Joined up Thinking: Aggregating archaeological datasets at an international scale

Julian D. Richards

The archaeological research community was an early adopter of digital tools for data acquisition, organisation, analysis, and presentation of research results of individual projects. However, the provision of e-infrastructure and services for data sharing, discovery, access, and reuse has lagged behind. The ARIADNE Research Infrastructure has sought to address this situation. Developed with European funding, ARIADNE has created an e-infrastructure that enables data providers to register and provide access to their digital resources through the ARIADNE data portal, facilitating discovery, access, and research. ARIADNE has aggregated resources from over 45 data providers, spanning over 40 countries and 4 continents. The portal now provides online access to over 3.9 million research resources. It is based upon Linked Open Data technologies and is underpinned by a flexible and extensible architecture, enabling multiple combinations and presentations of the same underpinning data. We have been keen not to 'make a great heap' of all the data and, learning from previous data aggregation projects, we have defined a subset of the CIDOC CRM to be used as a strict ontology and paid close attention to data standards and controlled vocabularies to achieve a high degree of interoperability. This article discusses some of the challenges of large-scale data integration and describes the approaches adopted to ensure that the ARIADNE Knowledge Base is an effective tool for archaeological heritage management and research at a national and international level.

1. Background: The Prehistory of Archaeological Data Aggregation

In 2013 there were no more than six digital repositories providing facilities for long-term preservation and access to archaeological data at a national level (Richards 2017 233). In the last decade that situation has improved, although it is

still patchy. A 2021 survey, undertaken under the auspices of ARIADNEplus and the SEADDA COST Action, collected information about 60 repositories, of which 43 were operational, and 17 were being set up (Geser *et al.* 2022). Nonetheless, there are still only a handful of archaeological data repositories holding the CoreTrustSeal accreditation (ADS, Canmore, DANS, SND, tDAR, and ARCHE). The SEADDA and European Archaeological Council State of the Art survey revealed a very mixed picture, with a lack of enforceable deposit mandates as well as of facilities in most countries (Richards *et al.* 2021). It remains clear, however, that in the heritage domain, national infrastructures are the appropriate level for data curation. This is less straightforward in the physical and chemical sciences. Where data are collected via shared international facilities (such as CERN) or where the data relate to phenomena that transcend Planet Earth (such as in Astrophysics) then national repositories make little sense.

In the heritage sector, however, data are generally collected within national legal frameworks and guidelines, and relate to features that have fixed geospatial coordinates that exist within political borders. Exceptions exist in some federal countries, including Germany where heritage protection operates at the state level (e.g. Bibby 2021), and in the United States (Nicholson *et al.* 2021) although in the latter the main sponsors of work that creates archaeology tend to be national federal agencies. But generally, archaeology is organised at a national state level, and national heritage agencies take a lead in the stipulation of data standards, as well as in directing research frameworks and priorities, even if they are enforced by regional bodies. In the university and research sector the majority of funding is subject to national rules about data deposit and Open Access, although these may mirror international trends. In addition, it is worth noting that the sustainability of digital repositories capable of dealing with the wide variety of archaeological data types and formats requires a critical mass of funding, staffing and expertise, which in many countries is not feasible at a regional level. Repositories also need to have organisational backing and long-term commitment, and if their resources are also to be 'trusted' in the more conventional sense then some reputable institutional imprint is essential.

There is certainly a risk that a network of repositories operating at national level creates a series of data silos. In fact, digital repositories need to do two things: they must bring resources together and make it easy for users to interrogate them via shared and user-friendly interfaces, but they must also open data up: via APIs, harvesting protocols such as OAI-PMH, and as Linked Open Data (LOD) so that it can be viewed and manipulated by multiple routes (May *et al.* 2015). It is also important to acknowledge concerns raised by Huggett (2015) that in aggregating data we take the data outside the context in which they were collected in the first place, and in the absence of paradata describing that context we risk misuse. As we aggregate data at a higher and higher level the risks increase, as we may divorce data from the cultural and societal values in which it was recorded. The metadata standards we employ were generally produced by archaeologists educated in a western positivist tradition, and we need to pay attention to the CARE as well as the FAIR data principles to avoid inadvertently imposing post-colonial perspectives. In this regard, the work of the TETRARCHs project in exploring other types of metadata to encourage different forms of data reuse is important. Nevertheless, archaeological research questions rarely coincide with modern political boundaries,

which were irrelevant for the vast majority of the timespan of the human past. Big research questions require big data, and it is our role to catalogue it with sensitivity, and with regard to all the nuances of human observation and different perspectives.

As early as 1992 Henrik Jarl Hansen expressed the need to join up digital resources across Europe (Figure 1) (Hansen 1992), and the European Commission has been a key funding agency in facilitating several projects in this area. The first was AQUARELLE (1996-99), which designed a resource discovery architecture enabling query broadcasting to distributed and heterogeneous data sources from the museums and heritage domain (Michard *et al*. 1998). The project collaborated with the CIMI consortium of American museums and libraries to use a Z39.50 Application Profile to enable query broadcasting and a software system to create and manage multilingual terminology resources.
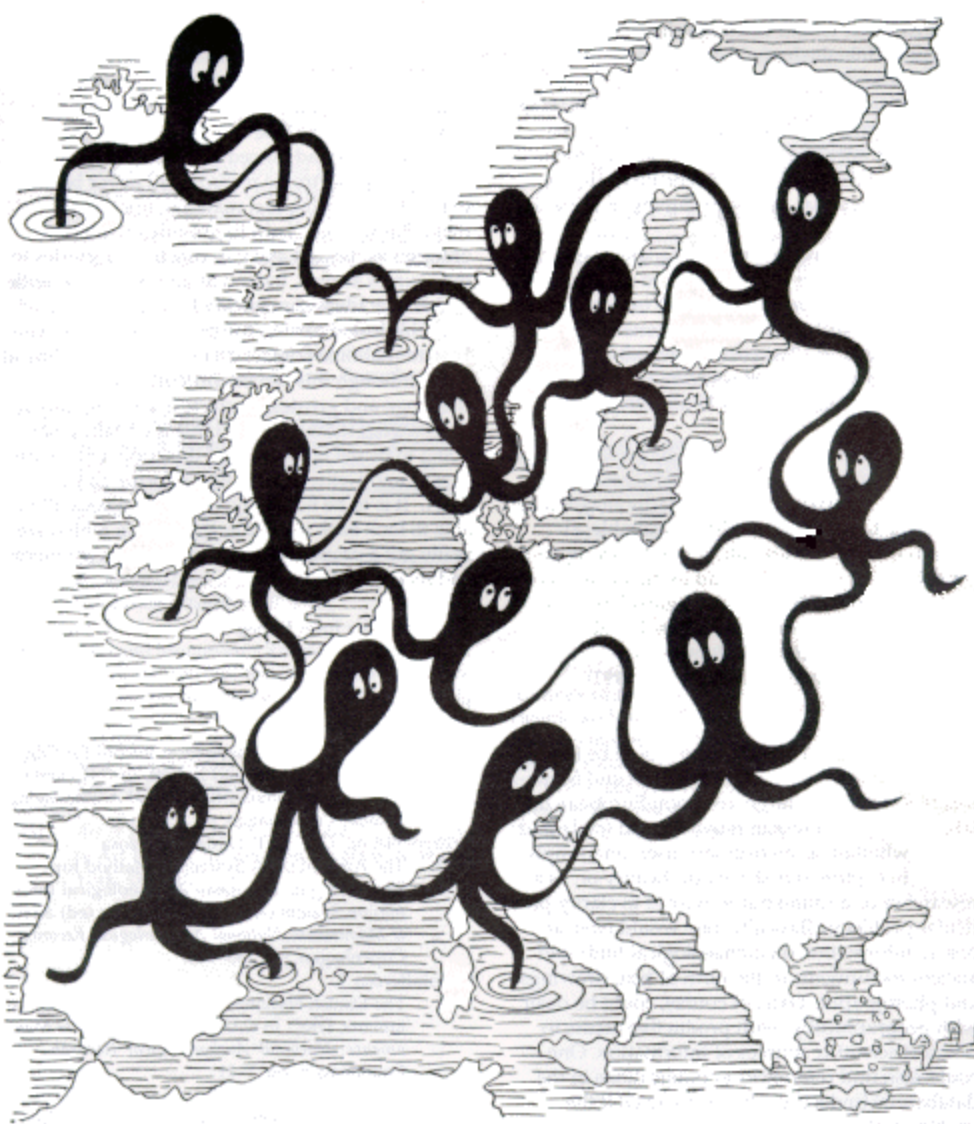
Figure 1: Hansen's octopi symbolising the networking of European resources (reproduced from Hansen 1992)

In 2002-4 the Archaeology Data Service (ADS) led a consortium of European partners on the ARENA project, funded under the Culture 2000 programme (Kenny and Richards 2005). One of the outcomes of the project was a portal that provided a distributed cross-search of sites and monuments records for six countries: Denmark, Iceland, Norway, Poland, Romania and the United Kingdom (Figure 2) (Dam and Hansen 2005). The ARENA portal was completed in 2004 and launched at the European Association of Archaeologists conference at Lyon.

