Enabling European Archaeological Research: The ARIADNE E-Infrastructure

Nicola Aloia, Ceri Binding, Sebastian Cuy, Martin Doerr, Bruno Fanini, Achille Felicetti, Johan Fihn, Dimitris Gavrilis, Guntram Geser, Hella Hollander, Carlo Meghini, Franco Niccolucci, Federico Nurra, Christos Papatheodorou, Julian Richards, Paola Ronzino, Roberto Scopigno, Maria Theodoridou, Douglas Tudhope, Andreas Vlachidis, and Holly Wright

Cite this as: Aloia, N. et al. 2017 Enabling European Archaeological Research: The ARIADNE E-Infrastructure, Internet Archaeology 43. https://doi.org/10.11141/ia.43.11

1. Introduction

In the last 20 years, e-infrastructures have become ever more important for the conduct and progress of research in all branches of scientific enterprise. Increasingly collaborative, distributed and data-intensive research requires the sharing of resources (data, tools, computing facilities) via e-infrastructure as well as support for effective co-operation among research groups (ESF 2011; ESFRI 2016). Moreover there is the expectation that with large datasets ('big data'), e-infrastructure and advanced computing techniques, new scientific questions can be tackled.

The archaeological research community has been an early adopter of various digital methods and tools for data acquisition, organisation, analysis and presentation of research results of individual projects. The provision of e-infrastructure and services for data sharing, discovery, access and re-use for the heritage sector is, however, lagging behind other research fields, such as the natural and life sciences. The consequence is a high level of fragmentation of archaeological data and limited capability for collaborative research across institutional and national as well as disciplinary boundaries (Aspöck and Geser 2014).

This situation is being addressed by ARIADNE: the Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. This e-infrastructure initiative is being promoted by a consortium of archaeological institutes, data archives and technology developers, and funded under the European Commission's Seventh Framework Programme (ARIADNE 2014a; Niccolucci and Richards 2013). ARIADNE enables archaeological data providers, large and small, to register and connect their resources (datasets, collections) to the e-infrastructure, and a data portal provides search, access and other services across the integrated resources. The portal puts into operation a proof of concept exemplar first developed under the ARENA (Archaeological Records of Europe Networked Access) project (Kenny and Richards 2005; Kilbride 2004), itself inspired by a proposal made by Hansen (1993).

ARIADNE integrates resource discovery metadata using various controlled vocabularies, e.g. the W3C Data Catalogue Vocabulary (adapted for describing archaeological datasets), subject thesauri, gazetteers, chronologies, and the CIDOC Conceptual Reference Model (CRM). Based on this integration the data portal offers several ways to search and access resources made available by data providers located in different countries. ARIADNE thus acts as a broker between data providers and users and offers additional web services for products such as high-resolution images, Reflectance Transformation Imaging (RTI), 3D objects and landscapes. Employing such services in research projects or for content deposited in digital archives will greatly enhance the ability of researchers to publish, access and study archaeological content online.

ARIADNE therefore represents a substantial advance for archaeology; in particular it provides a common platform where dispersed data resources can be uniformly described, discovered and accessed. It is also an essential step towards the even more ambitious goal of offering archaeologists integrated data, tools and computing resources for web-based research that creates new knowledge (e-archaeology).

The next section describes the current landscape of data repositories and services for archaeologists in Europe, and the issues that make interoperability between them difficult to realise. The results of the ARIADNE user surveys undertaken to match expectations and requirements for the e-infrastructure and data portal services are then presented. The main part of the article describes ARIADNE's overall architecture, core services (data registration, discovery and access) and other extant or experimental services. A further section presents the on-going evaluation of the data integration and set of services. Finally, the article summarises some lessons already learned in the integration of data resources and services, and considers the prospects for the wider engagement of the archaeological research community in sharing data through the ARIADNE e-infrastructure and portal.

2. The Archaeological Research Communities and their Requirements

2.1 Existing infrastructures, standards, best practices, services and data

Most European countries have provision for the documentation of archaeological sites and monuments through national or regional databases. Despite being created for management purposes, these can also be invaluable research tools, although public access is rarely provided. Several initiatives have begun to integrate archaeological datasets under a common portal, often national in scope. Some have responded to the need for research-focused services such as digital preservation and open access. The best known is the UK's Archaeology Data Service (ADS), established in 1996. The ADS is the mandated place of deposit for archaeological research data for a number of UK research councils and heritage organisations and makes all of its holdings freely available for download or online research. The ADS currently provides access to over 36,000 unpublished fieldwork reports and over 1000 data-

rich digital archives. It was the first archaeological digital archive in Europe, but there have been related initiatives in several other European countries, although so far these are concentrated in Northern Europe and Scandinavia. In 2007 the ADS was joined by EDNA, the e-depot for Dutch archaeology, which was established as part of DANS (Data Archiving and Networked Services). In 2007 agreements to deposit archaeological data at DANS were formalised in the quality standard for Dutch archaeology, making archaeology one of the largest components of the digital resources hosted by DANS. By 2016 DANS provided access to over 21,000 reports and 4,000 excavation archives, with collections growing daily. The Swedish National Data Service (SND), based at the University of Gothenburg, has also extended its collection policy to focus on archaeology. It archives a number of archaeological reports, including over 450 GIS files with excavation data from Östergötland. A second Swedish infrastructure, the Strategic Environmental Archaeology Database (SEAD) is based at Umeå University, and focuses upon access to data pertaining to environmental archaeology. After a three-year preparatory phase, begun in 2012, the German Archaeological Institute (DAI) is now developing IANUS, a digital archive for German archaeology. ARIADNE has provided additional impetus for other countries to develop their own infrastructures, including Austria, Hungary, Ireland and Slovenia, and there are also new initiatives in Denmark and Norway. Outside Europe, the United States has the best developed archaeological digital repository, tDAR, hosted at Arizona State University on behalf of the Digital Antiquity consortium. Open Context, hosted by the Alexandria Institute, provides an alternative option, although its focus is digital data publication rather than preservation.

There are other infrastructures that focus upon networked access rather than digital preservation. Classical archaeologists are relatively well provided for in this regard. Fasti Online has, since 2000, provided a database of archaeological excavation projects for classical archaeology.

The project originated in Italy, but now includes a further nine countries. At the level of artefacts rather than excavations, Arachne is a major resource. Arachne is the central object database of the German Archaeological Institute (DAI) and the Archaeological Institute of the University of Cologne. It provides archaeologists and classicists with an online research tool for quickly searching hundreds of thousands of records on objects and their attributes. Both Fasti Online and Arachne also supply data to Pelagios, an initiative supported by the Mellon Foundation, to use Linked Open Data to aggregate information about the classical world. Finally, although primarily aimed at the general public rather than researchers, there have been a number of European projects, including CARARE, LoCloud and 3D-ICONS, which have aggregated archaeological data for the European cultural portal, Europeana. These tend to focus on image data but provide a useful resource for research.

Many of the existing research infrastructures already recognise that while modern Europe is highly politically and institutionally fragmented, archaeological research questions often transcend modern political boundaries. It is unrealistic that such data will ever be brought together in a single database and, in any case, it is better maintained at national or regional level where there is ownership and often a legal responsibility to maintain archives. Therefore we should look to options for interoperability that allow cross-searching of distributed resources. ARIADNE seeks to provide a bridge between existing national services, and to foster new ones where they do not so far exist. It differs from the existing national infrastructures in that it seeks to provide an integrating layer that exists independently of any one service, and it should allow the development of new research questions that transcend national datasets. However, in order to integrate services on the European level and beyond, there are a number of issues related to differences in classifications and vocabularies, metadata and different languages, which make interoperability difficult to realise.

1. Past cultures and modern political borders rarely correspond, hence researchers carrying out investigations that span sites located in different countries face a number of problems when trying to place their discoveries in a broader context. They would like to easily compare features and items of their site with those of sites in other countries, yet these will usually be documented in a different language and in a different way.
2. Thematic datasets, on the contrary, may span different regions. Yet in many cases they are unrelated to the context (e.g. a pottery database does not enable users to access other data concerning the fieldwork context in which the pottery was found).
3. Harmonisation of vocabularies among different, but similar, datasets is usually modest. When it exists, it is more often the result of good archaeological practice than a design feature of the databases involved. This affects terms, names, geographic names, and time periods.
4. For excavation and artefact datasets, almost all archaeological data are stored using one language and refer mainly, if not exclusively, to one country or a part of it.
5. Metadata structures are usually different, even in datasets with similar content.

