SELECT DISTINCT ?object ?materialLabel (STR(?height) AS ?height_cm) ?title
WHERE {{
      ?object crm:P62_depicts wd:Q163709;
              crm:P45_consists_of aat:300011571; # white marble objects
              crm:P43_has_dimension ?dimension;
              crm:P102_has_title ?title .
      aat:300011571 skos:prefLabel ?materialLabel .
      ?dimension crm:P2_has_type wd:Q208826;
                 crm:P90_has_value ?height .
      FILTER(xsd:float(STR(?height)) <= 115) # maximum 115cm object height
      FILTER(lang(?materialLabel) = "en") # only english material label
      FILTER NOT EXISTS { ?title rdf:type crm:E35_Title }
  } UNION {
      ?object crm:P62_depicts wd:Q163709;
              crm:P45_consists_of ?material;
              crm:P43_has_dimension ?dimension;
              crm:P102_has_title ?title .
      ?material skos:broaderTransitive aat:300011571 . # white marble childs
      ?material skos:prefLabel ?materialLabel .
      ?dimension crm:P2_has_type wd:Q208826;
                 crm:P90_has_value ?height .
      FILTER(xsd:float(STR(?height)) <= 115)
      FILTER(lang(?materialLabel) = "en")
      FILTER NOT EXISTS { ?title rdf:type crm:E35_Title }
  }}
```

Filter Results

Object	Title	Height_cm
1063304	Hermenbüste eines Satyrmädchens	25
1069529	Oberkörper eines Satyrmädchens	48
1078623	Satyr - Statue	40
1072825	Satyr - Herme	27
1080394	Torso eines Satyr	78
1089335	Kopf eines Satyrn	14
1124151	Kopf eines Satyrn	40
1124238	Kopf eines Satyrn	20
1092747	Herme eines Satyrn	41
1091592	Torso eines Satyrn oder eines Herakles	43

« 1 2 3 »

Figure 6 User interface for object centric queries over several databases

After these standards were applied, it was possible to apply the same query for all integrated databases: this was the basis of a prototypical user interface, see Figure 6 User interface for object centric queries over several databases

The results provide us with several comparable obtained, reconstructed or fragmentary statues that already point to the fact that the satyr head from Chimtou can be assigned the type “Ludovisi”. The next step would be an evaluation of the most suitable objects by further research filters in a new query, or the matching of “pairs, copies and correspondences” to evaluate the most likely type, which in the context of the satyr found at Chimtou is the Ludovisi type, with many comparable exemplars from numerous countries - especially Italy, but also from the Greek island of Rhodes.

Enhancing literature research

Literature references are essential for research purposes. Normal bibliographical systems offer at best a thesaurus to tag the bibliographical dataset. This approach is limited by the extent of the used thesaurus, and for the specific use case one could only search for literature containing information about sculptures. The use of the knowledge graph could enhance literature queries and further investigation, as it allows a more precise query.

To exemplify this approach, a closer look at the marble is needed. In this case there is an important opportunity that is offered by queries relating to quarries. If the query deals with the question of an overall spectrum of one single marble, new parameters could be formulated. The Pentelic marble (http://archwort.dainst.org/de/term/6273) was chosen, as it is one of the most popular ancient marbles with a white surface. To get an overview of statues made of this marble, the following parameters could be formulated:

- Material: Pentelic marble
- Extent: bigger than 0,70 m
- Term: statue

In the quarry use case the knowledge graph is used to show all literature containing information about sculptures made of Pentelic marble, with 393 hits. Aggregate functions, like the counting of documented sculptures made out of Pentelic marble for the specific literature, offer a further hint about which literature might deliver the best results.

The query delivers many objects in an overview from every database that includes Pentelic marble. It becomes obvious that Pentelic marble has been used for a long time in Antiquity to produce statues and sculpture of every kind, mostly by the Athenian workshops. This can be a starting point for further queries.

