D12.4 – Final report on data integration

Version 1.0 (*final*)

2022-10-31

Grant Agreement number:	**823914**
Project acronym:	**ARIADNEplus**
Project title:	**Advanced Research Infrastructure for Archaeological Dataset Networking in Europe - plus**
Funding Scheme:	**H2020-INFRAIA-2018-1**
Project co-ordinator name, Title and Organisation:	**Prof. Franco Niccolucci, PIN Scrl - Polo Universitario "Città di Prato"**
Tel:	**+39 0574 602578**
E-mail:	**franco.niccolucci@pin.unifi.it**
Project website address:	**www.ariadne-infrastructure.eu**

The research leading to these results has received funding from the European Community's Horizon 2020 Programme (H2020-INFRAIA-2018-1) under grant agreement n° 823914.

Author: **Alessia Bardi**

CNR-ISTI

Johan Fihn Marberg

SND

Maria Theodoridou

FORTH

Contributing partners: **Miriam Baglioni, CNR-ISTI**

Vittore Casarosa, CNR-ISTI

Pablo Millet, SND

Enrico Ottonello, CNR-ISTI

Quality control: **Stephanie Williams, PIN**

Document History

- 22.09.2022 – Draft Version 0.1
- 08.10.2022 - Draft Version 0.2
- 14.10.2022 - Draft Version 0.3
- 31.10.2022 - Version 1.0, final

Table of Contents

1 Executive Summary

This deliverable describes the final results of Tasks 12.1, 12.2, 12.3 and 12.4 of the ARIADNEplus project. The tasks developed, delivered, and maintained the components of the ARIADNEplus infrastructure that support the integration and interoperability of the data provided by the members of the consortium and associate partners.

The catalogue data integrated by the ADI (the Aggregative Data Infrastructure developed in T12.2) are transformed into RDF records compliant with the CRM-based ARIADNEplus Ontology (AO-Cat) and are made publicly available in the ARIADNEplus Knowledge Base (T12.1), where they are enriched and linked to the Getty AAT thesaurus and the PeriodO gazetteer of time periods. The knowledge base exposes a SPARQL endpoint and its content is indexed on an OpenSearch index that serves the ARIADNEplus portal (T12.3). The domains of mortuary archaeology, numismatics and epigraphies were chosen for deeper integration of item level data (T12.4) compliant with specific Application Profiles developed in WP4. A searching interface, built on the ResearchSpace platform provides support for research questions that require information richer than what is available in AO-Cat.

The design, development and deployment activities have been guided by the requirements of all the members of the consortium, especially those involved in WP4, WP5 and the working group on portal requirements.

The ARIADNEplus ADI features the following main components for data collection, transformation, and harmonisation: (i) the 3M Editor developed and maintained by FORTH, operated by CNR; (ii) the Vocabulary Matching Tool developed and maintained by USW, operated by CNR; and (iii) the ARIADNEplus aggregator developed, maintained, and operated by CNR.

The knowledge base is implemented with GraphDB (free edition) and one Springboot application that acts as mediator for the interactions among the aggregator, GraphDB, and OpenSearch. As of October 2022, the KB contains 175M RDF triples describing 3M resources from 60K collections and 25 providers. Numbers are expected to increase by December 2022 with the aggregation of additional resources as planned in WP5.

OpenSearch provides a full-text index of the content available in the knowledge base, to be used by the ARIADNEplus portal. The ARIADNEplus portal is developed using PHP, Vue.js, Javascript, Vuex, Tailwind, and Font Awesome. The portal provides standard free-text and faceted search options, but also advanced features based on the concepts of temporal, spatial, and topical coverage.

In order to test the new functionality of the portal and enable data curators to check the quality of data before it is made available on the public portal, WP12 set up a staging environment accessible only to members of the consortium, where the collected data is aggregated, added to a staging knowledge base and indexed on a staging portal. Upon confirmation of data experts, data is then pushed to the production environment.

2 Introduction and Objectives

This deliverable describes the final results of the four tasks of WP12:

- T12.1 Implementing the ARIADNEplus AC (JRA1.1) - now called ARIADNEplus Knowledge Base (KB)
- T12.2 Implementing the ARIADNEplus Aggregative Data Infrastructure (ADI) (JRA1.2)
- T12.3 Implementing the ARIADNEplus Portal (JRA1.3)
- T12.4 Item-level data integration (JRA1.4)

The objectives are to develop, deliver and maintain the components of the ARIADNEplus infrastructure that support the integration and interoperability of the data provided by the members of the consortium and the associate partners. Figure 1 shows the main components and interactions implemented for data integration and interoperability. The ADI collects data from partners and transforms them into RDF records compliant with the AO-Cat model. Those records are stored in the ARIADNEPlus KB, where they are enriched and linked to Getty AAT and PeriodO. The KB exposes a SPARQL endpoint and its content is indexed on an OpenSearch index that serves the ARIADNEplus portal. The collection, transformation, storage, enrichment, and indexing phases are orchestrated by the ADI with the supervision of an aggregation manager.