Providing researchers with the ability to pose questions at a pan-European scale does not mean that there will always be single European answers. The importance of specific historical circumstances should not be underestimated: the limes of the Imperial Roman frontier system in Germania might be culturally and temporally equivalent to the Hadrian's Wall milecastles and the frontier fortlets of the Roman East, but the sheer scale of regional variation means that local factors will influence the particular form that these fortifications take. However, in order to appreciate the role of local circumstances one also needs to compare datasets that cross modern boundaries. During the Neolithic, many

European cultures developed megalithic tombs in order to commemorate their dead. Scholars who limit their research to the monuments of a single country – Britain or Ireland, say, or Denmark, France or Spain – will derive partial answers. Both archaeological and linguistic groups transcend political borders demarcated in the modern world. Nonetheless, the cultural context and different historical traditions within which archaeology has operated in the different European countries highlights the perils, as well as the benefits, of harmonisation. It will no doubt be easier to achieve interoperability in some areas than in others.

Some national systems have benefited from decades of investment in thesauri development and controlled vocabularies; others have grown organically and suffer from a lack of standardisation. Metadata are the key factor to guarantee interoperability among different data collections via mappings to common standards. They must be rich and specific enough to provide researchers with information useful and relevant for specific research questions. They must be simple to create and maintain, through automatic recording of machine-created or transformed data, or the use of standardised procedures and tools (thesauri and taxonomies, among others) when data are manually generated. The challenge here is to reconcile these apparently conflicting requirements and overcome the tension of simplicity vs richness and interoperability/generality vs specificity. This requires testing the effectiveness of metadata in research practice and expert evaluation of adequacy: a joint effort of archaeologists and information scientists. At a high level, substantial advances have been achieved through increased compliance of archaeological metadata schemas with the CIDOC-CRM. Within ARIADNE much effort has been invested in mapping between different national or regional-based time periods and subject classifications. At the level of individual file types and those metadata required to enable digital preservation and data re-use, then the online series of *Guides to Good Practice* initiated by the ADS has seen widespread adoption. The Guides

have been further developed in collaboration with the US-based Digital Antiquity consortium, and have been taken up by IANUS, with enhancements by ARIADNE partners.

2.2 Identification of user requirements

ARIADNE carried out several research activities to identify users' requirements for the e-infrastructure and portal services of the project. The objective was to ensure that ARIADNE addresses the existing and emerging needs of the archaeological research community in Europe and beyond. The research comprised an extensive literature review, 26 interviews with members of the ARIADNE partners and other stakeholders, two online questionnaire surveys with participation of over 600 archaeological researchers and repository managers, a survey of 25 content/data portals, and contributions by ARIADNE Special Interest Groups. Here we present selected study results, focusing on the surveys that allowed the project to acquire a good understanding of what users need and expect from the ARIADNE data infrastructure and services, which are being developed accordingly.

2.2.1 Online questionnaire surveys

Two international online surveys conducted in November/December 2013 collected needs and requirements of researchers and repository managers (ARIADNE 2014b, 69–143). The results made it clear that archaeological researchers in most countries lack appropriate data repositories and services for finding and accessing relevant data. The selected results presented below are based on between 470 and 590 survey responses per result.

The majority of researchers agreed that they often do not know what is available, because research data are scattered across many places and different databases. Consequently, 95% considered it to be very or rather important to have a good online overview of available datasets. About the

same percentage required datasets to be available online and in an uncomplicated way, not 'limited to specific persons/communities' or 'kept in private collections of other researchers'. Some 75% of respondents thought it important to have easy access to international datasets, suggesting a high interest in data that allows for comparative studies and integrative research.

Furthermore, 60% of the researchers said that their organisation (university, research institute or other) does not have an institutional repository that is managed by dedicated staff, and 66% perceived a lack of international archives. Indeed, most institutional repositories manage only documents. Consequently the survey found that data were made available through an institutional repository in a few projects only or not at all by 67% of the researchers. The figures for national and international repositories were 76% and 83%, respectively. Most researchers wanted ARIADNE to create a data portal that allows an overview of existing archaeological data resources and to provide search facilities across the resources, using novel mechanisms for data discovery and access. Asked which services they would benefit from most ('very helpful'), researchers responded: a portal that makes it more convenient to search for existing archaeological data that is stored in different archives/repositories (79%); innovative and more powerful mechanisms for data discovery and access (63%); a directory of European archaeological databases and repositories (52%); services for geo-integrated data (58%); and data recommendations based on collaborative filtering, rating and similar mechanisms (29%). Thus the capability to search and 'mine' distributed digital archives for relevant data was appreciated most. There was much less interest in typical features of Web 2.0 platforms, such as content filtered based on tags or ratings provided by other users. Researchers appreciate effective mechanisms that save time in identifying relevant data (e.g. clear licensing information); what they typically do not like are resources pre-selected by others.

The results of the online survey of managers of data repositories are only indicative owing to the small sample of 52 respondents. The main concern of the data managers is the quality of metadata, but they would also appreciate higher awareness of good practice in data management (e.g. available guides and recommendations) among researchers. Moreover the data managers more than the researchers expected much better data access through improvements in data/metadata extraction and indexing as well as Linking Data. Nonetheless, Web 2.0 features were also ranked last among this group.

2.2.2 Survey of existing data portals

Further insights for the development of the ARIADNE portal services have been acquired through a survey of various websites by a panel of 23 archaeological researchers and data managers involved in the project (ARIADNE 2015a). The panel members served as 'lead users' because they make intensive use of searchable archives and other websites and have a good understanding of the state-of-the-art and potential solutions that might serve their requirements even better.

The survey evaluated 25 archaeological websites, giving access to content/data of more than one institution or project, and some existing data portals of other domains. Most of the websites/portals were 'international' in that they provide access to content/data from research in more than one country. The survey participants looked for good practices and gave recommendations for services of the ARIADNE portal. The 34 suggestions of the survey report were then evaluated by 28 experts from 21 project partners in order to focus on the most relevant services in the short to medium term (ARIADNE 2015b, 278–89).

The highest scores were received by highly functional portals, e.g. with regard to overview of searchable data and portal navigation, and search and filter functionality based on geo-location (maps) and date ranges/chronologies. High relevance was also attached to deploying

Linked Open Data to integrate information within the portal and to link to external resources. Furthermore providing interfaces to allow external applications to exploit available data, metadata and terminologies was considered as important. Indeed, the ARIADNE infrastructure and portal should not be an 'island' but enable added value in the wider information ecosystem of archaeology and beyond.

Some suggested portal features were not ranked highly. These features concern personalised portal services (e.g. alerts on possibly relevant new data), linking of online professional information (e.g. researcher profiles) or networking and discussion on the portal. Portals for the latter exist (e.g. Academia.edu, ResearchGate and others) and are used by many archaeological researchers. Clearly the service portfolio of the ARIADNE portal should meet core requirements of data discovery, access, visualisation and re-use. There is little scope to invest limited funds in specific services that are not appreciated, are provided by other portals, or may run ahead of the needs of broad user segments.

The latter includes support for online research work (e-research), which is not an immediate need of the archaeological research community, but may emerge when more open data becomes available through digital archives and novel services provided by e-infrastructures. However, some specific ARIADNE services (e.g. for visual media) can be seen as a first set of services of a future virtual research environment for archaeologists.

2.2.3 Requirements for Visual Services

To complement the user requirements study described above, the project organised a workshop specifically aimed at gathering a clear view of user needs related to Visual Data technologies and services. The results made it clear that the community was already intensively producing visual data (2D, 3D, videos, terrains) and that the status of the related enabling technologies was considered sufficiently consolidated. Conversely, we discovered that one of the major limitations perceived was the lack of

knowledge and tools for easy sharing of these visual resources and to support remote visual analysis (web-based publication and visualisation). In response to these needs, two services have been designed and implemented as part of the ARIADNE Infrastructure: Visual Media Resources and Landscape Services, both described in section 3.4.

3. The Ariadne Research Infrastructure

This section describes the ARIADNE infrastructure starting with its architecture, and proceeding to its main services.

3.1 Rationale and overall architecture

Integration of data created by archaeological research and in the Cultural Heritage domain in general is a highly complex process. This complexity mainly results from the fact that, although they are often very similar to each other, the diverse institutions that create and use such information have to maintain varying types of collections that are documented in different ways, with no common language and different metadata schemas for their encoding. Very often, the way information is organised is influenced by the vision derived from related disciplines or by specific objectives related to the places and periods under study. However, managing this information in an interoperable way has become a vital necessity to ensure efficient use in order to unlock its full potential and to bring a significant contribution to the advancement of archaeological research. This can only take place in an integrated environment where different data are mutually interpretable and able to be consumed as if they were stored in a single archive. The retrieval of meaningful information on both a factual and space/temporal level will thus be ensured.

Integration in ARIADNE required a preliminary analysis of the archives, necessary for the identification of formats, standards and services already in use by the content providers in charge of supplying content to the

project. Descriptions of these analyses were collected in various ways and encoded using a data model, the ARIADNE Catalogue Data Model (ACDM), developed by ARIADNE specifically to produce a detailed, formal and unambiguous representation of the archaeological information of the legacy archives (and described in detail in section 3.2.2). Integration usually means a series of complex operations that takes place on multiple levels and at multiple depths. The core of any activity of this type is the identification of key elements, common traits that can identify objects and conceptual entities that could then be described through a common language.