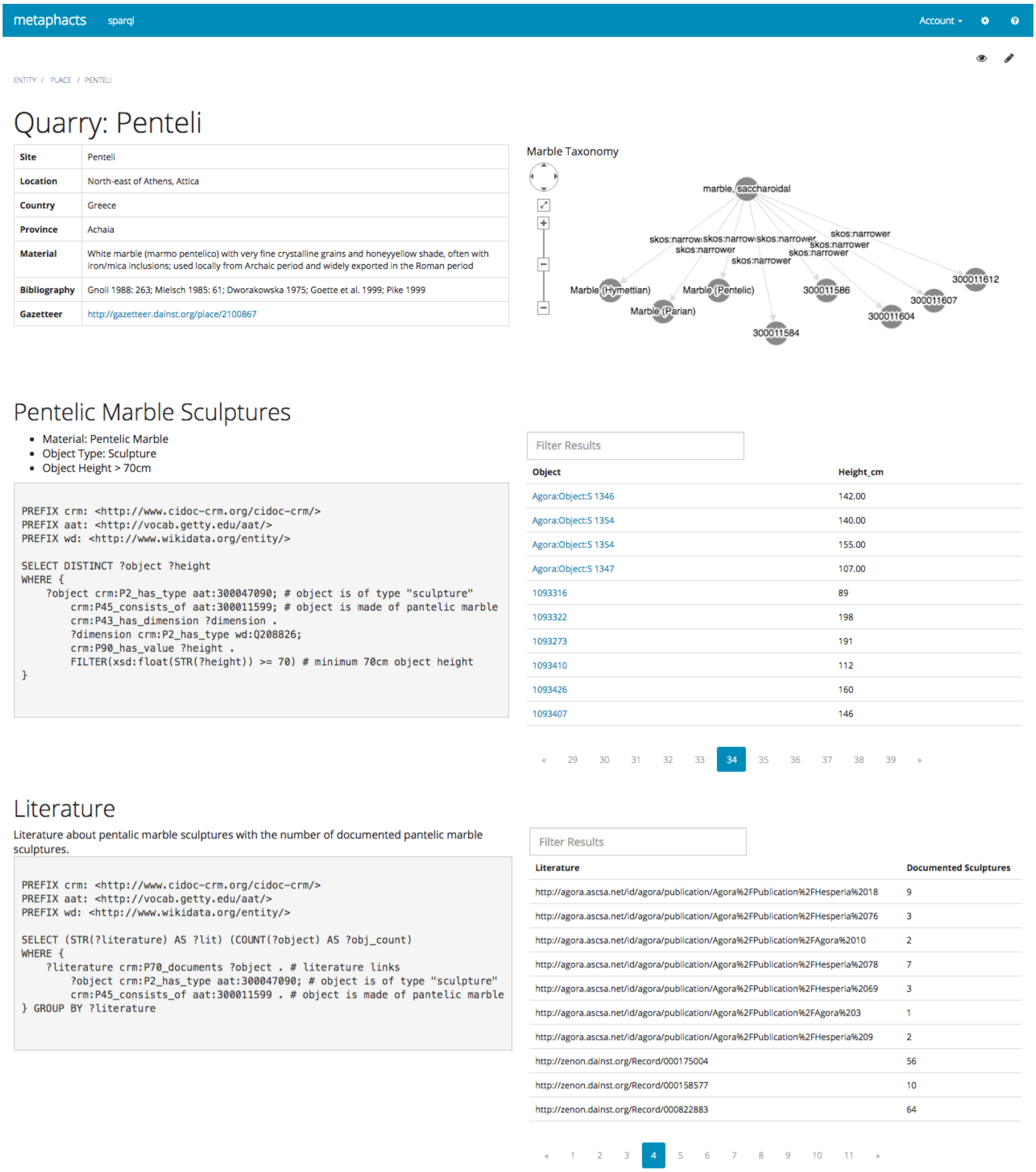


Figure 7 Searching for statues made of Pentelic marble

Bringing the data together faces a few challenges. For example, there are some remarkable differences in the terminology of the different datasets. We want to point out two problems we have observed. For ancient marbles there are not many differences that cause problems, but the huge quantity of marbles in the database are described using the modern and ancient terms alternatively. Also, while the Getty AAT and iDai.vocab provide useful categorized and multilingual overviews of many marbles, the British Museum Collection notably mentions only the parameter "marble" without further details. For example, a query with the parameters "sculpture" and "giallo antico" will deliver matches in every dataset, but if we are using "marmor numidicum" - the literary-based ancient name - instead of "giallo antico", the Arachne can cope with it but we won't get the data from the British Museum Collection. The same heterogeneous approach of the datasets is already present in the archaeological research literature. An already widely accepted collection of ancient marbles was recently published by Ben Russell in his "Gazetteer of Stone Quarries in the Roman World". However, a problematic aspect of Russell's study is the occasional absence of ancient names for some marbles that are known from the literary sources, while other studies that were fundamental for the datasets in the last decades consequently used this technical terminology (For instance Gnoli, 1971; Schneider, 1986; Maischberger, 1997). Thus we have a serious inconsistency in older and younger datasets.

These differences can cause huge problems in a query that will sometimes deliver faulty and incomplete results. It becomes apparent that there is a need for common standards and/or agreements about the archaeological terminology and appropriate tools to assist the mapping between different terminologies.

8. Scenario 3: Buildings

The scenario describes the application of the ontological models CRMba and CRMarchaeo to model an unstructured dataset about the excavation of a prehistoric archaeological site, in Crete. The two CIDOC CRM extensions were developed within ARIADNE to cover the specific needs of the archaeological investigation and the documentation of built structures, and to help archaeologists examine and understand the complex relationships between all entities and activities related to it. Indeed, archaeological investigations are not only focused on exploring the subsurface deposits of an area, but they also include the study of buildings and other built structures, which can be found standing on the soil surface or as fragments underneath.