Figure 2: The landing page of the ARENA search portal, developed during the ARENA project 2002-2004, provided a distributed cross-search of catalogues of sites and monuments data for Denmark, Iceland, Norway, Poland. Romania and the United Kingdom

ARENA provided a proof-of-concept demonstrator for a search paradigm later to be adopted by ARIADNE, based around a powerful triumvirate of 'Where', 'When' and 'What' parameters, enabling users to identify resources of interest in a few mouse clicks. Partners provided high-level subject-term mappings to the UK MIDAS data standard Thesaurus of Monument Types and used their local controlled vocabularies for period terms (Figure 3). There was no attempt, however, to provide a temporal search using absolute From and Until time spans.
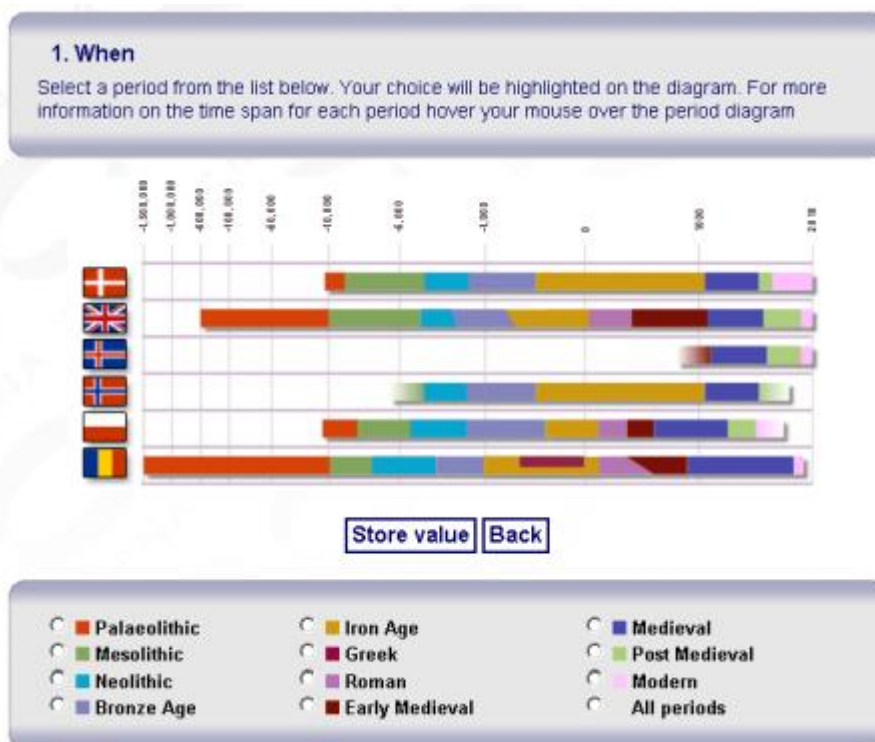
Figure 3: The ARENA temporal search provided a visualisation of the different start and end dates for cultural period terms applied by each of the partners and users could select all resources for which the same period term was used, but searches using absolute date ranges were not possible

The ARENA portal demonstrated that a live search across multiple European data providers was possible, but it relied upon dated technologies such as the Z39.50 communications protocol, which had been developed for cross-searching library catalogues. The same technology had been deployed initially for the UK portal for the Arts and Humanities Data Service catalogue in the late 1990s and had been further developed by the ADS for the UK HEIRPORT portal for Historic Environment Information Resources (Austin *et al*. 2002). In 2009-10 the ADS was able to work with DANS as part of the preparatory phase for DARIAH, the Digital Research Infrastructure for the Arts and Humanities, to migrate the ARENA portal into a more flexible web services architecture (Figure 4). ARENA2 changed the technical approach of ARENA to one based on using a Service-Oriented Architecture (SOA) over selected partner data centres to demonstrate the viability of using a SOA approach rather than the original Z39.50 and OAI metadata harvesting approach. A SOA approach differs from metadata harvesting approaches in that it allows direct, live access to remote databases making the most current data available to be queried and minimising harvesting and data management on the part of the aggregator.
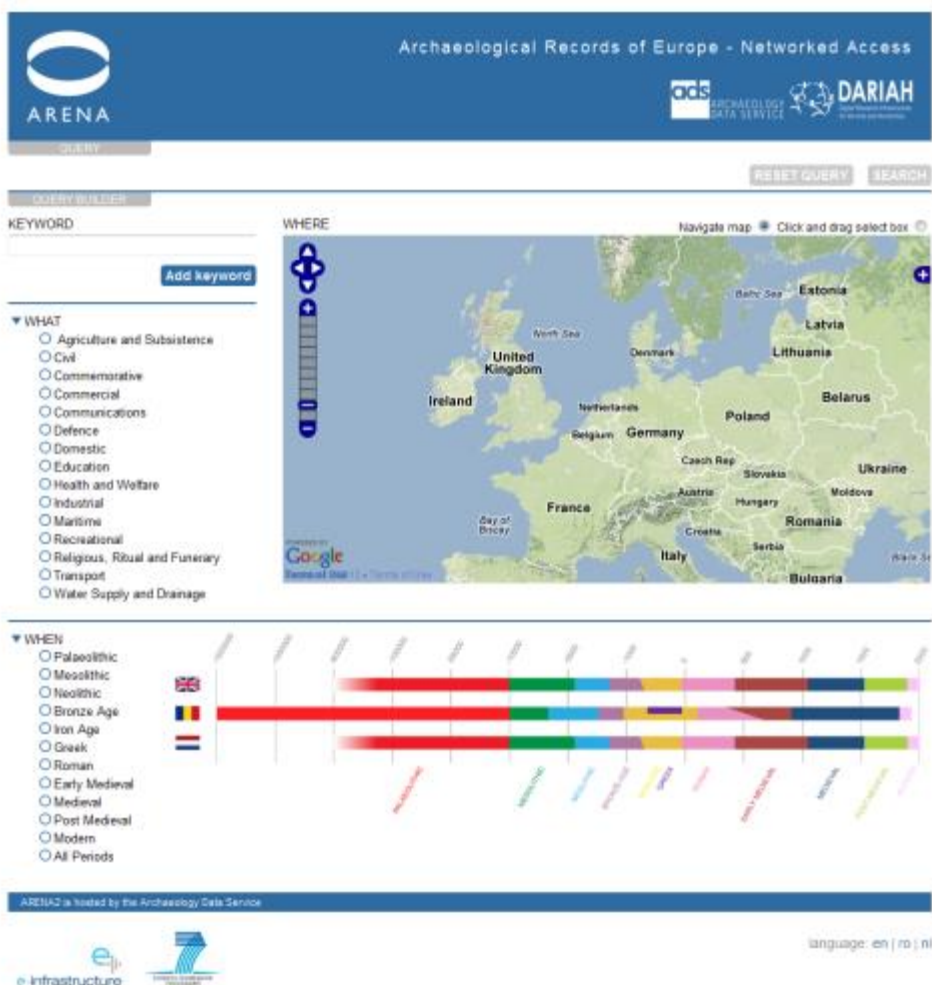
Figure 4: The ARENA2 portal: developed as a technical demonstrator for DARIAH

The same technological infrastructure was also employed in a collaborative project between the ADS and tDAR to build a Transatlantic Archaeology Gateway (TAG) searching across UK and US data archives (Jeffrey *et al*. 2012). The first stage was to create an infrastructure to enable basic cross-search of Dublin Core compatible metadata records for digital resources covering the archaeology of the USA and UK, with additional challenges with regard to periodisation and subject type (Figure 5). The second stage of TAG was an attempt to develop a much deeper and richer level of cross-searching for a specific subset of archived data: faunal data from North America and Europe. This sub-discipline was chosen as there is a relatively high level of agreement over basic classifications. The project provides an early example of the application profile approach to metadata harvesting that was later adopted in ARIADNEplus.