The top level of this integration starts at the conceptual level, where these fundamental elements can be detected in each archive and captured in accordance with the famous 'who, what, where, when' paradigm, in order to identify people, objects, places and time periods, elements of crucial importance especially in archaeology. Careful analysis of these elements demonstrated that integration based on these profiles was possible, if preceded by an appropriate reduction of the concepts themselves to a common shared language. ARIADNE has therefore devoted part of its activities to the identification of those key features and the proper encoding using existing and already well-accepted international standards and terminological tools.

Definition and encoding of key elements and high-level entities has constituted the basis for the creation of the ARIADNE Catalogue, a core resource intended to store metadata and other valuable information concerning the archaeological archives and services connected to them. The catalogue, and the detailed descriptions it contains, constitutes the core of the whole integration process, since it provides all the support necessary for the retrieval and analysis of integrated archaeological information and the resource discovery facilities.

The subjects to which the various datasets relate (e.g. excavations and archaeological surveys, monuments, burials, pottery and the like), which constitute the 'what' strand in our model, are described using terms drawn from the Art and Architecture Thesaurus (AAT) of the Getty Research Institute. The AAT forms the spine for the whole framework of terms in ARIADNE, not only with regard to the general subjects, but also for every other typological, morphological and functional description of archaeological objects and activities connected to them. The use of a shared thesaurus required a mapping of each terminological resource already in use by content providers to the AAT concepts.

Integrating spatial entities (the 'where') was also straightforward since many archaeological archives already contain detailed spatial data in a standard format. ARIADNE has recommended the use or the conversion of the spatial coordinates in WGS84 format to enable the browsing of archives through geographical tools. Specific resources, like the GeoNames gazetteer, were used to obtain spatial coordinates, starting from simple names of places in the case where these were the only geographic information present. As for the use of the historical names that a location may once have had, an invaluable collaboration with the Pelagios project was established in order to get geographic information from Pleiades (a thesaurus of past places built on a bibliographic database) and deploy it in Linked Open Data format to unambiguously identify such places.

Of particular interest was the time-based integration (the 'when'), including information concerning dates, times, time intervals and periods abundantly present in archaeological archives. The sharing of dates expressed in numeric format poses no problem, these being unambiguous. It should, however, be noted that very often time indications in databases only appear as simple names, without any reference to absolute dates; this may give rise to ambiguities in an integrated perspective, e.g. the Iron Age in Anatolia has a very different

time span from the Iron Age in the British Isles. It is evident, therefore, that the temporal definition of an 'age' in the absolute sense is impossible without a precise spatial reference.

An obvious and immediate solution to the problem of periodisation was to convert each period to absolute time spans by specifying start and end dates. However, this would not solve the semantic overlaps resulting from the need to keep the original time stamps as part of the documentation. Collaboration with the PeriodO project, whose aim is to manage collections of periods built as intersections of documented events on specific geographical areas, helped to solve this issue.

A deeper stage of interoperability has been reached with the integration of individual records coming from the legacy archaeological archives; this is what ARIADNE has defined as 'item-level integration'. Preparatory activities towards this goal include a broad conceptualisation, mappings and conversions of archaeological information and the construction of a repository with semantic capabilities to perform complex queries on aggregated data. The implementation of these features is based on the definition of mappings able to capture and express the semantic richness of archaeological data. Mappings are performed within the project through specific tools that allow individual partners to track complex correspondences between the entities contained in their databases and conceptual classes provided by the CIDOC-CRM and its extensions (CRMarchaeo in primis, see section 3.5). Conceptual mappings for each partner, applied to real data, enable the creation of semantic representations for individual items in RDF, in order to form a complex graph of relationships ready to be viewed, queried, integrated with semantic technologies and published in Linked Open Data format.

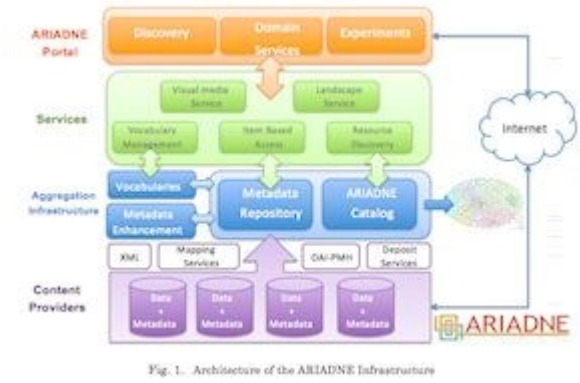

Fig. 1. Architecture of the ARIADNE Infrastructure

Figure 1: Architecture of the ARIADNE infrastructure

The integration platform designed and implemented by ARIADNE (shown in Figure 1) appears, in its final form, as a complex modular system, providing 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, and to present them to users in a coherent way by means of advanced services and tools to visualise, analyse and possibly use them as part of subsequent queries.

All the operations are constantly driven by the catalogue, which, in addition to detailed descriptions of the original files, contains data related to digital provenance and the complete record of all the 'addresses' through which legacy data can be browsed and harvested. Catalogue information is used to address queries to those archives that contain the information the user is interested in. A set of additional services, deployed on top of the integrated framework, will provide users with advanced features for using data in different ways, such as advanced visualisation and landscape analysis for the definition of use cases and scenarios potentially different from the ones in which the same data were created.

The access point to the whole infrastructure is the ARIADNE Portal, which represents the highest level of the architecture. Through it, users are able to browse, query, analyse all the available information, discover and activate the services, and trigger all the features provided by the system.

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 retrieve and display them in a user-friendly and meaningful way.

3.2 Resource discovery

Resource discovery is the basic service of the ARIADNE infrastructure, allowing researchers to (a) discover the data and services that populate the ARIADNE information space, (b) obtain basic information about them, and (c) access them. This service hinges on the ARIADNE Catalogue, a collection of descriptions of the resources, structured according to the ACDM. The descriptions in the catalogue are computed by the ARIADNE aggregation infrastructure, which takes the original descriptions from the holding institutions and transforms them into valid ACDM records.
The *Box* provides an introduction to metadata registries. The rest of this section gives the basics of the ACDM, then presents the ARIADNE aggregation infrastructure, and concludes with the search functionality enabling discovery, divided into querying and browsing. The current contents of the catalogue are described in section 4.1.

3.2.1 The ARIADNE Catalogue data model

The main goal of the ARIADNE project is 'to bring together and integrate the existing archaeological research data infrastructures so that researchers can use the various distributed datasets and new and powerful technologies as an integral component of the archaeological research methodology'. In order to achieve this goal, it is necessary (i) to gather information about the existing data resources and services in the archaeological domain, and (ii) to implement advanced search functionalities across this information in order to support the discovery of resources that make good candidates for integration. As a necessary step towards the realisation of the first objective, a data model is needed for representing archaeological resources that come in three different types:

Data Resources, including the resources that are containers of data such as databases and collections; Language Resources, including the resources related to the formal languages used in Data Resources, such as vocabularies and metadata schemas; and Services, including the resources offering some kind of functionality in the archaeological domain.

The ACDM was built around the DCAT vocabulary, which was expanded by adding classes and properties that were needed for best describing the ARIADNE assets. Its adoption therefore places ARIADNE in an ideal position to publish archaeological data resources as Open Data, and demonstrates the application of DCAT to research datasets. As illustrated in Figure 2, the central notion of the model is the class *ArchaeologicalResource*, which uses terms of the DCAT vocabulary, to which it adds properties for specifying the access policy and the original identifier of the resource. The class, as noted above, is specialised in:

1. *DataResource*, whose instances represent the various types of data containers owned by the ARIADNE partners and lent to the project for integration. This class is created for the sole purpose of defining the domain and the range of a number of associations. It is therefore an abstract class, whose instances are inherited from sub-classes.
2. *LanguageResource*, having as instances vocabularies, metadata schemas, gazetteers and mappings (between language resources). As new resources of a linguistic nature are added to the catalogue (such as subject heading systems and thesauri), the corresponding classes will be added to the model as a sub-class of this class. To describe language resources we have again used ISO/IEC 11179 (ISO 2004).
3. *Services*, whose instances represent the services owned by the ARIADNE partners and lent to the project for integration. (Each of these classes is described in some detail in this *Box*).

3.2.2 The aggregation infrastructure

The ARIADNE Catalogue aggregates metadata, such as descriptions for datasets, metadata schemas, vocabularies, etc. provided by the project partners utilising the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Content aggregation is inherently a content-driven task. This raises the importance of the data model, which needs to be robust and flexible in order to aggregate information for different domains and schemas. Therefore the metadata and object repository aggregator (MORe) has been utilised and customised (Isaac *et al*. 2013) in ARIADNE. The MORe aggregator has been used effectively in numerous projects and provides an easy and flexible way of aggregating metadata from multiple sources and in multiple formats.