8.1. Description of Datasets

Zominthos is a prehistoric archaeological site located on the north slope of Psiloritis, the highest mountain on the island of Crete, Greece - at an altitude of 1187 m above sea level. It lies between the modern village of Anogeia and the Idaean Cave, one of the most important sacred caves of Crete. It is the largest known residential center of Minoan times on the mountains. It was discovered in 1982 by professor Yiannis Sakellarakis. The excavations revealed a building (Figure 8 Central Building of the minoan settlement "Zominthos"

(http://www.efsyn.gr/arthro/galazia-kokkini-kai-leyki-zominthos)), conventionally referred as the "Central Building" and an extended Minoan settlement. Zominthos was occupied from the Minoan period to modern times.

Figure 8 Central Building of the minoan settlement "Zominthos" (http://www.efsyn.gr/arthro/galazia-kokkini-kai-leyki-zominthos)

The "Central Building" covers an area of 1,600 sq m. It is one of the most well preserved Minoan buildings, with walls standing at a height over 2,50 m, with at least 60 rooms on the ground floor, upper storey, staircases, paved floors, corridors, courtyards, windows, entrances, lightwell, storages, workshops, decorated private rooms with frescoes, ritual rooms with columns, drainage system, etc. It was built of large blocks of local limestone. During the excavations, a large number of masterpieces were revealed: bronze figurines, bronze cups, daggers, elaborate pins and pendants made of bone, ivory and bronze, jewelry with semi-precious stones or bone, seals, rhyta in the form of animals, chalices with floral motifs, cups, jars, pithoi, etc. The main phase is dated at the Late Minoan IA period (1700-1600 BC). It was destroyed by an earthquake around 1600 B.C., which resulted in a fire that devastated the vast majority of building's construction elements and materials, like collapsed parts of the upper storey fallen paved floors, large amounts of stone slabs, burnt wooden beam, cane fragments, mortar, frescoes, earth clay etc. (Sapouna-Sakellaraki, 2013).

The excavation is ongoing under the direction of Dr. Sapouna-Sakellaraki Efi and under the auspices of the Archaeological Society at Athens.

The documentation produced since 2005 consists of the archaeologist's excavation diaries, reports, photos and maps. All information is locally digitally stored and structured in folders. Some of the content is available online, published on the "interactive archaeology" website and in other articles and proceedings. Since the excavation is on going, the final publication of the excavation is not ready yet.

In the framework of a study about the construction methods and the materials of the Central Building, the possibility of using SYNTHESIS, a cultural information system developed by FORTH, is under discussion. The information system supports scientific and administrative documentation of museum objects and monuments, and can be extended to enable the documentation of archaeological sites and built structures.

8.2. Knowledge Graph

For the aim of this case study we selected a subset of information, which we modelled using CIDOC CRM core and its extensions, mainly CRMba, CRMarchaeo, and CRMsci. We used the 3M editor to edit the schema mapping. The graph representing the mapping of the selected subset (Room 37) to CIDOC CRM core, CRMba and CRMarchaeo is shown below (Fig.10).

Figure 9 Room 37, Central Building, Zominthos (http://interactive.archaeology.org/zominthos/2016/07/field-notes-2016-week-1/)

ROOM 37

Room 37 (Figure 9 Room 37, Central Building, Zominthos

(http://interactive.archaeology.org/zominthos/2016/07/field-notes-2016-week-1/)) is located in the southwest part of the Central Building. The excavation of one of the upper floors was completed. It contained mostly lepidha-earth and fallen stones. In the northeast corner, the removal of a clay structure from the upper storey was also continued. A concentrated mass of burnt clay-earth construction was uncovered. In the southeast corner a group of stones may belong to the southern wall of the room, which wasn't found during the previous excavation period. The southern edge of the Eastern Wall was also revealed and the works extended to the opening between the Rooms 37 and 39.

Several animal bones, some conical cups, two whetstones and a sealstone, from steatite with an carved iconographic theme, small pieces of burnt wood and some small groups of ceramic pots, two of which were found laying on schist slabs of the upper storey's paved floor, were collected.

