Integrating Archaeological Datasets: the ARIADNE Portal

Paola Ronzino[1], Achille Felicetti[1], and Sara Di Giorgio[2]

[1] PIN, Polo Prato, Italy
{paola.ronzino, achille.felicetti}@pin.unifi.it
[2]Istituto Centrale per il Catalogo Unico delle biblioteche italiane, Roma, Italia
{sara.digiorgio@beniculturali.it}

Abstract. One of the emerging needs of the archaeological community is represented by the importance of availing of systems that allow to tackle new research questions, by querying diverse available resources. Usually, archaeological digital data is stored in non-standardised individual databases with a limited possibility of integration and a high level of fragmentation. The EU-funded project ARIADNE, has developed an e-infrastructure which enables the integration of archaeological datasets from various different institutions, integrating resource discovery metadata using controlled vocabularies, thesauri, gazetteers and ontology (CIDOC CRM). This paper presents the ARIADNE infrastructure, describing the activities undertaken by the project to achieve interoperability of archaeological resources at the dataset and item level. Moreover, the architecture of the ARIADNE Infrastructure and the Portal, with the different ways to search and access the resources are described.

1 Introduction

In the recent years we have assisted to an increasing awareness of the importance of creating networks of data that allow integrated access to documentation and to digital archives of archaeological resources. An important condition for the development of such networked accesses lays in the definition of standards and guidelines that establish a degree of compatibility between the datasets that make these networks up.
Usually, data is stored in non-standardised individual databases with a limited possibility of integration and a high level of fragmentation of data. This is mostly due to the different needs of the various research communities who store and structure their data according to the standards that apply to their specific research domain. However, when the different communities agree to share their data with the wider community and for a broader purpose, the related problem of data interoperability arises. This is the challenge that ARIADNE is facing [1, 2]. The EU-funded project has developed an e-infrastructure that enables the integration of archaeological datasets from various different institutions. ARIADNE's main objective is to provide researchers with an integrated access and to guarantee the semantic interoperability of archaeological datasets distributed throughout Europe. The main expectation of the project is that researchers will make use of these resources and benefit from them through the use of technologies and services made available by the infrastructure itself, and to challenge

new scientific questions. In the following sections we will introduce the consortium and the content made available into the ARIADNE infrastructure, touching on the user requirements that influenced the structure of the portal. Moreover, the paper describes the overall architecture of the infrastructure and the core services, which make ARIADNE a powerful system for the archaeological research community for sharing, discovering, accessing and reusing available data.

2 The ARIADNE Research Infrastructure

The ARIADNE infrastructure is supported by a consortium of archaeological institutes and data archives providing content, and by technology developers.
The consortium consists of 23 partners in 16 countries, and a number of associate partners that ensure an almost complete coverage of the European territory.
Most of the archaeological institutes involved in ARIADNE started to integrate archaeological datasets under a common portal, responding to the need of digital preservation, open access and networked access. Part of the network (to cite a few) are the UK's Archaeological Data Service (ADS) [3], which currently provides access to over 36,000 unpublished fieldwork reports and over 1000 data digital archives, the Data archiving and Networked Services (DANS) [4] providing access to over 21,000 reports and 4,000 excavation archives of Dutch archaeology, the Swedish National Data Service (SND) [5], based at the University of Gothenburg, the CulturaItalia Portal of the Italian Ministry of Cultural Heritage (MIBACT-ICCU) [6].
Content provided by partners was necessarily created and documented in different ways, using different languages and encoded by means of different metadata schemas. This obviously makes data integration a complex process.

2.1 Gathering user requirements

Before starting the ARIADNE infrastructure design, the project carried out various surveys and interviews to the archaeological research community, to find out the existing and emerging needs, so that the infrastructure would build on their basis. The research consisted of an extensive literature review, numerous interviews involving members of the ARIADNE partners and other stakeholders, two online questionnaire surveys with participation of more than 600 archaeological researchers and repository managers [7]. Subsequently, a survey of existing data portal involved 23 ARIADNE archaeological researchers and data managers, to get further insight for the development of the ARIADNE portal services [8]. The result of the two international online surveys made clear that archaeological researchers lack appropriate data repositories and services that allow to find and access relevant data. The 95% of the responses expressed that the most important need is to have a comprehensive overview of the available datasets. Most researchers asked for a data portal that offers an overview of the resources available online, with the possibility to search across resources scattered in different places, using new mechanism to discover and access data. For space reason we limit the description of analysis result to the top-level needs. The detailed

analysis of the user requirements can be found at [7]. The recommendation and the user needs collected through the activities presented above, influenced and guided the technology partners in the design and implementation of the infrastructure architecture, of the portal interface and of the services.