Figure 5: The Transatlantic Archaeology Gateway (2009-11), showing a cross-search across the archives of the ADS and tDAR

These early demonstrators provided examples of the scale of research questions that the technology could support. However, all were funded as research and development projects, and while there was continued support for the underpinning repositories at a national level, there was no obvious funding source for international infrastructure, and none sat within a sustainable framework. The potential exception was Europeana, an initiative to bring together in a virtual manner the rich collections of Europe's museums, galleries and archives. The initial data model - the Europeana Semantic Elements (ESE) - was based on the Dublin Core metadata scheme and its treatment of temporal and spatial coverage was inadequate for the archaeological domain (Hansen and Fernie 2010). In 2010 this was upgraded to the Europeana Data Model (EDM), a semantically richer schema that takes account of the event-based principles underpinning the CIDOC Conceptual Reference Model (CIDOC CRM).

A number of aggregator projects were funded and some of these are discussed by Vassallo *et al.* (2023) but those with a major archaeological component were CARARE (2010-13) and LoCloud (2013-16). CARARE had a focus on archaeology and architecture, with an additional emphasis on 3D data (Hansen and Fernie 2010), while LoCloud emphasised local content from small and medium-sized museums. Both projects adopted a subset of the CIDOC CRM ontology as their core metadata model, although only a subset of this could be exposed via Europeana. While the Europeana portal now provides access to over 50 million items its

coverage is so broad and its search interface so generalised that it can be inadequate for archaeological research, and it is not well known among archaeologists. That situation may change as Europeana tries to address the requirements of the common European data space. As Kilbride (2004, 136) highlighted: 'technology is not, alone, the answer'.

There is also growing interest in Linked Open Data in archaeology, providing mechanisms for data aggregation including services such as Pleiades (a community-built gazetteer of ancient places) and Pelagios (also joining up places in the Classical world), which focus on geospatial approaches to data interoperability (Isaksen *et al.* 2014). We should also note the implementation of Linked Open Data approaches by Open Context (Kansa 2014). This presents itself as a web-based data publication platform rather than a data aggregator, although in 2012 the same technical architecture was used to develop the Digital Index of North American Archaeology (DINAA) platform, which aggregates archaeological and historical datasets largely harvested from state and federal government agencies, and provides access to over 880,000 records via a map-based interface. The distributed nature of Linked Open Data also makes it a suitable mechanism for aggregation of data that spans several domains, with data managed and curated by domain specialists and the creation of tailored interfaces that reflect specific sets of user needs. Several projects funded within the UK's Towards a National Collection programme have explored the use of Linked Open Data as offering the only viable solution to providing national data views.

ARIADNE provides a new chapter in the history of archaeological data aggregation and provides the focus of this article. Funded by the European Commission research infrastructure programme, it aimed to provide joined-up access to archaeological data drawn from multiple countries in a way that would allow researchers to address new questions. ARIADNE sought to learn from past lessons by returning to the archaeological focus of ARENA and emphasising a mapping process driven by archaeological research user needs. It has developed a powerful cross-search portal, based upon Linked Open Data (Aloia *et al.* 2017; Meghini *et al.* 2017; Richards and Niccolucci 2019). To achieve interoperability across different European languages and cultures is challenging and adherence to data standards is essential for any level of semantic interoperability and cross-search. In ARIADNE national subject terms have been mapped to a common core standard (in this case the Getty Art and Architecture Thesaurus) and archaeological period terms have been defined according to explicit criteria, working with the North American PeriodO initiative (Shaw *et al.* 2016).

2. Data Aggregation in ARIADNE

ARIADNE has aggregated resources from over 45 data providers, spanning over 40 countries and 4 continents. As of May 2023, the ARIADNE portal provides online access to over 3.9 million research resources, with the number continuing to increase. The portal is underpinned by a triple store managed in Graph DB. At the time of writing the triple store used by the public portal comprises 322,218,928 triples.

The ARIADNE Research Infrastructure was developed during the first ARIADNE infrastructure project, with funding from the European Commission under Framework Programme 7 for the period 2013-2017. ARIADNE's goal was to provide open access to Europe's archaeological heritage and to overcome the fragmentation of digital repositories, placed in different countries and compiled in different languages (Niccolucci and Richards 2019). Innovative vocabulary mapping techniques brought interoperability to a huge and heterogeneous collection of texts, drawings, images, videos, 3D models and maps. By the end of the original ARIADNE project, we had succeeded in cataloguing *c.* 1.9m archaeological datasets, using the ARIADNE Common Data Model (ACDM) (Figure 6) (Aloia et al. 2017).

Figure 6: Screen shot of the original ARIADNE portal. The landing page provided a free text search box and the opportunity to browse the catalogue by the three key parameters of 'Where', 'When' and 'What'

In 2019 we commenced work on the second iteration of ARIADNE, with an enlarged consortium and funding from Horizon Europe. The data model was revised and simplified, also adopting some of the lessons from the PARTHENOS project, and was released as the AO-Cat (Felicetti *et al.* 2023). We were keen not to 'make a great heap' of all the data and, learning from previous data aggregation projects, a revised ontology, the AO-Cat was defined as a subset of the CIDOC CRM as a strict ontology. We also paid close attention to data standards and controlled vocabularies to achieve a high level of interoperability. The catalogue was migrated from a traditional database structure to a Linked Open Data triple store, avoiding the creation of another data silo, and opening up the Knowledge Base for linkage from and to other online resources.

Dimensions of integration
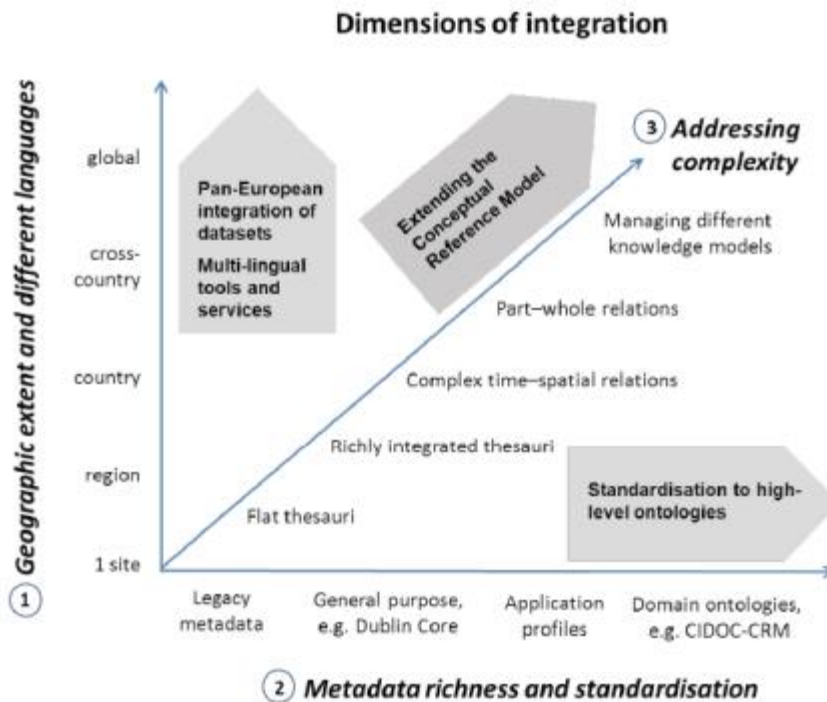
Figure 7: The ARIADNE approach for addressing the complexity of developing advanced integrated services for archaeological research communities (courtesy of ARIADNE project)