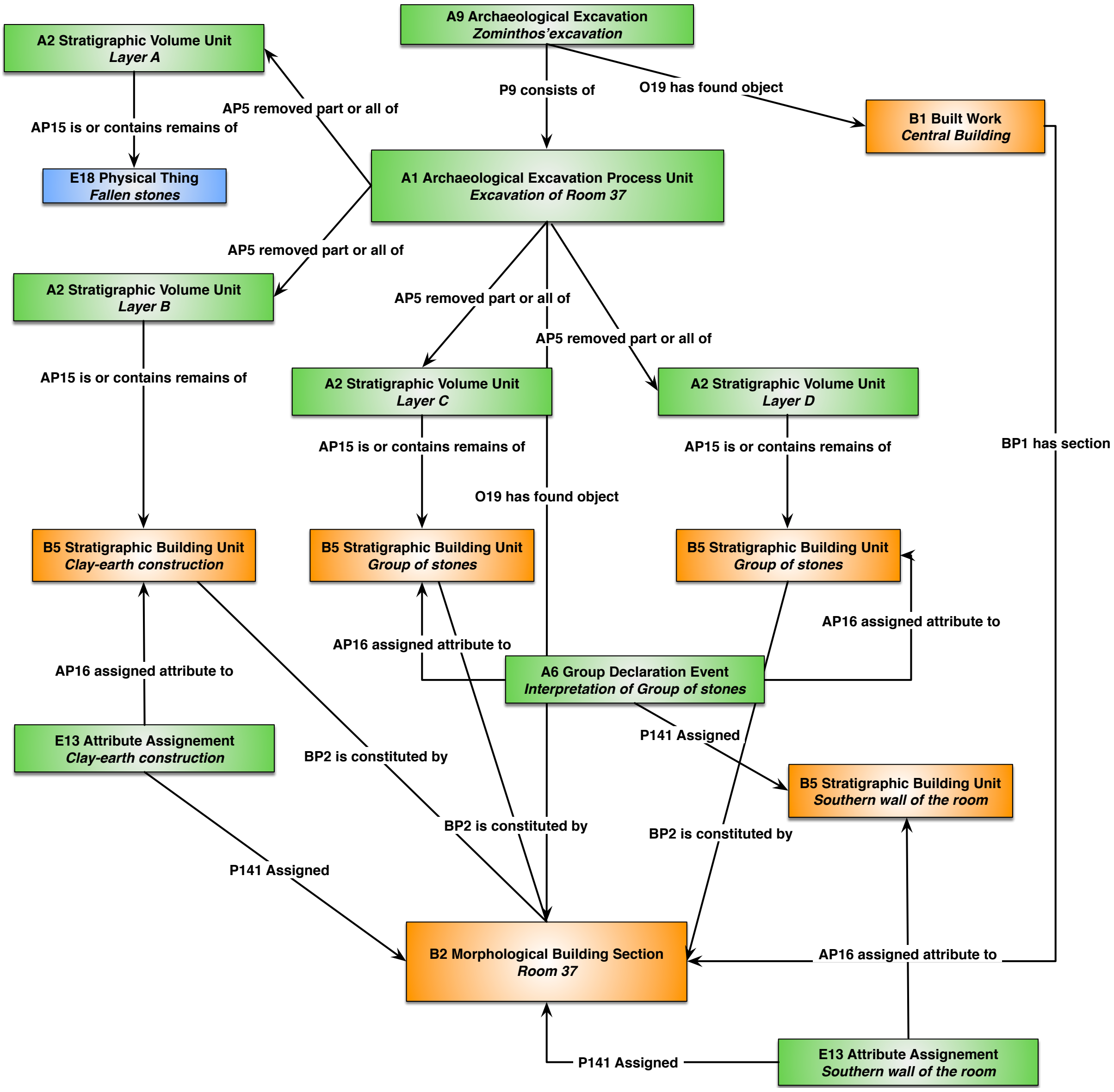


Figure 10 Graph of the mapping of Room 37 information to CIDOC CRM

9. Scenario 4: Faunal Remains

This scenario evaluates the potential of integrating different scientific datasets in the domain of archaeometry. Two different zooarchaeology datasets, which are published by the data centres "Archaeology Data Service" in York, and "IANUS" at the DAI in Berlin, were chosen for integration. They both focus on animal bones in archaeological contexts.

Zooarchaeology has a long experience in sharing their datasets and articles in community portals like Bone Commons (2016), which launched initially in 2006, the Zooarchaeology social network, the ZOOARCH email list, and so on (Kansa et Al 2001, 185ff). The discipline serves as a useful case study, as the terminology used is highly standardized, its materials and methodologies are global in scope and a lot of research questions can often be answered only by taking multiple datasets into account.

9.1. Description of Datasets

Animal Bone Evidence from Southern England

The "A Review of Animal Bone Evidence from Southern England" project, funded by English Heritage, aimed to review the animal bone evidence from Late Bronze Age-Late Iron Age sites from southern England. The Regional Review report (Hambleton 2008), for which this database serves as a freely available online appendix, provides a synthetic review of published faunal assemblages. Consequently, analyses (e.g. ageing, butchery, biometric data) focus on the exploitation and deposition of sheep, cattle, pigs, horses and dogs. Other taxa (e.g. wild mammals, birds, fish and amphibians) are also discussed.