MORe aggregates dataset items that consist of seven data streams:

1. The administrative metadata stream, which contains information about the provider, package, and general history of the item.
2. The technical metadata, which contains technical metadata regarding the contents of the item.
3. The native metadata, which contains the source representation (e.g. the native metadata as they were initially harvested).
4. The enriched native metadata, which contains a representation of the enriched version of the native metadata.
5. The target metadata which contains the representation to the target schema.
6. The enriched target metadata which contains a representation of the enriched version of the target metadata.
7. Preservation metadata, which is a log of PREMIS events.

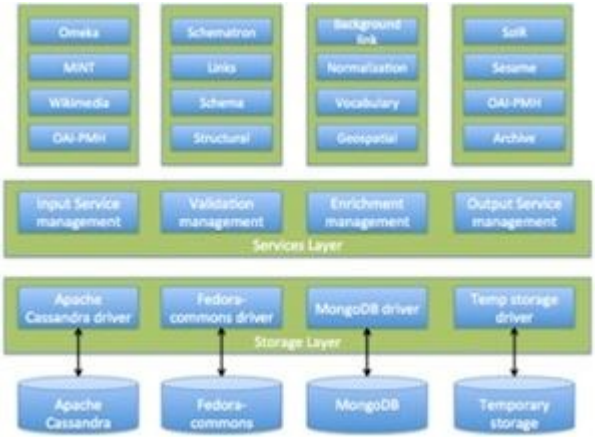

Figure 4: The architecture of the MORe aggregator

The overall architecture of MORe (see Figure 4) includes the following major elements:

- *A storage layer*. The storage layer provides an API that allows attaching virtually any CRUD based storage technology. For each storage technology a driver implementation is required and currently the Apache Cassandra, Fedora-commons and Temporary storage have been implemented.
- *A services layer*. The services layer consists of a number of core services, including:
  - Harvest: responsible for harvesting content from multiple sources;
  - Ingest: responsible for ingesting content into the appropriate storage;
  - Validation: responsible for validating content;
  - Indexing: responsible for indexing specific elements;
  - Quality: responsible for measuring metadata quality;
  - Transform: responsible for transforming content from one schema to another;
  - Enrichment: responsible for enriching content using specific enrichment micro-services;
  - Publish: responsible for publishing aggregated content to a specific target.

- *A set of micro-services*. Some of the above services follow the micro-services architecture, where a set of micro-services is used to increase the flexibility of certain tasks. One of the most important aspects of MORe is that it employs a number of curation/enrichment micro-services that can enrich metadata in various ways. Indicative micro-services that have been integrated/developed in MORe are:
    - Geocoding: a geocoding as well as a reverse geocoding micro-service based on GeoNames.
    - Rule-based thematic enrichment: a subject collections micro-service that allows the user to create thematic collections of concepts encoded in SKOS.
    - Automatic thematic enrichment: a vocabulary-matching micro-service that identifies SKOS concepts based on title, descriptions and subject-related information found in each metadata record.
    - Wikipedia and DBPedia automatic enrichment: a background links service that automatically identifies Wikipedia and DBPedia entries, based on title, descriptions and subject related information found in each metadata record.
    - Language identification: identifies languages based on a title or description using Apache Tika.
    - Thesauri mappings: allows loading and managing SKOS concepts mappings from SKOSified subject terms to a target SKOS thesaurus.

3.2.3 Querying the ARIADNE catalogue

Users can discover ARIADNE resources via the ARIADNE portal (see Figure 5), which also provides access to the services made available by the consortium members. The Portal was built using version 5 of Laravel, an open source, PHP-based, web application framework. Laravel follows the Model-View-Controller architectural pattern, separating the concerns of the data model, front-end views and business controllers. Composer is

used as a dependency manager to add third-party PHP packages to extend the framework.

At the heart of the portal lies the ARIADNE catalogue, comprising descriptions of all the resources in the ARIADNE information space, according to the ACDM data model described in section 3.2.2. The ARIADNE catalogue includes descriptions of millions of resources, as detailed in section 4.1.

Figure 5: The initial page of the Ariadne portal

The general discovery functionality is a free text search accessible from both the portal entry page as well as from a bar located in the menu of the portal. The free text search enables access to all metadata fields of the ACDM. The entry page search also gives the option of specifying a number of facets to narrow down the search. Using these facets a user can filter a search so that specific items only are displayed. The available facets are:

- *Resource type*. Every resource in the portal is categorised with a resource type, which can be any of the following options: Fieldwork archives, Event/intervention resources, Sites and monument databases or inventories, Scientific datasets, Artefact databases or image collections, or Burial databases;
- *Native Subject*. Subjects from a vocabulary used by the original owner of the resource. Associated with the *skos:Concept* class;

- *Derived Subject*. Subjects derived from mapping native subjects to Getty AAT vocabulary terms;
- *Keyword*. Keywords or tags describing the resource;
- *Contributor*. The agent responsible for describing the resource in the catalogue;
- *Publisher*. The agent responsible for making the resource accessible;
- *Place*. Place names the resource is connected with;
- *Period*. Time periods the resource is associated with;
- *Rights*. Access rights connected to the resource;
- *Language*. Language of the resource.

These facets are also available on the search result page to display more specific items only.

The underlying storage and search engine is Elasticsearch, a Lucene-based open source search engine (https://lucene.apache.org) ideal for a product like the ARIADNE Portal, providing near real-time search on resources within provided indices. Elasticsearch has the capability to be run as a distributed system by dividing the included indices into 'shards', which in turn can have one or more replicas. This approach facilitates an automatic load balancing that has been built into the system. The content is stored as denormalised documents in a javascript object notation (JSON) structure, ingested into the data store by the MoRe Aggregator. The JSON structure has been derived from the ACDM model and structured to serve the search and discovery interface optimally (Figure 6).

Figure 6: JSON structure of a resource in Elasticsearch

Elasticsearch also provides a JSONstyle query language used to execute queries on the stored documents. This query language provides facilities such as, among others, full text queries, term-level queries as ranges, exist, wildcards, fuzzy search etc., and geo-queries.

The searchable content is stored as de-normalised documents in a Javascript Object Notation (JSON), ingested into the data store via the aggregation infrastructure described above. The JSON structure has been derived from the ACDM model and structured as to serve the search and discovery interface optimally.

Two separate indices have been created in Elasticsearch to accommodate the portal:

- First and foremost is the resource index where metadata for all resources have been included.
- The second index is the AAT index which includes AAT subjects as well as mappings of these terms to native subjects from data provider thesauri.

3.2.4 Browsing the ARIADNE catalogue

In addition to searching for specific topics through a full-text search interface, users can also visualise and filter the contents of the catalogue

along geospatial, temporal and thematic lines, thereby allowing them to explore and dig into the available information resources.

3.2.4.1 *Where – map-based browsing*

The 'where' section of the browsing interface is realised as a full-screen map layout based upon OpenStreetMap and implemented with the help of Leaflet. The main challenge that had to be resolved in the implementation process was to develop a view that would provide both a dynamic visualisation of vast amounts of geographical data and the ability to narrow down the visible resources in order to be able to pick out the specific datasets of interest to the user.

Therefore the resources are first visualised as a heatmap that represents resource density. This view dynamically changes to markers representing single locations when the user has reduced the result set by filtering or zooming.

The implementation of the dynamic heatmap is realised with the help of Elasticsearch's aggregation feature. This allows the creation of 'buckets' that cluster similar resources based on indexed field values. The particular index used in this aggregation is based on the geohash representation of geographical coordinates and accelerates access to geographically similar objects. By being able to do this accumulation on the server, based on indexes already present for the search functions, we were able to greatly reduce the cost of data transmission and processing in the browser. This enabled us to visualise millions of datasets without major lag or performance issues for the user.

3.2.4.2 *When – timeline browsing*

A similar approach was taken in the realisation of the 'when' section of the browsing interface. The particular implementation involves the creation of dynamically defined buckets that cover date ranges distributed

over a logarithmic scale. These buckets are then visualised as an area graph that represents the distribution of the dates connected to the archaeological resources over time. The visualisation, which is based on D3.js, also makes use of the zoom metaphor users are acquainted with from map interfaces, and allows drilling down into smaller date ranges for increased details. The user can then select date ranges as a starting point for a search in the catalogue.

3.2.4.3 *What – subject browsing*

The 'what' section aims to present yet another starting point for discovering the contents of the ARIADNE catalogue. Its purpose is to provide a summary of the different thematic aspects of the registered resources and to offer an exploratory entrance into the available subjects, built upon the unifying mapping provided by the AAT. Additionally the thesaurus data collected in the subject index is used to provide autocomplete suggestions for the search field. These can then be used to discover resources connected to a particular theme present in the common thesaurus.