ARIADNEplus has brought together a much broader range of records than the predecessor ARIADNE project, upon whose success it has been able to build (Figure 7). The intention was to integrate datasets ranging widely in space and time, and across a wider range of archaeological sub-domains. By the end of the funded project, we have 7710 resources that pre-date 1,000,000 BP, and 517,535 resources that post-date AD 1900, demonstrating the full range of 21st-century archaeological research. We have resources from North and South America, Europe and Scandinavia, and Asia. The integrated data covers every ARIADNE resource type, from inscriptions and rock art to scientific analyses and dating resources. We have integrated over one million records for individual sites and monuments, over 855,250 artefacts and 470,000 coins. Our catalogue includes over 290,000 unpublished fieldwork reports, and over 110,000 full fieldwork archives. These represent a massive digital research infrastructure now available freely and easily as Open Access data to European scientists and researchers, teachers and students, as well as to interested members of the public.

The aggregation task will now be sustained through the work of the ARIADNE Research Infrastructure AISBL, a not-for-profit association, registered under Belgian law, but operating internationally. This is managed under a subscription model whereby members can pay an annual fee to update and add to their online resources. In this respect it is important to note that ARIADNE is a community as well as a technical infrastructure. Online research infrastructures can act as catalysts to bring together different communities of expertise and interest and ARIADNE has created a scholarly ecosystem, as well as a portal (see Benardou *et al.* 2017; Wright and Richards 2018).

3. The ARIADNE Ontology

The ARIADNE Ontology, AO for short, was developed at the start of the ARIADNEplus project for the purpose of integrating the archaeological data of the ARIADNE partners into a common information space. Building upon earlier projects, the CIDOC CRM, the standard ontology in the cultural heritage domain, was adopted as the conceptual backbone of the AO, providing a unified and coherent linguistic and axiomatic framework. Subsequently, the CRM has been specialised, under a specially devised namespace, to cater to the needs of the different subdomains tackled by the ARIADNEplus project. The first of those specialisations was the AO-Cat, the part of AO dealing with the representation of the resources in the ARIADNE Catalogue. The AO-Cat was used to build the ARIADNE Catalogue. The second round of specialisations are known as application profiles, which have been developed to support the integration of more specialised item-level data of the ARIADNE partners into the ARIADNE Knowledge Base.

The ARIADNE Catalogue provides a collection-level representation of the resources shared by the project partners. However, while the distinction between collection level and item level may seem obvious, it can be applied in many ways, according to how we define what we consider to be items. This in turn depends on the specific research context and research question. For example, at the level of landscape research, each archaeological site may be considered to be an item of observation, and the collection-level record refers to the database of national sites and monuments, whereas for an artefact-based study, the individual objects may be the items of interest, and the collection-level record now refers to the database of artefacts, potentially from across several sites.

In the first phase of ARIADNE we produced a catalogue that already included summary metadata records for archaeological sites and monuments at item level, and this approach was retained in ARIADNEplus. Experiments were also undertaken in more granular item-level integration, most notably investigating interoperability between several databases of coins held by different partners (Felicetti *et al*. [2015]). In ARIADNEplus such developments were extended by the adoption of application profiles.

Application Profiles are particular ontological models whose purpose is to describe the information relating to a specific research domain in a coherent and complete way. The philosophy behind the definition of an Application Profile is that of reusing the elements of existing ontologies and models to describe similar entities in the new research context, limiting the development of new elements only for the description of peculiar and typical aspects of the specific discipline. In ARIADNEplus, for instance, we selected classes of the CIDOC CRM ontology and from the Dublin Core standard to define application profiles for the description of scientific data and mortuary data at item level. However, it is worth noting that the AO-Cat is itself an application profile for what we might regard as higher level archaeological observations. It captures the basic 'What', 'When' and 'Where' information that powers the portal search interface, and provides an adequate representation for the discovery of resources relating to archaeological sites and monuments. It has also become apparent that it provides adequate core information for other sub-domains

where 'What', 'When' and 'Where' capture the core metadata and data, as in the case of archaeological artefacts, for example. In other cases, it can provide a collection-level record for related classes of data, such as scientific analyses, although an application profile may be needed to deal with the specific research questions of the sub-domain. The work of creating application profiles was delegated to special interest groups of sub-domain specialists. The activities and the discussions of the working groups have demonstrated that in many cases the AO-Cat is sufficient, without further modification and enhancement, to cover the needs of their sub-domains, even for item-level data aggregation. It is also notable that the addition of a large collection of resources to the Catalogue during the lifespan of ARIADNEplus has not required any change to the ontology. This indicates that the AO-Cat has reached a mature level of stability and is therefore ready for wider usage outside the boundaries of the ARIADNEplus project. Indeed, within the UK the AO-Cat has been adopted with little modification by [Unpath'd Waters](), a major project to characterise and aggregate digital data concerning the rich maritime heritage of the UK.

4. Data Mapping

Having designed the AO-Cat as a shared ontology, interoperability was dependent upon mapping the data models of heterogeneous partner datasets to the AO-Cat. For this the X3ML Toolkit developed by FORTH was used. The X3ML Toolkit is a set of small, open source, microservices that follow the [SYNERGY Reference Model]() of data provision and aggregation. It allowed us to define equivalences between the various attributes in partner-provided data with properties in the AO-Cat. However, to facilitate a useful cross-search it was also essential to achieve interoperability of the metadata used to describe the three main facets of 'Where', 'When' and 'What'. To support effective cross-search of metadata originating from many different countries requires some shared common understanding of the meaning of the metadata. The 'Where' and 'When' facets may be communicated using common data types and comparable values (e.g. spatial coordinates relative to a known location on the earth, date ranges relative to a known epoch). The 'What' (subject) aspect can be more difficult to define in a commonly understood and comparable format.

4.1. The Getty Art & Architecture Thesaurus (AAT)

Rather than attempt to map every subject vocabulary to every other subject vocabulary it was agreed to map each vocabulary to a common spine (Binding *et al*. [2015]()). The [Getty Art & Architecture Thesaurus]() (AAT) provides concepts and terms to describe cultural heritage concepts. The AAT is available as Linked Open Data, so each concept in the thesaurus has a unique identifier in the form of a URI (e.g. [http://vocab.getty.edu/aat/300211545]() is the identifier of the concept '*penannular brooches*'). ARIADNE adopted the AAT as a central hub for scalable interconnection of local subject vocabularies. Data providers were required to provide a set of mappings from their own local subject vocabulary (all terms used to describe subjects in their own metadata) to the AAT, and a vocabulary mapping tool was

developed to make this process easier (Figure 8). The aim of the subject mapping exercise was to improve recall and precision when later browsing and searching the data by subject. However, one of the issues faced was that for several countries no agreed subject vocabulary existed. In these cases, partners often undertook an extensive data enhancement exercise, adopting the AAT from the outset. For others this was not possible but it was decided to still include their data in aggregation, as long as it was still discoverable by the 'When' and 'Where' facets.

Figure 8: The Vocabulary Matching Tool, illustrating the ability to define Exact, Close, or Broad matches between the Getty AAT and target concepts in native vocabularies deployed by partners (courtesy of Ceri Binding, ARIADNE project)

4.2. PeriodO

PeriodO is a multilingual gazetteer of period definitions for describing historical and archaeological named periods. It comprises a series of 'authorities' (lists of named periods) where each individual period definition is associated with a particular geographical area and a start date and end date. Periods and authorities have 'permalink' identifiers (URIs) so they can be clearly and unambiguously referenced - e.g. the identifier http://n2t.net/ark:/99152/p0rrjd9gjx9 defines the period 'Moyen Âge' (Middle Ages) within the authority http://n2t.net/ark:/99152/p0rrjd9 'INRAP: Chronologie Generale. 2007' (Figure 9). All named periods in ARIADNE have been defined with reference to periods within PeriodO.