A total of 108 site reports were recorded in the database. These 108 'site' records correspond to reports from excavations at 101 separate monument locations. These sites generated 154 distinct 'assemblage' records for faunal assemblages. Bibliographic references for all zooarchaeological reports reviewed are listed in the database.

This database was reviewed and published by the Archaeological Data Service: https://dx.doi.org/10.5284/1000102

The CIDOC-CRM Mapping is online available in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping764

Holocene history of wildlife in Europe

From 1994 - 1998 the project "Holozängeschichte der Tierwelt Europas" researched the development of the animal world in Europe from the Late Pleistocene to the Middle Ages. Three different institutes have collected and examined published data on the occurrence of wild and domestic animals (mammals, birds, amphibians, reptiles, fish) and pets from over 8000 archaeological complexes. The evaluation covered issues such as the influence of the climate on the animals in the transition from the Pleistocene to the Holocene, spatial distribution shifts of species and the influence of human intervention in nature. 4500 publications were examined and integrated in a database, which documents the faunal remains in more than 8500 archaeological sites.

This database was reviewed and published by the research data centre "IANUS": https://dx.doi.org/10.13149/001.mcus7z-2

The CIDOC-CRM Mapping is online available in the 3M Mapping Memory Manager: http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping765

9.2. Knowledge Graph

The knowledge graph derived from the mapping and alignment of the two datasets described above. To overcome the language barrier between the German and English datasets, the common standard Encyclopedia of Life (2016) was used, which provided biological definitions of species in a classification tree. As the thesaurus is not available in semantic web formats, the terms used were described with RDF/XML.

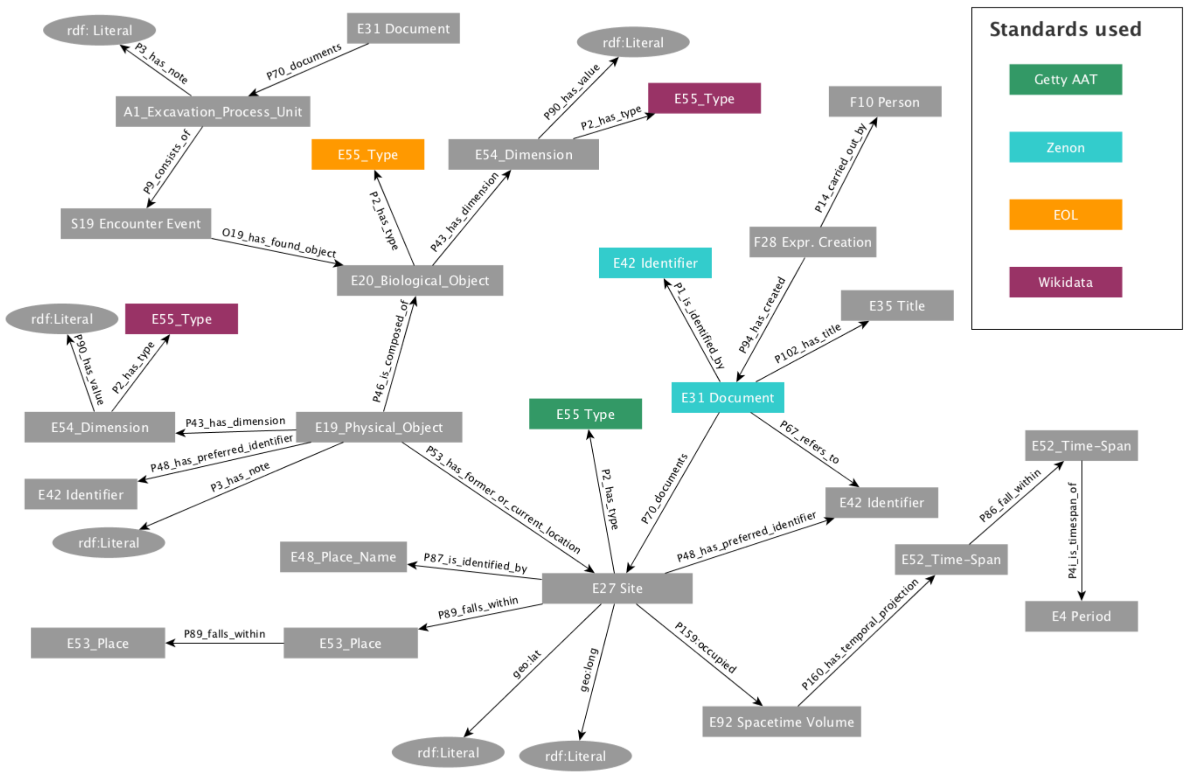


Figure 11 Graph representation of the CIDOC-CRM mapped datasets in the "Animal Remains" scenario

9.3. Query Possibilities

The integration of the zooarchaeology datasets leads to an environment where users are able to specify queries that will be run on a common aggregated repository and will combine the results coming from the different datasets.