3.3 Vocabulary resources and services

For subject access, the ACDM *ArchaeologicalResource* class has two kinds of subject property. The property, native-subject, associates the resource with one or more items from a controlled vocabulary used by the data provider to index the data. However, there are a large number of partner vocabularies in several different languages. Cross search and semantic interoperability is rendered difficult, as there are no semantic links or mappings between the various local vocabularies. Standard ontologies for metadata schemas, such as the CIDOC-CRM, do not have vocabulary coverage so there is a need to complement the ontology with the terminology contained in subject vocabularies. Trivial variations in spelling or different synonyms for the same concept can result in failure to find

relevant results. This problem is exacerbated when subject metadata may be in different languages, which is clearly the case when providing an infrastructure for European archaeology. Not only may useful resources be missed when searching in another language from the subject metadata but there is also the problem of false results arising from homographs where the same term has different meanings in different languages. For example, 'vessel' has different archaeological meanings in the English language, while 'coin' is French for corner, 'boot' is German for boat and 'monster' is Dutch for sample.

The established solution to this problem is to employ mapping between the concepts in the different vocabularies. However, the creation of links directly between the items from different vocabularies can quickly become unmanageable as the number of vocabularies increases. A scalable solution to this mapping problem is to employ the hub architecture, an intermediate structure where concepts from the ARIADNE data provider source vocabularies can be mapped (ISO 2013). In the portal, retrieval based on a concept from one vocabulary (in a search or browsing operation) can use the hub to connect to subject metadata from other vocabularies, possibly expressed in other languages. In the ACDM, *ariadne-subject* is used for shared concepts from the hub vocabulary (the AAT), which have been derived via the various mappings from source vocabularies. This underpins the MORe enrichment services augmenting the data imported to the registry with mapped hub concepts (see section 3.2.2). These derived subjects in turn make possible concept-based search and browsing in the ARIADNE Portal (see section 3.2.3). It is hoped that the mappings will also form one of the stepping stones towards a multilingual capability in the Portal.

The AAT was chosen as an appropriate hub vocabulary, following a prototype mapping and retrieval exercise involving five ARIADNE vocabularies (in three different languages). The AAT had recently been made available as Linked Open Data by the Getty Institute, which fit well

with ARIADNE's strategy for semantic interoperability. The AAT linked data is expressed in the standard SKOS RDF representation and the appropriate representation for the mappings is via SKOS mapping relationships (http://www.w3.org/TR/skos-reference/#L4138). The next step was to produce the mappings from the subject vocabularies employed to index the various datasets selected for the ARIADNE catalogue. This is not a trivial exercise. It requires domain experts to make quality mappings, who may not have expertise in computing semantic technologies. The vocabularies themselves vary from a small number of keywords from a picklist for a particular dataset to standard national vocabularies with a large number of concepts.

Table 1: Example of vocabulary mapping

sourceLabel	SourceURI	matchURI	targetLabel	TargetURI
DITCHED ENCLOSURE	http://purl.org/heritagedata/schemes/eh_tmt2/concepts/70361	skos:broadMatch	agricultural settlements	http://vocab.getty.edu/aat/300008420
CROFT	http://purl.org/heritagedata/schemes/eh_tmt2/concepts/68617	skos:closeMatch	small holdings	http://vocab.getty.edu/aat/300000211
etc.				

Two different tools were developed to support the domain experts doing the mapping between vocabulary concepts, orientated to different contexts for the vocabularies. An interactive mapping tool was developed for ARIADNE orientated to major vocabularies already expressed as Linked Data via local or national initiatives. The mapping tool generates SKOS mapping relationships in JSON and other formats between the source vocabulary concepts and the corresponding AAT concepts. To assist the production of quality mappings, the mapping tool displays the source concepts and the AAT concepts side by side, together with contextual evidence, and allows the person making the mappings to browse related concepts in either vocabulary to fine-tune the mapping.

The mapping tool is a browser-based application working directly with linked data, querying external SPARQL endpoints directly (Binding and Tudhope 2016). The mapping tool is open source and will be made available via the ARIADNE portal. The first complete mapping exercise was performed by ADS on UK HeritageData vocabularies. Analysis of results from a pilot mapping informed an iteration of the mapping guidelines and the mapping tool user interface. A complete set of mappings was then produced for the subject metadata used in the ADS data imported by the ARIADNE catalogue. These were reviewed by a senior archaeologist and the final mappings were communicated to the DCU Registry team as RDF/JSON statements. The mapping guideline revisions included recommendations on the appropriate SKOS mapping relationship to employ in different contexts and, when appropriate, to specify more than one mapping for a given concept. The revised guidelines were employed in the mappings of vocabularies from the other partners.

The second mapping tool was orientated to cases where the source vocabularies were not expressed as linked data and included simpler 'flat list' vocabularies. Since many of the simpler vocabularies were already available or easily expressed in spreadsheet format, the most flexible solution was to design a standard spreadsheet with example mappings that domain experts from the partners could use to specify the mappings. A CSV transformation produced the RDF/JSON format required by the catalogue. The spreadsheet was accompanied by a set of guidelines informed by the pilot mapping exercise (together with support from the vocabulary team on problematic mappings or precedents from other partner mappings). In some cases, data cleansing was required before the mapping exercise could proceed. The template used contained a tab to record metadata for the mapping. The mappings have potential to underpin various options in the search functionality and user interface, offering a cost-effective route towards different multilingual functionality.

In future work, making the mappings (and mapping services) fully available as outcomes in their own right, with appropriate metadata for the mappings would be desirable, as more than one mapping may be produced for large vocabularies.

The information from the mapping tool is passed to MORe, which associates it with the provider of the vocabulary. It updates the property *derived-subject* and enriches an ACDM record (see Figure 7), adding a broader term, or a *skos:altLabel* to correlate a term via the 'use for' relationship, or adds multilingual labels (*skos:prefLabel and skos:altLabel*) in order to facilitate multilingual search.

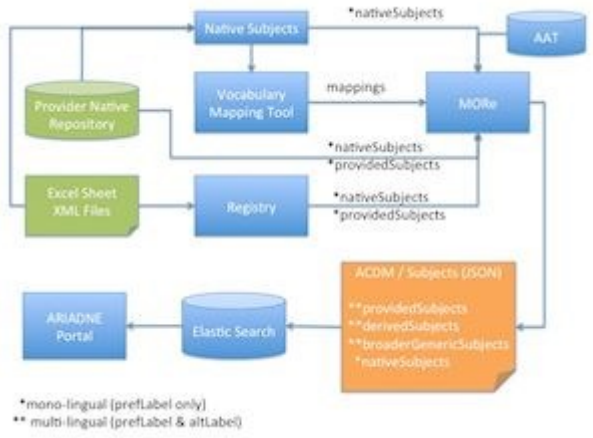

Figure 7: MORe enrichment

Prototype experiments have shown the potential of working with the URI identifiers of AAT concepts rather than the ambiguous strings of term labels. Using the URI identifier for the concept avoids the problem (discussed above) common with multilingual data of terms that are homographs in different languages. Working at the concept level also makes possible hierarchical semantic expansion, making use of the broader generic ('IS-A') relationships between concepts in a hierarchically structured knowledge organisation system, such as the AAT. Thus a search expressed at a general level can (if desired) return results indexed at a more specific level, for example a search on *settlements* might also return *monastic centres*. In some cases, ARIADNE has contributed to updated or even new subject vocabularies. One example is the ongoing

initiative to develop a multilingual SKOS vocabulary to be used for documenting data resulting from dendrochronological analysis. In other collaborations, ARIADNE has assisted with the generation of SKOS representations for national vocabularies.

3.4 Visual services

As pointed out at the end of section 2, ARIADNE included two services in its infrastructure: the Visual Media Service and the Landscape Service.

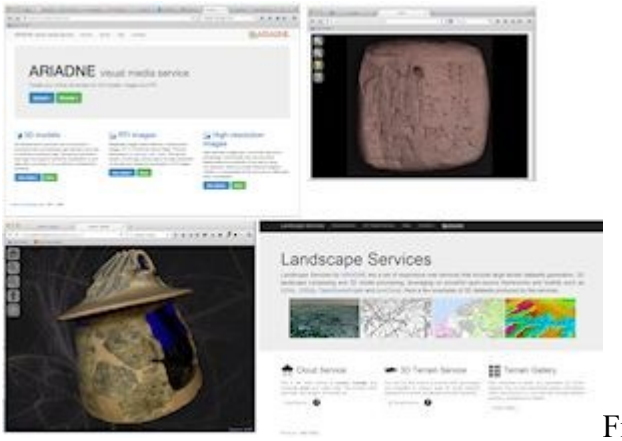

Figure 8: Some snapshots from the services (from top to bottom, left to right): the home page of the Visual Media Service; an example of an RTI image visualised with the RTI browser; an example of 3D model visualised with the provided 3D browser; the home page of the Landscape Services.