Figure 9: PeriodO - period definitions for France provided by Inrap

The purpose of using PeriodO in ARIADNE differs from the use of AAT data enrichment; rather than aligning data to common identifiers, PeriodO is utilised in a subsequent data enrichment stage to enhance records already indexed with named periods (such as 'Bronze Age', 'Iron Age' etc.) where the dates associated with these periods are not already made explicit in the input data - providing start/end dates to make these records comparable and searchable in the same way as other date information.

Again, there were some cases where period names were not available, although sometimes providers already had From and Until dates for their resources, so the PeriodO enhancement could be bypassed. In a few cases, however, no temporal data was available, but resources were included if they were discoverable via the other core facets.

4.3. Geographic Data

To deal with the spatial dimension, all spatial coordinates in ARIADNE are expressed using the World Geodetic System 1984 (WGS84). This is a standard geographic coordinate system, comprising a global horizontal and vertical datum and a coordinate system used (particularly by GPS systems) to express global positioning on the surface of the Earth. For ARIADNE purposes all spatial coordinates are expressed using WGS84. Coordinates in local datasets sometimes required normalisation/transformation prior to aggregation in order to improve opportunities for cross-searching the integrated datasets. WGS84 coordinates are expressed in degrees – the horizontal position is the longitude (having values between -180° and +180°, relative to a fixed datum of 0° at Greenwich, UK) and the vertical position is the latitude (having values between -90° and +90°, relative to a fixed datum of 0° at the equator). As an example, the global location of the Leaning

Tower of Pisa (Torre di Pisa) is given by WGS84 coordinates latitude 43.723056 (43° 43′ 23″ N), and longitude 10.396389 (10° 23′ 47″ E).

5. The Aggregation Workflow

Partners were offered two options for aggregation: the standard approach using a suite of tools for the semi-automated aggregation of large datasets, and a basic approach for the manual upload of small numbers of records. The majority of partners used the standard approach, but a small number use *FastCat*, a tool provided for upload of a few records. This also proved invaluable for the addition of bespoke Collection records for harvested resources.
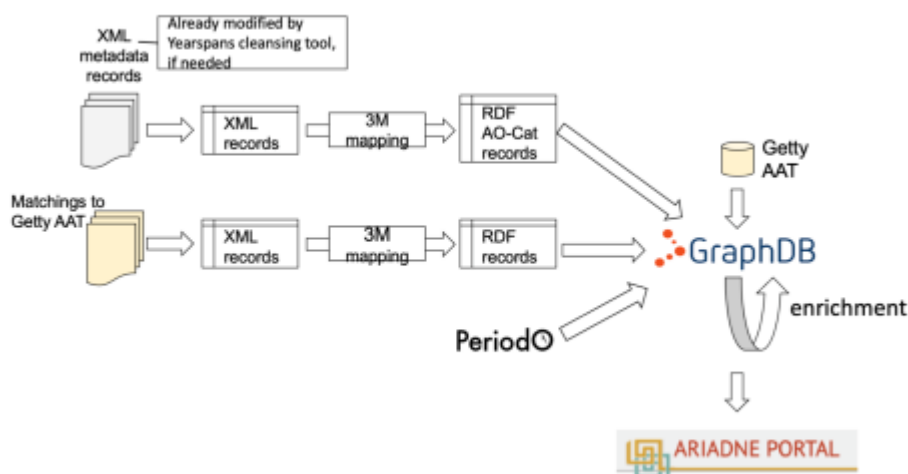
Figure 10: The aggregation workflow (courtesy of Alessia Bardi, ARIADNE project)

Partners following the standard approach had to:

1. Describe their data according to the AO-Cat using the X3ML tool, usually with one mapping per partner
2. Map subject terms to the Getty AAT using the Vocabulary Matching Tool
3. Define any period terms used so that they are uploaded to PeriodO.

Where temporal data needed cleaning to create consistent use of date ranges and periods, partners use an additional tool, *Yearspans*, to normalise date ranges (Binding and Tudhope 2023). They also had to ensure that their spatial data was compliant with WGS84.

Partners using *FastCat* instead manually entered their data records in the spreadsheet-like tool, where the column headings already corresponded to AO-Cat core mandatory fields, so that there could be a single X3ML mapping covering multiple partners.

Data providers supplied their data to ARIADNE either as XML or JSON data dumps or, where regular updates might be required then OAI-PMH feed was preferred. Data

were then converted into the AO Cat Ontology as RDF files and stored on the ARIADNEplus knowledge base (based on the GraphDB technology). The supplied data might already include mappings to Getty AAT or PeriodO, or these could be added to the knowledge base at this data aggregation stage (Figure 10). From here the ARIADNE data resources can be reused for multiple purposes e.g., in portals or in VREs, or by external consumers, avoiding the creation of a data silo. The knowledge base also facilitates linkage with external Linked Open Data sources, such as Wikidata (Figure 11). Data were first published on a staging portal, to allow a quality check, before release to the public portal.
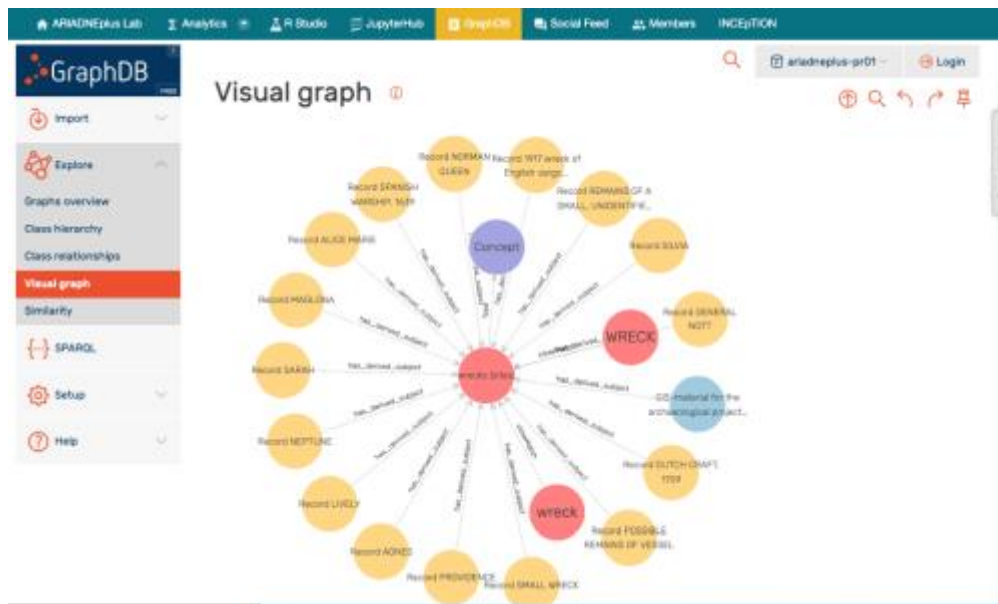
Figure 11: Visual graph showing part of the ARIADNE triple store in GraphDB

6. Using the ARIADNE Portal for Archaeological Research and Heritage Management

Part of the strength of the ARIADNE portal is provided by the indices that sit on top of the triple store and provide a rapid and effective means of filtering the Knowledge Base. These were initially implemented in *Elastic Search* but were migrated to *Open Search* to maintain the emphasis on open-source software; they offer a powerful faceted browse facility. Reflecting its pedigree in Dublin Core, the AO-Cat allows for the definition of various agents associated with a data resource: Creator, Contributor, Owner, Publisher, Scientific Responsible. These can be individuals or organisations. In many cases the majority of the agents declared by data providers were the same entity but the range of possible agents allowed for the attribution of a wide range of roles. As Publisher and Contributor each had filters in the portal interface these were the most important, and one of the challenges of the aggregation process was to ensure that the categories were applied consistently. The AO-Cat scope notes therefore defined Publisher as the agent (usually the data provider or partner)

responsible for publishing the metadata online, while the Contributor was generally the same as the Owner or Creator (Figure 12).