The first presented example is a quite species-centric query, where all sites which do containing horse remains are shown. For a researcher, who is focussed on the distribution on a specific species, the literature references could be a desired starting point.

The second example shows the possibilities on performing statistics with the help of the query language SPARQL over a common mapped dataset. The bar chart shows the frequency distribution of selected species in the archaeological contexts.

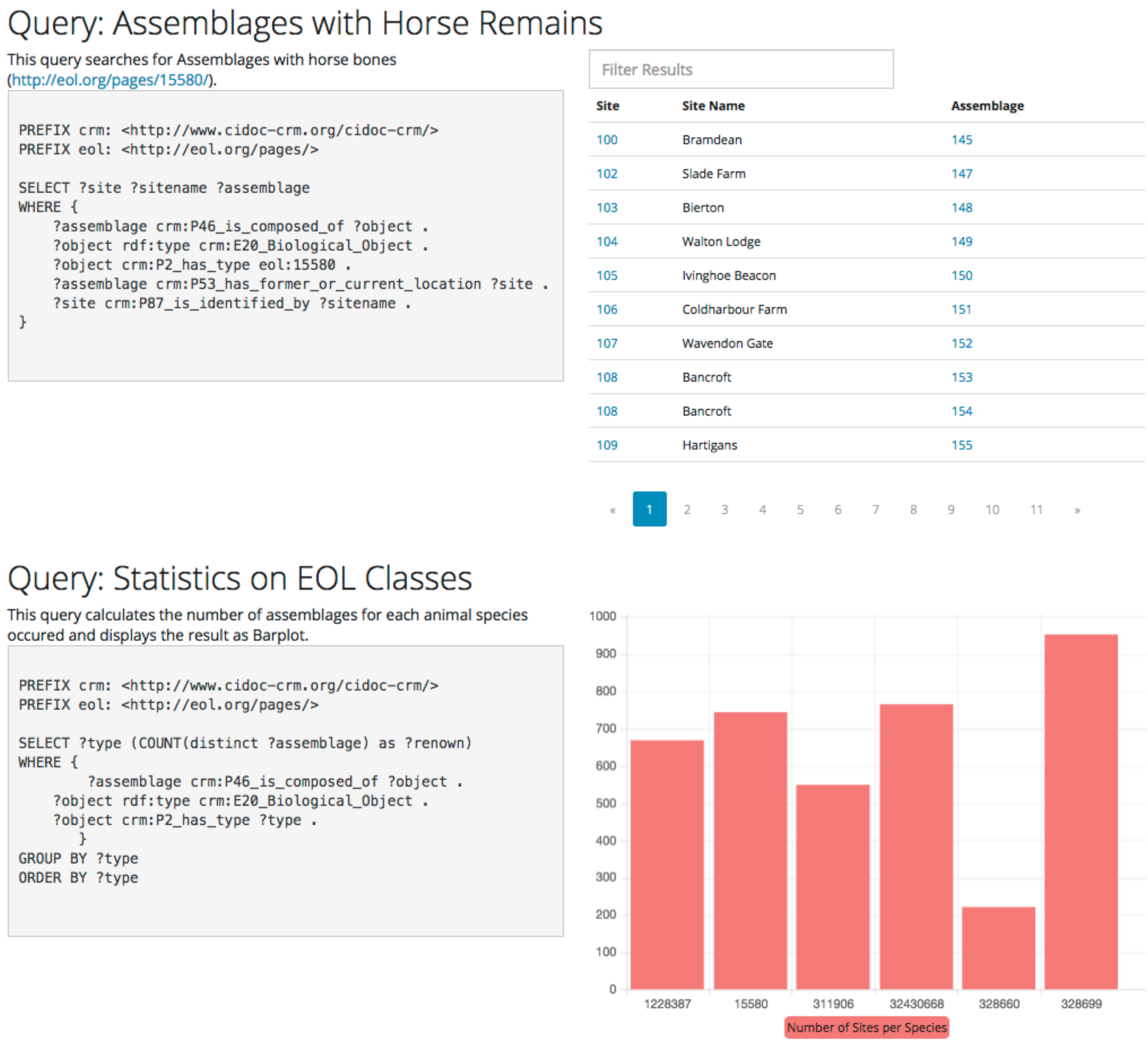


Figure 12 Visualization of example queries of the "Animal Remains" scenario.

10. Conclusions

The four pilot experiments show a successful integration of diverse datasets, where the methodology and the tools used have shown their practicability and usefulness even for very diverse prerequisite conditions of the datasets in terms of language, data structure and formats. With the data flow that we have described, the origin of the source data does not matter as long as it is possible to export the data in a structured dataset using a non-proprietary open format.