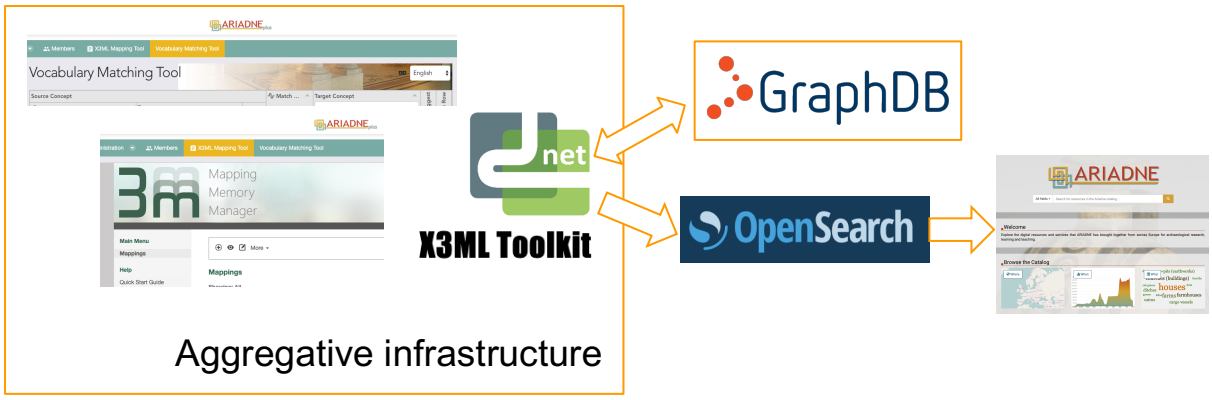

Figure 1 Services and tools for data integration and interoperability in the ARIADNEplus infrastructure

The aggregative infrastructure (T12.2) includes services and tools required to perform data collection, transformation, and enrichment according to domain-specific vocabularies and ontologies:

- 3M Editor: definition of the mappings from local metadata format to AO-Cat format.
- Vocabulary Matching Tool: definition of mappings from local subject terms to terms of Getty AAT.
- D-NET: the ARIADNEplus aggregator is based on the D-NET software toolkit. It collects the providers' XML records and integrates the X3ML toolkit for the execution of 3M mappings. It is configured to implement the aggregation workflows defined in collaboration with WP5.

The ARIADNEplus Knowledge Base (KB) (T12.1) (previously named ARIADNEplus AC) is a knowledge graph developed according to the Resource Description Framework (RDF), supporting the Semantic

Web principles (Linked Data). It comes with a SPARQL access point, providing triples of the knowledge graph, and reasoning capabilities that can be exploited for the implementation of advanced knowledge discovery services.

The portal (T12.3) is the main entry point for humans willing to search, browse and access the aggregated resources.

Finally, the deeper integration of item level data (item-level integration) is investigated in task 12.4 for the domains of mortuary archaeology, numismatics and epigraphies. Item-level integration is achieved using specific Application Profiles developed in WP4 for each of the selected domains, in addition to the AO-Cat. A searching interface, built on the ResearchSpace platform provides support for research questions that require information richer than what is available in AO-Cat.

All services are operational in production (TRL9) from December 2020. Their operation and performance is monitored with the monitoring infrastructure hosted at CNR based on Prometheus and Grafana.

Final Results

T12.1 ARIADNEplus data and knowledge Cloud (AC) - aka ARIADNEPlus Knowledge Base (KB)

The ARIADNEplus KB has been designed, implemented and deployed in two instances: one staging instance for data curation and one production instance available to the public. Each instance includes a GraphDB (free edition, v9.8.0) server and a "publisher" component: a SpringBoot application in charge of the communication with the ADI and OpenSearch. All services are monitored via Prometheus and Grafana.

The KB is accessible via the ARIADNEplus Lab Virtual Research Environment (VRE)[1]. Programmatic access is also available via a SPARQL endpoint. More information can be found in the documentation, available in Jupyter Notebook, PDF, and MarkDown formats[2], slides[3] and sample Java code[4].