Publisher	Hits
Archaeology Data Service	1080189
British Museum	945228
Historic Environment Scotland	334636
Museum of Cultural History, University of Oslo	295674
Archaeological Information System of the Czech Republic (AIS CR)	229999
National Monuments Service, Department of Housing, Local Government and Heritage, Republic of Ireland	161383
Nara National Research Institute for Cultural Properties	141159
Data Archiving and Networked Services (DANS)	99291
Hungarian National Museum (HNM)	63019
CEIPAC - Universitat de Barcelona	57089
Historic England	50680
Institut national de recherches archéologiques préventives (Inrap)	37616
Swedish Rock Art Research Archives	25223
National Heritage Institute - Romania	23888
THANADOS	23580
Israel Antiquities Authority	23240
Aarhus University	16826
International Association for Classical Archaeology	14279
The Institute of Archaeology, Iceland	10344

Get 20 more results..

Figure 12: The Publisher filter in the ARIADNE portal. This provides a rapid identification of the number of records from each provider, and in many cases also acts as a country search.

Of equal importance is the Resource Type filter. This is effectively a specialisation of the What or Subject category, drawn from a short list of 13 pre-determined categories. Although the terms can be mapped to the broader AAT, they provide a pragmatic set of resource types that were defined 'bottom-up' according to the most common types of data provided by the partners. Again, consistency of application was important but the filter provides an effective means of defining the level of granularity required of a search (Figure 13). Records within a Resource type tend to be more homogeneous than between Resource Types, ensuring a higher degree of interoperability within categories. If more providers were to join the ARIADNE RI the categories are capable of extension, although this could have implications for the need to re-categorise existing collections. They are also dependent upon the level of information provided. Therefore, it was relatively easy to distinguish between the 470,677 coins and the 474,552 other artefacts provided by the British Museum Portable Antiquities Scheme (PAS), and some data providers, such as the RGK, only provided information about coins, but for the Danish DIME database the coin records were embedded within the broader dataset.

Resource type	Hits
Site/monument	1389759
Artefact	855253
Coin	471403
Fieldwork	421697
Fieldwork report	293825
Fieldwork archive	113941
Maritime	86509
Inscription	83365
Rock art	26805
Burial	11573
Dating	9015
Scientific analysis	7766
Building survey	1407

Figure 13: The Resource Type Filter in the ARIADNE portal allows users to extract a subset of interoperable data

The collaboration with THANADOS, an Austrian anthropological and archaeological database of cemeteries and graves provides a particularly good example of the value of resource types. The THANADOS data model provides information at four levels: cemetery, grave, burial and artefact, and this is reflected in the ARIADNE application profile for mortuary data (Aspöck *et al.* 2023). The AO-Cat mapping enables the use of ARIADNE resource types 'Site/monument', 'Burial' and 'Artefact' to mirror the relational structure of the THANADOS database, allowing users to filter according to the level of granularity of interest to them. In turn, the Mortuary application profile can be used to encode the details of each of the unique aspects of this archaeological activity.

For more detailed subject-based searching the use of the AAT as a common spine vocabulary has provided very effective, and the Vocabulary Mapping Tool has allowed partners to undertake the mapping relatively easily. Although the Getty thesaurus does not define equivalent terms in all languages it does allow a level of multilingual searching. Having specified an AAT subject term search, for example, the users can enter 'Swords' (em), 'Zwaarden' (nl), 'Épée' (fr), 'Spada' (it) or 'Espadas' (es) and achieve the same results (Figure 14). Furthermore, so long as the provider has mapped their own native language term for the equivalent to 'Swords' to the Getty AAT concept then their records will also be retrieved. The search also uses its dictionary to provide an auto-complete function as a further aid to the researcher.

Figure 14: The ARIADNEplus portal landing page, showing the auto-complete and multilingual features provided by the application of the Getty AAT

The power of the hierarchical Getty AAT is also demonstrated by the ability of the user to find objects defined via narrower terms, even if the broader term has not been included in the metadata (Figure 15). Therefore, a search for 'Swords' also retrieves 'Cutlasses', 'Foils', 'Rapiers' and any of the 22 narrower categories. In some cases, the Getty AAT did not go down to the level of granularity that researchers might want, for example in detailed brooch or coin types, although where there is domain backing and authority a future case could be made to extend the thesaurus for specialist usage.

Figure 15: A search on 'weapons' without narrower terms (top) yields only 753 resources, whereas when narrower terms are included (bottom) 24,834 resources are retrieved

One of the most popular and most used features of the portal interface has proved to be the map-based search (Figure 16). The heat map provides a rapid visualisation of the density of records, but resolves into individual clickable points and polygons as the researcher uses the zoom function. Additional tools include the option to use different map bases and to search via a bounding box or a user-defined polygon. The menu provides the same filters as available on the main results page, allowing users to select specific records of interest.

Figure 16: The map browsing function uses a heat map to visualise the density of records

Support for polygons was incorporated within the AO-Cat, as these can be important to display the full extent of archaeological sites or fieldwork areas. The polygons are best viewed on the maps included within landing pages of individual records (Figure 17), which also provide the option to view nearby resources. The metadata displayed on individual resource landing pages also generally provide links to the landing page for the primary resource. This is of considerable importance as the links may give access to much richer and more detailed information, including downloadable text reports and, in some cases, complete online digital archives. Ideally the link should be via permanent identifiers, such as Digital Object Identifiers (DOIs), contributing to the portal providing FAIR data access. This also provides an important 'shop window' function for the data providers, drawing researchers to their own web sites, about which they might have been previously unaware. In a few cases, however, the data have been published online for the first time via the ARIADNE portal, and there is no additional information to link to.

Figure 17: Landing page for an individual data resource, illustrating the ability to display the geospatial extent of the resource as a polygon

The temporal browse function provides an equivalent functionality for the 'When' facet. It allows rapid visualisation of the temporal range of the resources selected, providing a valuable research tool in its own right, but the use of PeriodO to define ranges in absolute years also allows the user to select a specific time span of interest and view the results (Figure 18). A second option, added during the project, facilitates a filter by time periods, according to the temporal regions selected. For example, this allows the user to select sites classified according to the cultural label 'Iron Age' across different temporal regions, irrespective of the From and Until dates of the Iron Age in those regions

Figure 18: The temporal browse facility in the portal

The AO-Cat ontology proved to be adequate for the needs of most data providers during the lifetime of the project and extensions were only required in two areas. The first came from a desire to include images among the metadata displayed on the resource landing pages, and thumbnail images within the results list. This arose particularly from artefact researchers wanting to be able to compare individual objects (Figure 19). Clearly this could also be enhanced using the International Image Interoperability Framework (IIIF) and appropriate viewers in a future project.

Figure 19: The landing page for a coin held by the Bank of Cyprus Cultural Foundation showing the value of including images among the metadata harvested by the portal