3.4.1 The ARIADNE Visual Media Service

The ARIADNE Visual Media Service (Ponchio *et al*. 2015) is a resource providing easy publication and presentation of complex visual media assets via a web browser. It is an automatic service that allows the user to upload visual media files to an ARIADNE server and to transform them into an efficient web format, making them ready for web-based visualisation. The user is asked simply to complete a short form and upload the raw file; all processing required to transform the data in a web-compliant and efficient format is done automatically by an ARIADNE server.

This service, released in January 2015, was extended in January 2016 and supports the publication on the web and browsing of the following three types of visual media:

- High resolution 2D images (input images are converted in a multi-resolution format and can be browsed in real time, zooming in and out);
- Reflection Transformation Images (RTI), also known as Polynomial Texture Maps (PTM) images, i.e. dynamically re-lightable images (Mudge *et al*. 2008);
- 3D models (triangulated meshes, point clouds and textured models).

For each media type, automatic conversion to an efficient multi-resolution representation is supported, offering data compression, progressive transmission and view-dependent rendering; each data type has a specific web-browser, implemented using Web-GL and appearing in a standard web page (see Figure 8).

The new features also allow for further personalisation of the page: it is now possible to change the navigation paradigm and the style of the page. Moreover, new tools (i.e. creating sections and for taking point-to-point measurements) have been made available, and they can be added to the visualisation page.

3.4.2 The ARIADNE Landscape Service

The Landscape Service is a set of online services for the processing, management and publication of large, multi-resolution 3D interactive terrain datasets within a collaborative workflow (see Figure 8). The goals within this service are to: (1) aid and support 3D landscape reconstruction tasks and projects in Virtual Archaeology and (2) provide online services for dissemination of interactive landscapes, through several devices. The Landscape Services are designed for responsiveness,

thus adapting to both desktop and mobile devices such as smartphones and tablets. Data management is performed through a cloud service, allowing fine-grained access control on input/output data and collaborative approaches among research institutions and professionals, with specific focus on input DTMs/DEMs, imagery and shape files. The Terrain service allows generation and publication of 3D datasets by presenting different options to control format, resolution and dissemination segment; the service will then take care of multi-resolution, geometry/texture compression and much more. The Gallery service allows the user to control, update or delete their projects.

The WebGL front-end provides a high level of customisation and several features including:

- Paged multi-resolution on desktop and mobile browsers for efficient streaming; camera and Point-of-View management;
- Embed options;
- Metadata presentation;
- Support for touch and multi-touch devices;
- Multi-texturing and spherical panoramas.

3.5 Item-level integration

Among the emerging needs of the archaeological research community is the ability to answer a research question by using relevant information from several available heterogeneous sources. This can be achieved only with the integration of rich structured information from all such sources through a common, consistent representation of data that have a potential bearing on questions beyond their local context of creation and use, so that directly and indirectly related facts can be filtered out effectively from the mass in order to support further interpretation by the researcher.

In order to address the complexity of archaeological data integration, the main challenge for ARIADNE was to develop a global, extensible schema in the form of a formal ontology that allows for integration without loss of meaning. The CIDOC-CRM (Doerr 2003) (version 6.2 available from http://83.212.168.219/CIDOC-CRM/Version/version-6.2) was chosen as the backbone of the ARIADNE reference model and a suite of extensions was developed to address the complexity of archaeological data integration.

The CIDOC-CRM (ISO21127) is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It started from the bottom up, by re-engineering and integrating the semantic contents of more and more database schemata and documentation structures from all kinds of museum disciplines, archives and, recently, libraries as an empirical base. The CIDOC-CRM contains the most basic relationships to describe what happened in the past at a human scale, i.e. people and things meeting in space-time, parts and wholes, use, influence and reference. More detailed kinds of discourse require extensions (See Box).

Having defined the ARIADNE Reference Model (RM), integration is accomplished by creating an advanced knowledge base (target, aggregation database) based on the common reference model. The integrated knowledge base is the aggregation of several existing archaeological databases that were transformed by mapping their individual schemata (source schemata) into the ARIADNE RM (target schema). The mapping process was supported by the X3ML Mapping

Framework (See Box), ensuring the integrity of the initial data and preserving their initial 'meaning'.

In order to demonstrate the item-level integration process of archaeological datasets, ARIADNE has chosen as a use case the numismatics field, a highly standardised field with widely available data. Five datasets were selected (See Box). Four of them have been mapped to the ARIADNE RM and transformed to RDF using the X3ML framework while the fifth is already in CIDOC-CRM RDF form, compatible with the ARIADNE RM, and was extracted via OAI-PMH. As a common thesaurus for the aggregated knowledge base, the AAT and nomisma.org were adopted as the most appropriate resource in numismatics.

The mapping and transformation workflow is presented in Figure 10. 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. The ARIADNE portal will provide a main access point to the integrated repository and an intuitive interface will guide the user to formulate queries, browse the results and refine the search with facet view. We plan to implement a query interface that will take advantage of the principles of the Fundamental Categories and Fundamental Relationships (Tzompanaki and Doerr 2012; Tzompanaki *et al*. 2013). Potential research questions that need to be supported include:

- 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

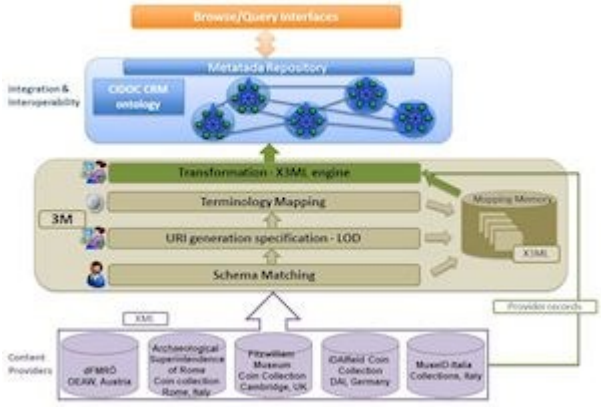

Figure 10: The mapping and transformation workflow

Such queries might appear trivial if answered by each dataset separately; however, they become important if they can be addressed by the aggregated repository. Results from our first experimental aggregated repository are quite promising (Meghini *et al*. 2015).

3.6 Natural Language Processing services

The archaeological domain generates vast quantities of text, including journal articles and reports of fieldwork or specialist analysis not formally published (grey literature). This text information is frequently difficult to access and opaque to computer-based tools for cross searching or meta analysis. This has become recognised as a significant problem for archaeological research. ARIADNE is addressing this issue, particularly as regards grey literature, by experimenting with text analysis methods based on Natural Language Processing (NLP) techniques for information extraction. The ultimate aim within ARIADNE is to extract additional relevant subject metadata from these reports and express it using the same ontological (CIDOC-CRM) and vocabulary standards as those used to describe archaeological datasets within the catalogue. This is a long-term goal beyond the reach of the immediate project. Nonetheless, some initial investigations point the way for further research.

Information Extraction is a specific NLP text analysis technique that extracts targeted information from context. This technique analyses textual input to form a new textual output capable of further

manipulation. There are two types of NLP information extraction techniques: rule-based and machine learning (Richards *et al*. 2015). The aim within ARIADNE is to investigate both approaches; each has its respective strengths and weaknesses and the ARIADNE partners will explore both to assess their usefulness within the archaeological domain.

Rule-based techniques have been employed with available archaeological vocabularies from Historic England (HE) and Rijksdienst Cultureel Erfgoed (RCE). This builds upon previous work with the grey literature digital library from the ADS, which proved capable of semantic enrichment of English language grey literature reports conforming both to archaeological thesauri and corresponding CIDOC-CRM ontology classes representing archaeological entities, such as Artefacts, Features, Monuments Types and Periods. The current pilot system has achieved some promising semantic enrichment of Dutch grey literature reports, for example artefacts such as 'pottery/aardewerk' (via the RCE Archeologische artefacttypen vocabulary) and other concepts including time periods. Work extending the techniques to develop a Swedish language pipeline is underway. The resulting NLP tools will be available via the ARIADNE portal.

Machine-learning work has focused on an English language user interface that can be used by archaeological practitioners to automatically generate metadata related to uploaded, text-based content on a per file basis, or using batch creation of metadata for multiple files. Two sets of training data have been used, one produced by human annotators and the other using a rule-based machine annotator. Human annotations are considered to have high potential for providing detailed examples for the machine-learning algorithms but are very resource intensive to produce. To address this, a web application interface has been developed, which will allow domain experts to annotate reports, generate resource discovery metadata where none exists, and generate metadata that can be used to further train the classifiers. This will be a useful feature, which can be

used to produce more training data in the future, and also provide an intuitive interface for users to correct results, which can then be used by the training classifier.