2.3 The ARIADNE Catalogue

The first step towards data interoperability consisted in a preliminary analysis of the available archives, in order to identify formats, standards, and services in use by the content providers [9]. Key elements, common to all archives, were identified and encoded using existing international standards and terminological tools, and referring to the "what, where, when" paradigm. The descriptions of content have been encoded using the ARIADNE Catalogue Data Model (ACDM) [9], developed by ARIADNE with the aim to produce a detailed representation of the archaeological information of the legacy archives made available by the consortium.
The ACDM was built on the DCAT vocabulary [10] and was extended with classes and properties needed to better describe the ARIADNE archaeological resources.
The main classes, which reflect the ARIADNE assets, are: *DataResources* (including the resources that are the containers of the data, like databases and collections), *LanguageResources* (such as vocabularies and metadata schemas) and *Service* (owned by the ARIADNE partners and offered to the project for integration).
The ARIADNE Catalogue aggregates metadata describing datasets, metadata schemas, vocabularies, which were harvested either manually or trough the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The metadata and object repository aggregator (MORe), was customized to meet the ARIADNE content needs and used to aggregate metadata from multiple sources and in multiple formats. The Catalogue, and the detailed information it contains, represents the core of the entire integration process. The Catalogue lays at the base of the portal, and provides the necessary support for retrieval, analysis and resource discovery facilities.

2.3 Datasets integration

The highest level of integration started at the conceptual level, by identifying key elements, common to each archive. The "what, where, when" paradigm was used to identify objects, places and time periods, which are fundamental elements in the archaeological domain.

2.3.1 The WHAT

The "what" represents the subjects of the various datasets. These are described using terms derived from the Art and Architecture Thesaurus (AAT) [11] of the Getty Research Institute, which was adopted as the spine for the whole framework of terms in ARIADNE. Each of the terminological resources used by ARIADNE content providers was mapped to the AAT concepts to demonstrate the semantic and conceptual similarity between the different archives. The mapping activities were facilitated by a

mapping tool developed to establish correspondences between concepts coming from different vocabularies [12].

2.3.2 The WHERE

The "where" represents the spatial entities. Most of the archaeological archives had already standardized spatial information. To enable browsing the archives, ARIADNE recommended content providers to provide geographic information in the WGS84 format. When the only information available was the name of a place, the spatial coordinates were retrieved through the GeoNames gazetteer.
Geographic information about historic names were retrieved from Pleiades [13], through a collaboration between ARIADNE and the Pelagios project.

2.3.3 The WHEN

The "when" represents temporal entities. When the dates were expressed in numeric format, the temporal integration was easy to manage. When, instead, the periods were indicated as names, for example Bronze Age, this caused a lot of ambiguities because these were not referred to absolute dates. Beside converting each period in absolute time spans, the collaboration with the PeriodO project [14], allowed to manage collections of periods as intersections of documented events on specific geographical areas providing unique identifiers for each of those periods as Linked Open Data.

2.4 Item-level Integration

The possibility to answer a research question by using relevant information from several available heterogeneous sources, is one of the emerging needs of the archaeological community. To address the complexity of archaeological data integration, ARIADNE developed a global, extensible schema as a formal ontology to allow for integration without loss of meaning. The CIDOC CRM ontology [15] was chosen as the backbone of the ARIADNE Reference Model, which includes a suite of extensions developed to address the complexity of archaeological data integration.
CIDOC CRM (ISO21127) is a formal ontology created to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It contains the basic relationships needed to describe what happened in the past, as for example people and things meeting in space-time, parts and wholes, use, influence and reference.

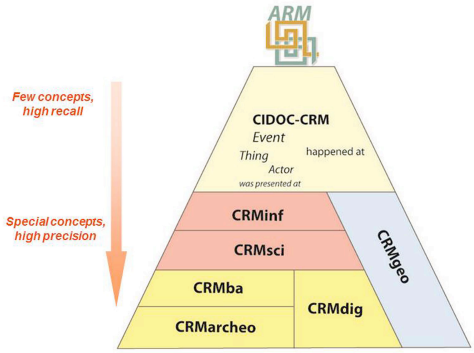

Fig.1 The ARIADNE Reference Model

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

- CRMinf [16]: 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 [17]: 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 [18]: is a spatiotemporal model that provides a link between the standards of the geospatial and the Cultural Heritage community in particular between GeoSPARQL and CIDOC CRM.
- CRMdig [19]: is an ontology to encode metadata about the steps and methods of production (provenance) of digitization products and digital representations such as 2D, 3D or animated models created by various technologies.
- CRMba [20]: 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 [21]: the Excavation Model is an ontology to encode metadata about the archaeological excavation process.