The second need for an extension to the AO-Cat arose because some data providers asked that information about the spatial region covered by a data resource be accompanied by a specification of the precision of the information. This was for two main reasons. Firstly, in some cases the data provider only held data for site location at an imprecise level, so it was necessary to make clear, when plotting locations on a map, that the point might represent the site location to the nearest 10km, for example. Once this was encoded within the AO-Cat and the data stored in the triple store, the portal dealt with it using a different symbology. Secondly, there were cases where the data provider had sensitive spatial data that they did not wish to make publicly available, maybe to protect the precise location of a site. In these cases, spatial data were supplied at the resolution that they are willing to make available, and the location was represented as a bounding box, which included the actual find spot at some random point. For example, spatial data for the UK PAS database was supplied to either a 1km or 10km bounding box, according to the sensitivity of the finds spot (Figure 20).
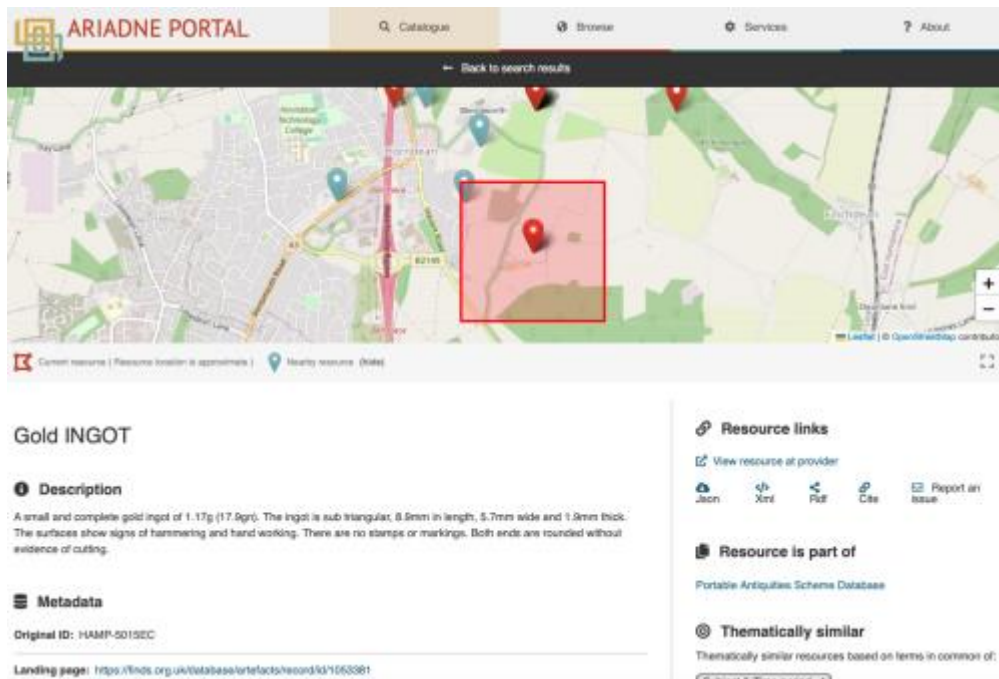
Figure 20: The location of the find spot of a gold ingot displayed to a 1km resolution, such that the actual find spot may be anywhere within the red bounding box

Finally, the ARIADNE Knowledge Base and portal architecture means that it can easily be adapted to provide a broad range of more specialist interfaces. For example, as part of the Towards a National Collection Unpath'd Waters project the ADS was able to develop a portal that provides a distributed cross-search of the many agencies that hold data for the maritime heritage of UK coastal waters by implementing an instance of the ARIADNE portal (Figure 21). In this case the search was 'hardwired' so that it limited the search to those resources for which the Resource Type was 'Maritime', and so that only records from UK data providers were retrieved. As the metadata source is the ARIADNE RI triple store the metadata is only held once, and resources added for the Unpath'd Waters project simultaneously appear in the ARIADNE portal, thereby providing a European view of the UK dataset. The approach could be extended in any direction, for example, providing national views for agencies that do not wish (or have the capacity) to develop their own search interface, and avoiding duplication, or for thematic data views, at an international level.
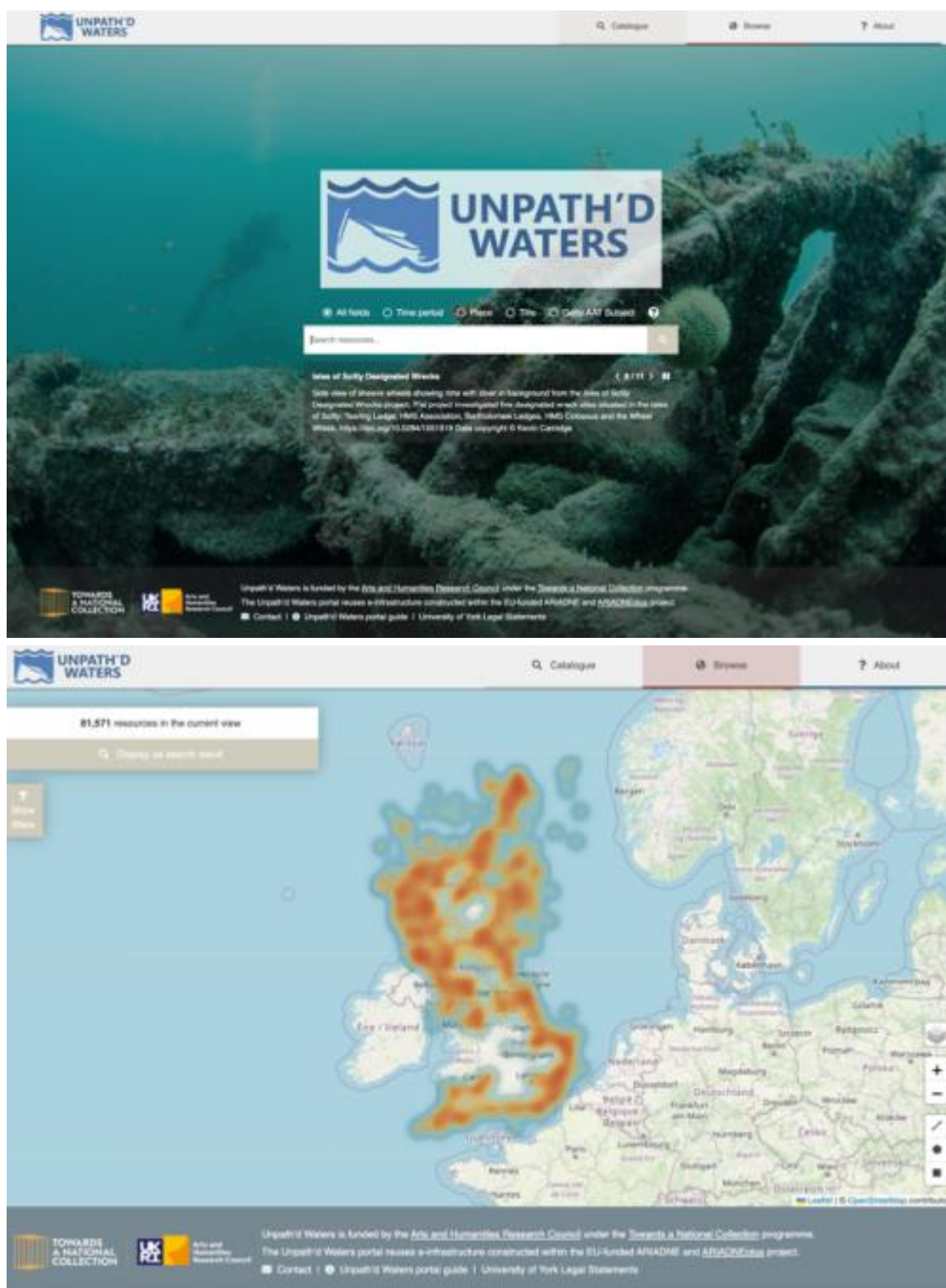
Figure 21: The Unpath'd Waters Portal uses the ARIADNE Knowledge Base and Portal interface to provide a hard-wired UK maritime heritage search

7. Conclusion

In conclusion, the ARIADNE portal provides access to the largest international archaeological dataset available online, and there is nothing comparable in other Arts and Humanities disciplines. It is essential to note, however, that the role of the ARIADNE aggregation service is not to duplicate services already available from data providers' own web sites, but rather to provide interoperability and resource discovery across multiple resources. Nonetheless in some cases the portal provides a powerful research tool in its own right (Richards 2023). The service was built upon

the foundations laid by a number of preceding projects but further lessons have been learnt. Firstly, to develop a useful cross-search it is clear that consistent use of an ontology is essential. The CIDOC CRM was broad enough to cover all the sub-domains involved, but it was important to map to it consistently, and the development of community-led application profiles was essential in this regard. Secondly, to achieve interoperability vocabulary control is also essential, and we were fortunate in being able to build on decades of investment in data standards and thesauri in archaeology. The Getty AAT provided a powerful common spine but mappings again require oversight to ensure consistency in partner mappings. Thirdly, it is essential to realise that, when dealing with such heterogeneous data it is impossible to make everything interoperable, and it is necessary to consider the role of aggregation and how the data will be used by researchers. In ARIADNE, building on previous projects, What, When, Where proved to be extremely effective, although in hindsight Who might also have proved useful, particularly if the portal were to be extended to include more authored reports and books. However, for this we currently lack authority lists to allow us to define individuals. Greater adoption of ORCID identifiers might provide a way forward for living people, but we need to examine solutions and identifiers offered by Wikidata to deal with people in the past.