As of October 2022, the KB contains 175M RDF triples describing 3M resources from 60K collections and 25 providers. Numbers are expected to increase by December 2022 with the aggregation of additional resources, as planned in WP5.

T12.2 Implementing the ARIADNEplus Aggregative Data Infrastructure (ADI) (JRA1.2)

The ADI of ARIADNEplus has been designed, implemented and it is fully operational. Its main components are:

- The D-NET framework toolkit, configured to implement the aggregation and enrichment workflows devised in collaboration with WP5. As of October 2022, the aggregator processed

[1] https://ariadne.d4science.org/group/ariadneplus_lab
[2] https://data.d4science.net/EvVX
[3] https://docs.google.com/presentation/d/1B2eFNRCi_jGVSXmwIzqrZRrwuZUvBbKq_ynNL1tZpF8/edit?usp=sharing
[4] https://github.com/ARIADNE-Infrastructure/sample-code

3.6M XML records from 33 providers. The performance of the aggregator is continuously monitored with Prometheus and Grafana.
Source code: https://code-repo.d4science.org/D-Net/AriadnePlus

- 3M Editor, updated to allow the generation of RDF records compliant with the AO-Cat model.
  Source code: https://github.com/isl/Mapping-Memory-Manager
- Vocabulary Matching Tool, deployed on the D4Science infrastructure for the generation of mappings from local subject terms to terms from the Getty AAT thesaurus.
  Source code: https://github.com/cbinding/VocabularyMatchingTool

T12.3 ARIADNEplus portal

The ARIADNEplus portal[5] has been designed, implemented and deployed in two instances: one staging portal for checking data quality and testing new functionality, and one production portal publicly available and linked from the project website.

The previously available Elasticsearch server for indexing and search has been migrated to OpenSearch.

T12.4 Item-level integration

Item-level integration unfold in three domains that were mature and suitable for deeper integration:

- Mortuary archaeology: integration based on Mortuary Application Profile v1.1.
- Numismatics: integration based on CIDOC CRM core and nomisma.org.
- Epigraphies: integration based on Inscriptions, marks and graffiti Application Profile (CRMtex v1.1)

The task proved the feasibility of integration in the above three application domains and the need for specialized VREs and vocabularies to support querying and access. An experimental searching interface is built on the ResearchSpace platform for the exploration of the Coins' and Inscriptions' collections.

[5] https://portal.ariadne-infrastructure.eu/

3 Activities

3.1 T12.1 – JRA 1.1 Implementing the ARIADNEplus Knowledge Base

The ARIADNEplus data and knowledge Cloud (ARIADNEplus AC) is a knowledge base (KB) where datasets of the Archaeological Research Communities are integrated, structured according to the CIDOC CRM-based ARIADNEplus Ontology (AO-Cat), and interlinked with standard ontologies and gazetteers of the archaeological domain, such as PeriodO for historical periods and Getty AAT for arts and archaeological subject terms classification.

The objectives of task 12.1 were to design, develop, deliver and maintain the KB using standards and state-of-the-art web and Semantic Web languages and protocols (e.g. REST, RDF, XML, JSON, SPARQL) to maximize interoperability and foster re-usability.

3.1.1 Work performed

During the first part of the project, GraphDB was selected as the technology for the implementation of the ARIADNEplus Knowledge Base and two instances were delivered to the project: a "staging" instance (available since M15, March 2020) accessible only to consortium members to evaluate the output of the aggregation, and a "public" instance (available since M24, December 2020) providing the openly accessible endpoint to the KB (see Figure 2). Details on how the data is organised in GraphDB to support continuous aggregation and update of the KB are available in the previous D12.2 mid-term report on data integration [1].

A publisher component was designed and developed for the communication between GraphDB, the ARIADNEplus aggregator and the OpenSearch index server. One instance was deployed in the staging environment and one in the production environment.