From an archaeological viewpoint, it becomes apparent that it is necessary to work with standards and declarations that are as similar as possible to each other, or at least to map their own value lists and thesauri to these international standards. Common standards and archaeological terminology are fundamental for data integration and comparability.

Combining open structured and semantically described data and transparent terminology gives rise to powerful tools in the daily work of an archaeologist working on excavations, survey data or museums and collections. The scenarios have illustrated the potential for an easy access to a huge amount on comparable data, which could be searched with one single query, which offers great opportunities to all archaeological and interdisciplinary areas of expertise.

Bibliography

Benecke, N., von den Driesch, A. and Heinrich, D., (2016) Holozängeschichte der Tierwelt Europas. Forschungsdatenzentrum IANUS, Berlin. (doi: 10.13149/001.mcus7z-2)

Bone Commons (2016) Bone Commons – An online community building and sharing resources for archaeozoology, Available at: http://alexandriaarchive.org/bonecommons/ (Accessed: 08.11.16)

Bödefeld, H. and Vacano, O. (1978) Elektronische Datenverarbeitung in der antiken Numismatik. Ein Projekt zur Erfassung griechischer Münztypen am Althistorischen Institut der Universität Düsseldorf. Chiron, 8:587–604, 1978.

Clark, M. K. (1935) A gazetteer of Roman remains in East Yorkshire, Leeds

Doerr, M. and Aspöck, E. (2015) Mapping archaeological databases to CIDOC-CRM, Available at: http://www.ics.forth.gr/_publications/DOERR_MappingCIDOC_1_wFig.pdf (Accessed: 03.11.2016)

Doerr, M., Theodoridou, M., Aspöck, E. and Masur, A. (2015) Mapping archaeological databases to cidoc-crm. In: Proc. 43rd Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2015 SIENA), Siena, Italy, April 2015. Archaeopress.

Encyclopedia of Life (2016) Encyclopedia of Life – Global access to knowledge about life on Earth, Available at: http://eol.org/ (Accessed: 08.11.2016)

Gnoli, R. (1971) Marmora Romana, Roma

Gruber, E. and Smith, T. J. (2015) Linked Open Greek Pottery, In: Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Oxford, 205-214, Available at: https://caa2014.sciencesconf.org/46119/document (Accessed: 03.11.2016)

Hambleton, E. (2009) *A Review of Animal Bone Evidence from Southern England*[data-set]. York: Archaeology Data Service [distributor] (doi:10.5284/1000102)

Kansa, E., Kansa, S., Watrall, E. (2011) Archaeology 2.0: New Approaches to Communication and Collaboration. Los Angeles.

Maischberger, M. (1997) Marmor in Rom. Anlieferung, Lager- und Werkplätze in der Kaiserzeit, Wiesbaden, Dr. Ludwig Reichert Verlag

Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., Flouris, G., & Doerr, M. (2016). X3ML Mapping Framework for Information Integration in Cultural Heritage and Beyond. International Journal on Digital Libraries (IJDL), Special Issue on "Extending, Mapping and Focusing the CIDOC CRM", 1-19.

Minadakis, N., Marketakis, Y., Kondylakis, H., Flouris, G., Theodoridou, M., de Jong, G., & Doerr, M. (2015). X3ML Framework: An effective suite for supporting data mappings, Extending. Proceedings of the Workshop on Extending, Mapping and Focusing the CRM co-located with 19th International Conference on Theory and Practice of Digital Libraries (2015), Poznań, Poland, September 17, 2015. CEUR Workshop Proceedings 1656, CEUR-WS.org 2016, 1-12.

Nomisma.org (2016) Nomisma.org. Available at: http://nomisma.org/ (Accessed: 03.11.2016)

Nomisma Ontology (2016) Nomisma Ontology. Available at: http://nomisma.org/ontology (Accessed: 03.11.2016)

Oldman, D., Doerr, M., de Jong, G., Norton, B. and Wikman (2014) T.. Realizing lessons of the last 20 years: A manifesto for data provisioning & aggregation services for the digital humanities. D-Lib Magazine, 20(7/8). doi:10.1045/july2014-oldman.

Ore, C., Doerr, M., Le Bœuf, P. and S. Stead (2015) Definition of the cidoc conceptual reference model version 6.1. Available at: http://www.cidoc-crm.org/Version/version-6.1 (Accessed: 03.11.2016)

Ronzino, P., (2015). CIDOC CRMba A CRM Extension for buildings archaeology information modelling. (Unpublished doctoral thesis). The Cyprus Institute, Nicosia, Cyprus

Ronzino, P., Niccolucci, F., Felicetti, A., & Doerr, M. (2016). CRMba a CRM extension for the documentation of standing buildings. International Journal on Digital Libraries. Vol. 17, n.1, March 2016, pp. 71-78

Ronzino, P., (2016). Harmonizing the CRMba and CRMarchaeo models. International Journal on Digital Libraries, Special Issue on Extending, Mapping and Focusing the CIDOC CRM.