Overall, our work on data aggregation has underlined the importance of interdisciplinary teams in this endeavour and a requirement for archaeologists to work closely with information scientists. It is necessary to take key decisions about the granularity and scale of data aggregation, and what is useful to aggregate at item level, or where it is only sensible to aggregate at collection level. In this regard, our approach needs to be driven by research, and the questions of researchers and domain specialists.

Acknowledgements

Bibliography

Aloia, N., Binding, C., Cuy, S., Doerr, M., Fanini, B., Felicetti, A., Fihn, J., Gavrilis, D., Geser, G., Hollander, H., Meghini, C., Niccolucci, F., Nurra, F., Papatheodorou, C., Richards, J., Ronzino, P., Scopigno, R., Theodoridou, M., Tudhope, D., Vlachidis, A. and Wright, H. 2017 'Enabling European archaeological research: the ARIADNE E-infrastructure', *Internet Archaeology* **43**. https://doi.org/10.11141/ia.43.11

Aspöck, E., Eichert, S., Theodoridou, M., Richards, N. and Felicetti, A. 2023 'Integrating Data on Early Medieval Graves: Mapping the THANADOS database to the ARIADNE infrastructure with the Mortuary Data Application Profile', *Internet Archaeology* **64**. https://doi.org/10.11141/ia.64.11

Austin, A., Richards, J.D., Pinto, F. and Ryan, N. 2002 'Joined up writing: an Internet portal for research into the historic environment' in G. Burenhult (ed) *Archaeological Informatics: Pushing the Envelope CAA2001*, British Archaeol. Rep. International Series **1016**. 243-51.

Benardou, A., Champion, E., Dallas, C. and Hughes, L. 2017 'Introduction: a critique of digital practices and research infrastructures' in A. Benardou, E. Champion, C. Dallas and L. Hughes (eds) *Cultural Heritage Infrastructures in Digital Humanities*, London: Routledge. 1-14.

Bibby, D. 2021 'Digital archaeological archiving in Baden-Württemberg, Germany: an evolving system', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.3

Binding, C., Charno, M., Jeffrey, S., May, K. and Tudhope, D. 2015 'Template based semantic integration: from legacy archaeological datasets to linked data', *International Journal on Semantic Web and Information Systems* **11**(1), 1-29. https://doi.org/10.4018/IJSWIS.2015010101

Binding, C. and Tudhope, D. 2023 'Automatic normalization of temporal expressions', *Journal of Computer Applications in Archaeology* **6**, 24-39. https://doi.org/10.5334/jcaa.105

Dam, C. and Hansen, H.J. 2005 'The European digital resource in archaeology: sites and monuments data as a common European web resource', *Internet Archaeology* **18**. https://doi.org/10.11141/ia.18.4

Felicetti, A., Gerth, P., Meghini, C. and Theodoridou, M. 2015 'Integrating heterogeneous coin datasets in the context of archaeological research' in Ronzino, P. (ed) *EMF-CRM@TPDL. Extending, Mapping and Focusing the CRM*. 13-27. https://ceur-ws.org/Vol-1656/paper2.pdf

Felicetti, A., Meghini, C., Richards, J.D. and Theodoridou, M. 2023 *The AO-Cat Ontology*. https://doi.org/10.5281/zenodo.7818375

Geser, G., Richards, J.D., Massara, F. and Wright, H. 2022 'Data management policies and practices of digital archaeological repositories', *Internet Archaeology* **59**. https://doi.org/10.11141/ia.59.2

Hansen, H.J. 1992 'European archaeological databases: problems and prospects' in J. Andresen, T. Madsen and I. Scollar (eds) *Computing the Past. Computer Applications and Quantitative Methods in Archaeology*, Aarhus: Aarhus University Press. 229-37.

Hansen, H.J. and Fernie, K. 2010 'CARARE: connecting archaeology and architecture in Europeana', *International Euro-Mediterranean Conference (2010)*. https://doi.org/10.1007/978-3-642-16873-4_36

Huggett, J. 2015 'Digital haystacks: open data and the transformation of archaeological knowledge' in A.T. Wilson and B. Edwards (eds) *Open Source Archaeology: Ethics and Practice*, Warsaw, Poland: De Gruyter Open. 6-29. https://doi.org/10.1515/9783110440171-003.

Isaksen, L., Simon, R., Barker, E. and de Soto, C. 2014 'Pelagios and the emerging graph of ancient world data', *Proc. ACM Conference on Web Science (WebSci'14)*, ACM. 197-201. https://doi.org/10.1145/2615569.2615693

Jeffrey, S., Xia, L., Richards, J., Bateman, J., Kintigh, K., Pierce-McManamon, F. and Brin, A. 2012 'The transatlantic archaeology gateway: bridging the digital ocean' in M. Zhou, I. Romanowska, Z. Wu, P. Xu, and P. Verhagen (eds) *CAA2011- Revive the Past: Proceedings of the 39th Conference in Computer Applications and Quantitative Methods in Archaeology, Beijing, China, 12-16 April 2011*, Pallas Publications. 198-208

Kansa, E. 2014 'Open context and linked data' in T. Elliott, S. Heath and J. Muccigrosso (eds) *Current Practice in Linked Open Data for the Ancient World*, ISAW Papers **7**(10) http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/kansa/

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

May, K., Binding, C. and Tudhope, D. 2015 'Barriers and opportunities for linked open data use in archaeology and cultural heritage', *Archäologische Informationen* **38**, 173-84.

Meghini, C., Scopigno, R., Richards, J.D. *et al.* 2017 'ARIADNE: a research infrastructure for archaeology', *Journal of Computing and Cultural Heritage* **10**(3), 1-27. https://doi.org/10.1145/3064527

Michard, A., Christophides, V., Scholl, M., Stapleton, M., Sutcliffe, D. and Vercoustre, A-M. 1998 'The Aquarelle resource discovery system', *Computer Networks and ISDN Systems* **30**(13), 1185-1200.

Niccolucci, F. and Richards, J. 2019 'ARIADNE and ARIADNEplus' in J. Richards and F. Niccolucci (eds) *The ARIADNE Impact*, Budapest: Archaeolingua Foundation. 7-25. https://zenodo.org/record/4319058

Nicholson, C., Fernandez, R. and Irwin, J. 2021 'Digital archaeological data in the Wild West: the challenge of practising responsible digital data archiving and access in the United States', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.22

Richards, J.D. 2017 'Twenty years preserving data: a view from the UK', *Advances in Archaeological Practice* **5**(3), 227-37. https://doi.org/10.1017/aap.2017.11

Richards, J.D. 2023 'Getting it together: combining information about archaeological sites and artefacts in ARIADNE', *Internet Archaeology* **64**. https://doi.org/10.11141/ia.64.14

Richards, J. and Niccolucci, F. (eds) 2019 *The ARIADNE Impact*, Budapest: Archaeolingua Foundation. https://zenodo.org/record/4319058

Richards, J.D., Jakobsson, U., Novák, D., Štular, B. and Wright, H. 2021 'Digital archiving in archaeology: the state of the art. Introduction', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.23

Shaw, R., Rabinowitz, A., Golden, P. and Kansa, E. 2016 'A sharing-oriented design strategy for networked knowledge organization systems', *International Journal on Digital Libraries* **17**, 49-61. https://doi.org/10.1007/s00799-015-0164-0

Vassallo, V., Theodoridou, M., Felicetti, A. and Avgousti, A. 2023 'Tools and Ontologies for the Aggregation and Management of Cypriot Archaeological Datasets', I*nternet Archaeology* **64**. https://doi.org/10.11141/ia.64.10

Wright, H. and Richards, J.D. 2018 'Reflections on collaborative archaeology and large-scale online research infrastructures', *Journal of Field Archaeology* **43**, S60-S67, https://doi.org/10.1080/00934690.2018.1511960