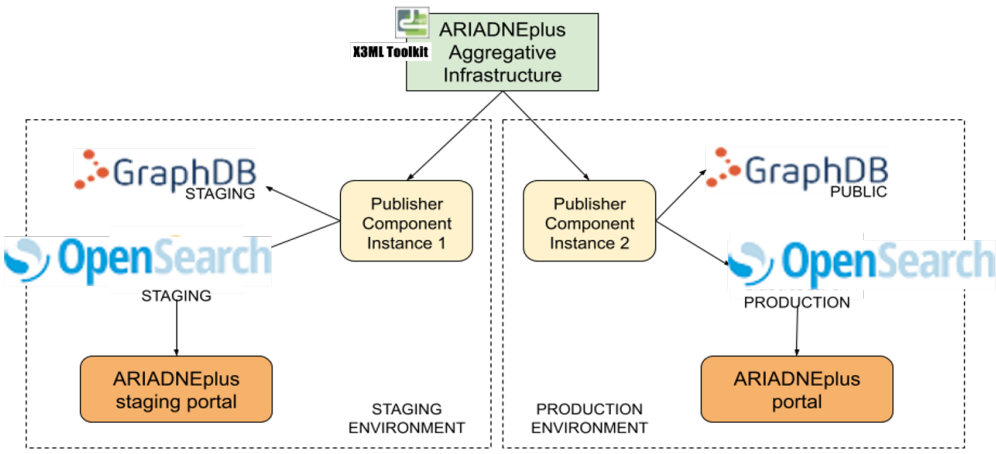

Figure 2 Staging and production environments

During the second period of the project, the task monitored the performance of the GraphDB servers and the publisher components, increasing the hardware resources and fine tuning their configuration when needed. GraphDB was updated from version 9.0 to 9.8.0. The monitoring is performed continuously by the monitoring infrastructure deployed at CNR's premises, based on Prometheus and Grafana.

The task also focused on activities to increase the level of FAIRness of the KB and its content:

Documentation & support

The task produced the following documentation and support material for the usage of the knowledge base:

- User guide for the ARIADNEplus Knowledge Base (GraphDB): an "actionable" user guide on how to use the GraphDB Web GUI to query the SPARQL endpoint programmatically. It is available in Jupyter Notebook, PDF, and MarkDown formats[6].
- Sample SPARQL queries and Java code: available in the ARIADNEplus GitHub repository[7]
- Slides for demo sessions of the ARIADNEplus KB[8].
- Additional material to support the hackathon at the Linked Pasts Symposium is under preparation and will be available in December 2022.

Accessibility

The task also ensured that the URIs associated with the resources aggregated by ARIADNE are resolvable and accessible, as suggested by the best practices for Linked Data. The objective was achieved by defining and implementing a specific policy for URI resolution based on content negotiation.

The Knowledge Base has been integrated within the ARIADNEplus Lab Virtual Research Environment[9]. The integration brought two main benefits:

1. In a research environment, users of the KB can work with several tools and documentation to carry out their research activities and share their results with other colleagues in a controlled way (for details on the benefits of a VRE please refer to D13.2 [2]).
2. ARIADNEplus can track the accesses to the GUI and the SPARQL endpoint of the KB and, therefore, monitor the usage trends of the KB.

Data enrichment and alignment with standards, ontologies, policies

The "provenance" graph, automatically created and updated by the ARIADNEplus aggregator to track which datasets come from which providers, initially used ad-hoc predicates. This task analysed two standard ontologies for provenance information, namely PAV [3,4] and Prov-O[5], so as to align the

[6] https://data.d4science.net/EvVX
[7] https://github.com/ARIADNE-Infrastructure/sample-code
[8] https://docs.google.com/presentation/d/1B2eFNRCi_jGVSXmwIzqrZRrwuZUvBbKq_ynNL1tZpF8/edit?usp=sharing
[9] https://ariadne.d4science.org/group/ariadneplus_lab

ad-hoc predicates to a standard and increase interoperability. Based on the features that the ARIADNEplus aggregator requires to model the provenance, the PAV ontology was considered to better fit our use case. The following triples have therefore been added to GraphDB ontology to align the provenance information to PAV:

```
<http://www.d-net.research-infrastructures.eu/provenance/isApiOf>
<http://www.w3.org/2000/01/rdf-schema#subPropertyOf>
<http://purl.org/pav/importedFrom> .

<http://www.d-net.research-infrastructures.eu/provenance/insertedInDate>
<http://www.w3.org/2000/01/rdf-schema#subPropertyOf>
<http://purl.org/pav/createdOn> .
```

Another relevant alignment to a standard has been done on the fields *aocat:from* and *aocat:until*, which contain temporal information. In fact, some values were fed into the KB as literal strings or non-valid years. The invalid values have been corrected so that SPARQL queries on temporal information can now work properly and return a complete set of results.

Finally, to avoid possible problems with privacy, any email address contained in the collected data was discarded and removed from the KB.

Findability via the ARIADNEplus portal

In order to support the implementation of additional features of the ARIADNEplus portal and improve the findability of the resources, the Publisher component was updated to feed the index server with more information from the KB, especially related to spatial and temporal information.