With the definition of the ARIADNE RM, and the creation of an integrated knowledge base, the aggregation of several existing archaeological databases were transformed by mapping their individual schemas to the ARIADNE RM. The mapping process was supported by the X3ML Mapping Framework, a tool that ensures the integrity of the initial data and preserving the original meaning [22].
An advanced level of the interoperability was achieved with the integration of individual records of legacy archaeological archives. This item-level integration experiment had the aim to reach the deepest integration of archaeological data. Preparatory activities towards this goal included mappings with specific tools which allowed individual partners to track complex correspondences between the entities contained in their databases and the conceptual classes provided by the CIDOC CRM and its extension. Conceptual mappings for each partner's archives enabled the creation of

semantic representations for individual items in RDF, to form a complex graph of relationships to be queried, integrated with semantic technologies and published in Linked Open Data format. ARIADNE chose as a use case the numismatics field, to demonstrate the item-level integration process of archaeological datasets. Five datasets were selected. Four of them were mapped to the ARIADNE RM and transformed to RDF using the X3ML framework, while the fifth was already in CIDOC CRM RDF form, and therefore, compatible with the ARIADNE RM. As a common thesaurus for the aggregated knowledge base, the nomisma.org, and the AAT thesauri were adopted. The mapping and transformation workflow is presented in Figure 2. The main goal of the integration of the diverse coin datasets was to create a system enabling users to specify queries that will be evaluated on the common aggregated repository and will be able to combine results coming from the different datasets. The ARIADNE portal provides a main access point to integrated repository and an intuitive user interface will guide the user to formulate the query, browse the results and refine the search with facet view.

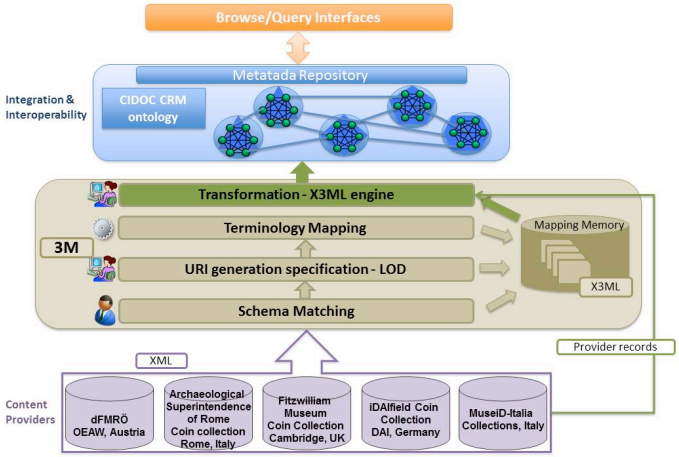

Fig.2 Mapping and transformation workflow

3 The Integrated Infrastructure and the ARIADNE Portal

The ARIADNE Portal represents the highest level of the infrastructure. It is the access point of the whole infrastructure, where users can browse, query and analyse available data and use the services to activate all the features provided by the system.
The integration platform designed and implemented by ARIADNE appears as a complex modular system, with advanced interfaces and features and an architecture able to interact with distributed archives, in a transparent way. The system is able to query and extract integrated information concerning legacy archives, to present them to users by means of advanced services and tools to visualize, analyse and possibly use them as part of subsequent queries. The search and browse operations are driven by the Catalogue, which, in addition to detailed descriptions, contains data related to

digital provenance. Catalogue information is used to address queries to the appropriate archives, which contain the information the user is interested in. A complex network of services will provide users with advanced features for using data in new ways, such as advanced visualisation, definition of use cases and scenarios potentially different from the ones in which the same data were created. Advanced interfaces for querying the item-level semantic network are also provided, so as to obtain relevant information about objects, places, events, people and types according to semantic criteria and to retrieved and display them in a user-friendly and meaningful way.