Russell, B. (2013) Gazetteer of Stone Quarries in the Roman World, Available at: http://oxrep.classics.ox.ac.uk/databases/stone_quarries_database/0to20/ (Accessed: 03.11.2016)

Sapouna-Sakellaraki, E. 2013. The Wealth of Psiloritis. Zominthos. *Archaeology & Arts*. http://www.archaeology.wiki/blog/2013/11/18/the-wealth-of-psiloritis/, Accessed 25 October 2016.

Schäfer, F. (2011) "iDAI.field - Ein modulares Dokumentationssystem für archäologische Feldforschungsprojekte." Köln und Bonner Archaeologica 1: 201-208

Scheding, P. (2013) Der Kopf eines Satyrs Typus „Ludovisi" aus Simitthus (Chimtou, Tunesien), Kölner Jahrbuch für Vor- und Frühgeschichte 46, Köln, 133-140

Schneider, R.-M. (1986) Bunte Barbaren, Worms, Wernersche Verlagsgesellschaft

Thacker, J. (2000) Standardization and Libraries, In: Study on International Sharing of Japanese Scholarly Information

K. Tzompanaki, K. and Doerr, M. (2012) Fundamental categories and relationships for intuitive querying CIDOC-CRM based repositories. Technical Report TR-429, ICS-FORTH.

Tzompanaki, K., Doerr, M., Theodoridou, M. and Fundulaki, I. (2013) Reasoning based on property propagation on CIDOC-CRM and CRMdig based repositories. Online Proceedings for Scientific Workshops, 2013.

Vondrovec, K. (2007) digitale Fundmünzen der Römischen Zeit in Österreich. Available at: http://www.oeaw.ac.at/antike/index.php?id=358 (Accessed: 03.11.2016)

Appendix A

Link collection to all CIDOC-CRM mappings in the 3M Mapping Memory Manager sorted by scenarios.

Scenario 1: Coins

The Fitzwilliam Museum, Coins and Medals:
http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping484

Archaeological Superintendence of Rome Coins Database:
http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping168

Mapping of iDAIfield_Muenzen to CIDOC-CRM:
http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping9

OEAW Coins DB:
http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping209

Scenario 2: Sculptures

Athenian Agora DB:

- Sculptures:
  http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping487
- Publications:
  http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping511

iDAI.field Chimtou:

http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping303

Gazetteer of Stone Quarries in the Roman World:

http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping647

Scenario 4: Animal Remains

Holozängeschichte der Tierwelt Europas:
http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping765

Animal Bone Evidence from Southern England:
http://139.91.183.3/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping764

Appendix B

The ARIADNE Reference Model version 1.0 in detail can be found at the ARIADNE website:

http://www.ARIADNE-infrastructure.eu/Resources/ARIADNE-Reference-Model

CRMinf version 0.7

Reference document: http://www.ics.forth.gr/isl/CRMext/CRMinf/docs/CRMinf-0.7.pdf

RDFS encoding: http://www.ics.forth.gr/isl/CRMext/CRMinf_v0.7.rdfs

CRMsci version 1.2.3

Reference document: http://www.ics.forth.gr/isl/CRMext/CRMsci/docs/CRMsci1.2.3.pdf

RDFS encoding: http://www.ics.forth.gr/isl/CRMext/CRMsci_v1.2.3.rdfs

CRMba version 1.4

Reference document: http://www.ics.forth.gr/isl/CRMext/CRMba/docs/CRMba_v1.4.pdf

RDFS encoding: http://www.ics.forth.gr/isl/CRMext/CRMba_v1.4.rdfs

CRMarchaeo version 1.4

Reference document: http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo_v1.4.pdf

RDFS encoding: http://www.ics.forth.gr/isl/CRMext/CRMarchaeo_v1.4.rdfs

CRMdig version 3.2

Reference document: http://www.ics.forth.gr/isl/CRMext/CRMdig/docs/CRMdig3.2.pdf

RDFS encoding: http://www.ics.forth.gr/isl/CRMext/CRMdig_v3.2.rdfs

CRMgeo version 1.2

Reference document: http://www.ics.forth.gr/isl/CRMext/CRMgeo/docs/TR435-CRMgeo.pdf

RDFS encoding: http://www.ics.forth.gr/isl/CRMext/CRMgeo_v1.2.rdfs