A common problem in text mining is the issue of gaining 'false positive' hits from statements which actually assert an absence of evidence, such as "No remains dating to the Roman period were identified". English language NLP research has investigated the issue of negation detection in archaeological grey literature reports, with a view to distinguishing a finding of evidence, for example, of Roman activity from statements reporting a lack of evidence, or no sign of Roman remains. A rule-based technique previously used in the biomedical domain was adapted to archaeological vocabulary and writing style. This technique was applied to detecting negated instances of the CIDOC CRM entities, Physical Object, Time Appellation, Place and Material. An evaluation exercise on ten grey literature documents from a range of UK archaeological units gave promising results, with overall Recall at 80% and Precision 89% (Vlachidis and Tudhope 2015). Further research on the semantic integration of archaeological datasets with grey literature reports, would be valuable, including negation detection (e.g. a negative finding) and the ability to discriminate in reports between important findings of archaeological evidence and less important information.

4. Evaluation

The effectiveness of the ARIADNE infrastructure in providing services to its research community will be evaluated in time by measuring the quantity and quality of usage of the services by archaeological researchers. However, during the project lifetime an initial evaluation is underway. This section describes this evaluation, focusing on two central respects: the adequacy of the catalogue, which provides an overview of the ARIADNE information space and supports the discovery functionality of the infrastructure; and the plan for evaluating the remaining services.

4.1 Contents of the ARIADNE catalogue

The ARIADNE consortium consists of 24 partners in 16 countries including Sweden, United Kingdom, Ireland, Germany, Austria, Hungary, Czech Republic, Slovenia, France, the Netherlands, Italy, Spain, Greece, Cyprus, Romania and Bulgaria. The ARIADNE discovery service has been developed to create a single global access point, which provides open access to integrated archaeological information and supports researchers and professionals, educators and students as well as the wider interested public.

After ingesting the metadata of the ARIADNE consortium into the catalogue, this service has been evaluated several times. Adaptations of the ACDM took place in order to maximise the effect that the services created by ARIADNE will have on the different stakeholders. The different metadata schemas have commonalities that allowed mapping to each other. Crosswalks to establish and provide an integrated approach can be made. Improved thesauri are helping to overcome linguistic barriers by linking related terms expressed in different languages.

Table 2: Contents of the ARIADNE catalogue (as of March 2016)

Data Resources	Data Resource Properties	Data Resource Types
		Sites and monuments databases or inventories 1,529,498
	Spatial 98%	Event/intervention resources 51,820
Datasets 1,534,375	Temporal 100%	Artefact databases or image collections 40,726
Collections 43,182	Native Subject 97%	Scientific datasets 4,904

Textual Documents 50,807	Derived Subject 48%	Fieldwork archives 1,340
Total 1,628,364	Publisher 100%	Burial databases 76

Table 2 gives an overview of the current contents of the catalogue. All descriptions provided by the ARIADNE partners could be mapped to the ACDM and therefore inserted into the catalogue. The numbers are already significant, covering a large percentage of the data made available by the ARIADNE partners. Further additions are expected before the end of the project. Above all, it is expected that opening the catalogue to the whole archaeological community will bring other descriptions, further enlarging the ARIADNE information space.

4.2 Planned service evaluation

The French National Institute for Preventive Archaeological Research (Inrap) is in charge of testing the integrated services produced within ARIADNE. The evaluation was undertaken using two complementary methods: predefined testing scenarios and open evaluation questionnaires. The aim of the questionnaire, related to a specific service, was to determine whether the service meets the expectations of the users. The questionnaire asks precise questions about usability of the service; open comments (usability, request for improvements and so on); a score (from 1 to 5); and quantitative data about usage (e.g. number of downloads, number of running processes, number of files uploaded, etc.). A quarterly analysis of the data results is planned. The evaluation process for any service requires the full availability of the service, in a stable version. It also requires the availability of a comprehensive set of data appropriate to using the service.

A first testing phase was completed in early 2016. A panel of 30 testers evaluated three main services developed within the project, namely the Portal, the Visual Media Services, and the Landscape Factory. The first

results demonstrated a great interest in the services. The tests showed a high approval rating: 3.86/5.0; 4.14/5.0; 4.0/5.0, respectively.A second testing phase was conducted from July to December 2016. This focussed on the usability of the ARIADNE infrastructure as a whole, the ARIADNE portal being the entry point. Informal tests were undertaken within a community of internal power users, who were all experienced professional archaeologists. This revealed considerable interest in the services provided for visualisation of 3D, RTI and HR images and for treatment of geographic data (DEM, lidar and so on), but also the need for enhanced visualisation and analysis tools, especially measuring and conversion tools.

5. Conclusions and Outlook

This article has presented the ARIADNE infrastructure, an ongoing European initiative that aims to create a single information space, where the data and the services owned by European archaeological institutions can be discovered and accessed through a single search facility. After reviewing the current landscape of research infrastructures in archaeology, a set of requirements has been distilled and addressed by technological developments. At the heart of the ARIADNE infrastructure is the ARIADNE catalogue, which describes the elements of the ARIADNE information space and supports their discovery and access. The catalogue is the result of a major effort; first, an adequate data model has been created and validated by the members of the consortium. An aggregation infrastructure has then been set up for populating the catalogue, by transforming the descriptions collected from the contributing partners. Finally, the discovery functionality has been implemented by relying on the catalogue contents and on the state-of-the-art search engine provided by Elasticsearch. Thanks to this effort, the archaeological community has a unique access point where its resources can be found. Having tested the infrastructure with a broad range of archaeological content drawn from the ARIADNE partners, the catalogue and its supporting technology will be

opened up for contributions from the broader archaeological community, thereby becoming a global knowledge source for archaeology.

ARIADNE has also started to address the item-level integration of archaeological data, by conducting Linked Data experiments on data related to coins. For this, it has relied on the pivotal role of a well-known ontology for cultural data integration, the CIDOC-CRM, and on the associated technology for mapping and transformation. The experiment has been described here in some detail, since it is key to an important future development, namely the creation of knowledge aggregations where researchers can find answers to research questions spanning several datasets. The experiment has had encouraging results, and will form an important item for the future. ARIADNE has also begun to tackle the sharing of services, making services on visual data and on reference resources accessible through the infrastructure.

Nonetheless, it is clear that much remains to be done before archaeology has a mature research infrastructure.

- Data integration needs to be undertaken more systematically, by making the available tools and resources available to the community through a state-of-the-art Virtual Research Environment, where domain experts can convene and collaborate to develop the necessary transformation rules and to apply those rules to create integrated data.
- A permanent conduit needs to be created through which archaeologists can channel their requirements to the relevant technological research and development communities, who can then respond with matching technology.
- The work of researchers in archaeology needs to be supported in a more substantial way, by endowing the infrastructure with the ability to understand and manage the knowledge generation process. This support requires the provenance of the data found in

the infrastructure to be tracked, and the possibility of defining, sharing and executing complex workflows for processing the data.

All of this is within the reach of current ICT technology, but it requires investment and institutional support in order to achieve it. ARIADNE has been made possible through European funding, which has allowed archaeologists and information scientists throughout Europe to collaborate on solving some major issues, and it has already gained widespread recognition. The European Archaeological Council (EAC) has strongly encouraged organisations to participate in the ARIADNE initiative. The EAC comprises heads of national services responsible under law for the management of the archaeological heritage in the Council of Europe member states. In their Amersfoort Agenda, setting the agenda for the future of archaeological heritage management in Europe, the Council emphasised 'the need to share, connect and provide access to archaeological information with the help of digital technologies. The key to this aspiration is to improve collaboration – we need to share rather than exchange. It is essential to encourage the development of European data-sharing networks and projects in the field of archaeology. The ARIADNE project is an excellent European initiative in this regard and participation in this project should be strongly encouraged' (EAC 2015, 21).

The European Strategy Forum on Research Infrastructures (ESFRI) in its 2016 Roadmap has also acknowledged the success of ARIADNE in building a (digital) research community, 'quickly growing in the field of archaeology', and its role as the leading integrator of archaeological research data infrastructures: 'In the archaeological sciences the ARIADNE network developed out of the vital need to develop infrastructures for the management and integration of archaeological data at a European level. As a digital infrastructure for archaeological research ARIADNE brings together and integrates existing archaeological research data infrastructures so that researchers can use the various distributed datasets and technologies' (ESFRI 2016, 52, 175).

With the inclusion of Heritage Science within the 2016 ESFRI Roadmap, and the strong engagement of ARIADNE with the nascent European Research Infrastructure for Heritage Science (E-RIHS), the foundations have been laid to place archaeology at the forefront of European research infrastructures, ensuring a sustainable future for the ARIADNE portal and services.