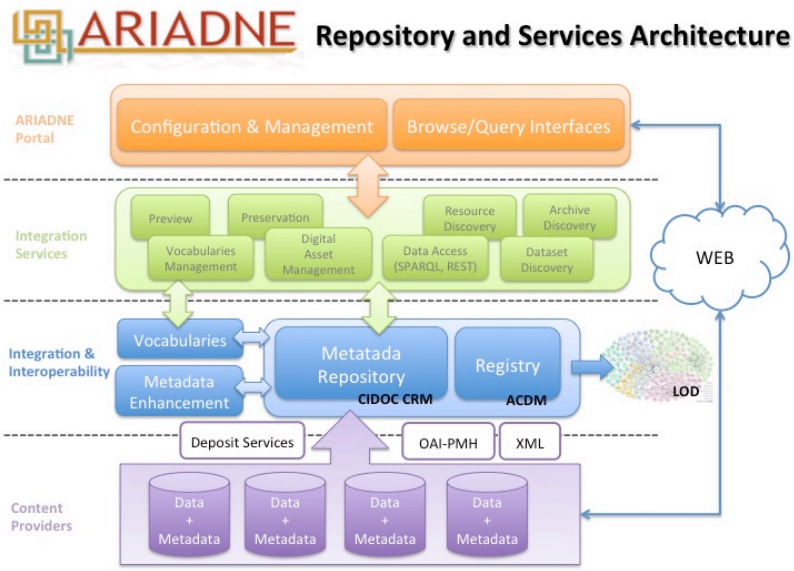

Figure 3: The ARIADNE Architecture

4 Conclusions

The ARIADNE portal was developed to create a unique global access point to provides researchers, repository managers, and the wider interested public, open access to integrated archaeological information. ARIADNE, in its scope, represents a substantial innovation for archaeology, as it provides a common platform where dispersed data resources can be homogenously described, discovered and accessed. It is also a crucial step towards the ambitious goal of offering archaeologists integrated data, tools and computing resources for web-based research and data reuse, which will pave the floor to the creation of new knowledge. The main ambition of ARIADNE is to achieve a wider engagement of the archaeological research community in sharing and reusing data through the ARIADNE portal.

Acknowledgments

The present work has been supported by the ARIADNE project, funded by the European Commission (grant 313193) under the FP7 INFRA-2012-1.1.3 call. The authors opinions do not necessarily reflect those of the European Commission.

References

1. Niccolucci, F., Richards, J.D.: ARIADNE Advanced Research Infranstructure for Archaeological Dataset Networking in Europe. International Journal of Humanities and Art, 7(1-2), pp.70–88, (2013).
2. ARIADNE 2016: www.ariadne-infrastructure.eu
3. ADS: www.archaeologydataservice.ac.uk
4. DANS: www.dans.knaw.nl/en
5. SND: www.snd.gu.se/en
6. CulturaItalia: www.culturaitalia.it
7. Selhofer, H., Geser G.: First Report on Users Needs. ARIADNE Deliverable D2.1 (2014). Available at http://www.ariadne-infrastructure.eu/Resources/D2.1-First-report-on-users-needs.
8. Selhofer, H., Geser G.: Second Report on Users Needs. ARIADNE Deliverable D2.2 (2015). Available at http://www.ariadneinfrastructure.eu/content/view/full/1188
9. Papatheodorou, C., et ali: Initial report on standards and on the project registry. ARIADNE Deliverable 3.1 (2013).
10. Goedertier, S.: DCAT application profile for data portals in Europe (2013). https://joinup.ec.europa.eu/asset/dcatn application profile/
11. AAT: http://www.getty.edu/research/tools/vocabularies/aat/
12. Binding, C., Tudhope, D.: Improving Interoperability using Vocabulary Linked Data. International Journal on Digital Libraries 17, 1 (2016), 5–21.
13. Pleiades: https://pleiades.stoa.org
14. PeriodO: http://perio.do
15. CIDOC CRM. Current Official Version of the CIDOC Conceptual Reference Model (2015). Available at http://www.cidoc-crm.org/docs/cidoc crm version 6.2.pdf
16. CRMinf: the Argumentation Model, version 0.7 (2015). Available at http://www.ics.forth.gr/isl/CRMext/CRMinf/docs/CRMinf-0.7.pdf
17. CRMsci: the Scientific Observation Model, version 1.2.3 (2016). Available at http://www.ics.forth.gr/isl/CRMext/CRMsci/docs/CRMsci1.2.3.pdf
18. Doerr, M., Hiebel, G.: CRMgeo: Linking the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement. TECHNICAL REPORT: ICS-FORTH/TR-435, (2013
19. Doerr, M., Theodoridou, M.: CRMdig: A Generic Digital Provenance Model forScientific Observation. In USENIX workshop on the Theory and Practice of Provenance (TaPP). Heraklion, Crete (2011).
20. Ronzino, P.: CIDOC CRMba A CRM Extension for buildings archaeology information modelling. (Unpublished doctoral thesis). The Cyprus Institute, Nicosia, Cyprus (2015)
21. CRMarchaeo: the Excavation Model, version 1.4 (2016) Available at http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo_v1.4.pdf
22. Minadakis, N., et ali: X3ML Framework: An Effective Suite for Supporting Data Mappings. Proceedings Workshop EMF-CRM2015, Poznań, Poland, September 17 (2015) CEUR-WS.org, online CEUR-WS.org/Vol-1656/paper1.pdf