Bibliography

3m2 2015 *3M The Mapping Memory Manager*. http://www.ics.forth.gr/isl/3M

ARIADNE 2014a *The Way Forward to Digital Archaeology in Europe*, November 2014. http://www.ariadne-infrastructure.eu/Media/Files/Ariadne-Booklet-The-Way-Forward-to-Digital-Archaeology-in-Europe

ARIADNE 2014b *First Report on Users Needs*, Deliverable D2.1. http://www.ariadne-infrastructure.eu/Resources/D2.1-First-report-on-users-needs

ARIADNE 2015a *Second Report on Users Needs*. Deliverable D2.2. http://www.ariadne-infrastructure.eu/fre/Resources/D2.2-Second-report-on-users-needs

ARIADNE 2015b *Preliminary Innovation Agenda and Action Plan*, Deliverable D2.3. http://www.ariadne-infrastructure.eu/Resources/D2.3-Preliminary-Innovation-Agenda-and-Action-Plan

Aspöck, E. and Geser, G. 2014 'What is an archaeological research infrastructure and why do we need it? Aims and challenges of ARIADNE', *Proceedings of the 18th International Conference on Cultural Heritage and New Technologies (CHNT 18), Vienna, November 2013*, http://www.chnt.at/wp-content/uploads/Aspoeck_Geser_2014.pdf

Binding, C. and Tudhope, D. 2016 'Improving interoperability using vocabulary linked data', *International Journal on Digital Libraries* **17**, 5–21. https://doi.org/10.1007/s00799-015-0166-y

Caplan, P. 2003 *Metadata Fundamentals for all Librarians*, Chicago: American Library Association.

CRMarchaeo 2014 *CRMarchaeo: the Excavation Model, version 1.2.1*. http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo1.2.1.pdf

CRMdig 2014 *CRMdig: an Extension of CIDOC-CRM to support provenance metadata, version 3.2*. http://www.ics.forth.gr/isl/CRMext/CRMdig/docs/CRMdig3.2.pdf

CRMinf 2015 *CRMinf: the Argumentation Model, version 0.7*. http://www.ics.forth.gr/isl/CRMext/CRMinf/docs/CRMinf-0.7.pdf

CRMsci 2014 *CRMsci: the Scientific Observation Model, version 1.2.2*. http://www.ics.forth.gr/isl/CRMext/CRMsci/docs/CRMsci1.2.2.pdf

dFM 2007 *dFMRÖ – digitale Fundmünzen der Römischen Zeit in Österreich*. http://www.oeaw.ac.at/antike/index.php?id=358

Doerr, M. 2003 'The CIDOC Conceptual Reference Model: an ontological approach to semantic interoperability of metadata', *AI Magazine* **24**, 75–92.

European Archaeological Council (EAC) 2015 'Amersfoort Agenda – setting the agenda for the future of archaeological heritage management in Europe' in P.A.C. Schut, D. Scharff and L.C. de Wit (eds) *Setting the Agenda – Giving New Meaning to the European Archaeological Heritage*, EAC Occasional Paper **10**, Budapest: Archaeolingua. 15-24.

European Science Foundation (ESF) 2011 *Research Infrastructures in the Digital Humanities*, Science Policy Briefing 42, Strasbourg, September 2011. archives.esf.org/fileadmin/Public_documents/Publications/spb42_RI_DigitalHumanities.pdf

European Strategy Forum on Research Infrastructures (ESFRI) 2016 *Strategy Report on Research Infrastructures: Roadmap 2016*. StR-ESFRI and ESFRI Secretariat. http://www.esfri.eu/sites/default/files/20160309_ROADMAP_browsable.pdf

Hansen, H.J. 1993 'European archaeological databases: problems and prospects' in J. Andresen, T. Madsen and I. Scollar (eds) *CAA92. Computing the Past. Computer Applications and Quantitative Methods in Archaeology*, Aarhus: University Press. 229-37.

Heery, R. 1997 'Naming names: metadata registries', *Ariadne* **11**, http://www.ariadne.ac.uk/issue11/metadata

Heery, R., Gardner, T., Day, M. and Patel, M. 2000 *DESIRE metadata registry framework*, Deliverable 3.5. http://web.archive.org/web/20080607052045/http://www.desire.org/html/research/deliverables/D3.5/

Hiebel, G., Doerr, M. and Eide, Ø. 2016. 'CRMgeo: A spatiotemporal extension of CIDOC-CRM, *International Journal on Digital Libraries* **17**, 1-9. https://doi.org/10.1007/s00799-016-0192-4

Isaac, A., Charles, V., Fernie, K., Dallas, C., Gavrilis, D. and Angelis, S. 2013 'Achieving interoperability between the CARARE schema for Monuments and Sites and the Europeana Data Model' in *Proceedings of the International Conference on Dublin Core and Metadata Applications, DC-2013*, Lisbon, Portugal. 115–125.

ISO 2004 *Framework for the Specification and Standardization of Data Elements*, Technical Report ISO 11179 Part 1.

ISO 2013 *Information and Documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies*, Technical Report 25964-2:2013.
ISO. http://www.iso.org/iso/catalogue_detail.htm?csnumber=53658

Jarrett, J., Zambanini, S., Hüber-Mork, R. and Felicetti, A. 2011 'Coinage, digitization and the world-wide web: numismatics and the COINSProject', *New Technologies in Medieval and Renaissance Studies* **3**, 459–89.

Jeong, D., Baik, D. and Park, S. 2003 'A practical approach: localization-based global metadata registry for progressive data integration', *Journal of Information and Knowledge Management* **2**, 391–401. https://doi.org/10.1142/S0219649203000577

Kenny, J. and Richards, J.D. 2005 'Pathways to a shared European information infrastructure for cultural heritage', *Internet Archaeology* **18** https://doi.org/10.11141/ia.18.6

Kilbride, W. 2004 'The Danube in prehistory in the digital age: towards a common information environment for European archaeology', *Archeologia e Calcolatori* **15**, 129-44.

Kondylakis, H., Flouris, G., Theodoridou, M., Doerr, M., Minadakis, N., Marketakis, Y. and de Jong, G. 2015 'X3ML framework: an effective suite for supporting data mappings' in *Proceedings of the Workshop: Extending, Mapping and Focusing the CRM*, TPDL 2015. Poznan, Poland.

Meghini, C., Theodoridou, M., Felicetti, A. and Gerth, P. 2015 'Integrating heterogeneous coin datasets in the context of archaeological research'

in *Proceedings of the Workshop: Extending, Mapping and Focusing the CRM*, TPDL 2015. Poznan, Poland.

Mudge, M., Malzbender, T., Chalmers, A., Scopigno, R., Davis, J., Wang, O., Gunawardane, P., Ashley, M., Doerr, M., Proenca, A. and Barbosa, J. 2008 'Image-based empirical information acquisition, scientific reliability, and long-term digital preservation for the natural sciences and cultural heritage', *Eurographics Tutorials*, The Eurographics Association. https://doi.org/10.2312/egt.20081050

Nagamori, M. and Sugimoto, S. 2006 'A metadata schema registry as a tool to enhance metadata interoperability', *TCDL Bulletin* **3**. http://www.ieee-tcdl.org/Bulletin/v3n1/nagamori/nagamori.html

Niccolucci, F. and Richards, J.D. 2013 'ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe', *International Journal of Humanities and Arts Computing* **7**, 70–88. https://doi.org/10.3366/ijhac.2013.0082

Ponchio, F., Potenziani, M., Dellepiane, M., Callieri, M. and Scopigno, R. 2015 'ARIADNE Visual Media Service: easy web publishing of advanced visual media' in S. Campana, R. Scopigno, G. Carpentiero and M. Cirillo (eds) *CAA2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, Oxford: Archaeopress. 433-42.

Richards, J.D., Tudhope, D. and Vlachidis, A. 2015 'Text mining in archaeology: extracting information from archaeological reports' in J. Barcelo and I. Bogdanovic (eds) *Mathematics in Archaeology*, Boca Raton: CRC Press. 240-54. https://doi.org/10.1201/b18530-15

Ronzino, P., Niccolucci, F., Felicetti, A. and Doerr, M. 2016 'CRMba a CRM extension for the documentation of standing buildings', *International*

Journal on Digital Libraries **17**, 71-8. https://doi.org/10.1007/s00799-015-0160-4

Tzompanaki, K. and Doerr, M. 2012 *Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM Based Repositories*, Technical Report TR-429, Crete: ICS-FORTH.

Tzompanaki, K., Doerr, M., Theodoridou, M. and Fundulaki, I. 2013 'Reasoning based on property propagation on CIDOC-CRM and CRMdig based repositories' in V. Alexiev, V. Ivanov and M. Grinberg (eds) *Practical Experiences with CIDOC CRM and its Extensions. Proceedings of the Workshop Practical Experiences with CIDOC CRM and its Extensions*, Valetta, Malta. 37-47.

Vlachidis, A. and Tudhope, D. 2015 'Negation detection and word sense disambiguation in digital archaeology reports for the purposes of semantic annotation', *Program* **49**, 118-34. https://doi.org/10.1108/PROG-10-2014-0076