3.1.2 Deviation from work plan

No deviation from the work plan.

3.1.3 Plans for the next period

Additional material to support the hackathon at the Linked Pasts Symposium is under preparation and will be available in December 2022.

3.2 T12.2 – JRA1.2 Implementing the ARIADNEplus Aggregative Data Infrastructure

The Aggregative Data Infrastructure (ADI) includes services and tools required to perform data collection, transformation and harmonisation according to domain-specific vocabularies and ontologies.

- 3M Editor: definition of the mappings from local metadata format to the format of AO-Cat. Source code: https://github.com/isl/Mapping-Memory-Manager
- Vocabulary Matching Tool: definition of mappings from local subject terms to the terms of the Getty AAT thesaurus. Source code: https://github.com/cbinding/VocabularyMatchingTool
- D-Net: the ARIADNEplus aggregator is based on the D-Net software toolkit [6]. It collects the providers' XML records and integrates the X3ML toolkit for the execution of 3M mappings. It is configured to implement the aggregation workflows defined in collaboration with WP5. The aggregator is continuously monitored by the monitoring infrastructure deployed at CNR's premises, based on Prometheus and Grafana. Source code: https://code-repo.d4science.org/D-Net/AriadnePlus

The 3M Editor and the Vocabulary Matching Tool are available via the ARIADNEplus Mapping Virtual Research Environment[10], where content providers can also find disk space to upload their metadata in case they cannot expose their records via an application programming interface.

3.2.1 Work performed

During the second part of the project, T12.2 focused on the maintenance and operation of the aggregative infrastructure to serve the aggregation tasks carried out in the context of WP5.

The structure of the workflows has not changed with respect to what was reported in D12.2 [1].

All aggregation workflows have been updated and now use an updated URI generation policy[11].

Dedicated code to collect data from the THANADOS API was implemented and deployed. Based on the list of archaeological sites updated/added to the THANADOS website from a given date[12], the aggregator performs a request (one per site) to get the metadata description. The source code of the collector module is available at: https://code-repo.d4science.org/D-Net/AriadnePlus/src/branch/master/dnetariadneplus/src/main/java/eu/dnetlib/data/collector/plugins/ariadneplus/thanados.

As of October 2022, the ARIADNEplus aggregator collects metadata from 134 endpoints from 33 providers (datasets from 25 of them are already available in the public environment, the others will be by December 2022). For each provider there is at least one endpoint to collect the metadata, one endpoint for the integration of the Getty AAT matching and one endpoint for the integration of the PeriodO collection. In some cases, the provider has more than one endpoint for collecting metadata. For example, ADS has 34 endpoints (one for each collection, exported as a separate dump).

Based on the information provided in the ARIADNEplus form used to gather information for metadata aggregation [7], only 10 providers out of a total of 31 (32,3%) export their metadata with an online

[10] https://ariadne.d4science.org/group/ariadneplus_mappings
[11] ARIADNEplus Generator Policy v1.10
[12] https://thanados.net/sites/sitelist

service (Figure 3), while 21 (676,7%) uploaded their files to a dedicated folder of the ARIADNEplus Mappings VRE.

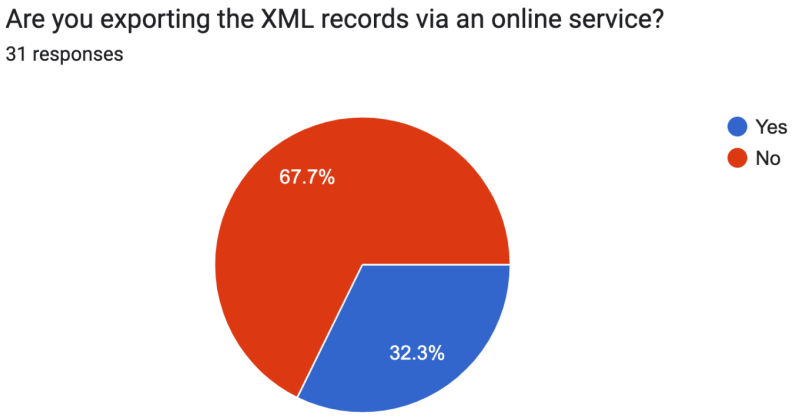

Figure 3 Replies of the providers: are you exporting the XML records via an online service?

Among the 10 providers with an online service for metadata export (Figure 4), 6 of them have an OAI-PMH server, 3 have HTTP APIs, 1 has both an OAI-PMH server and an HTTP API.

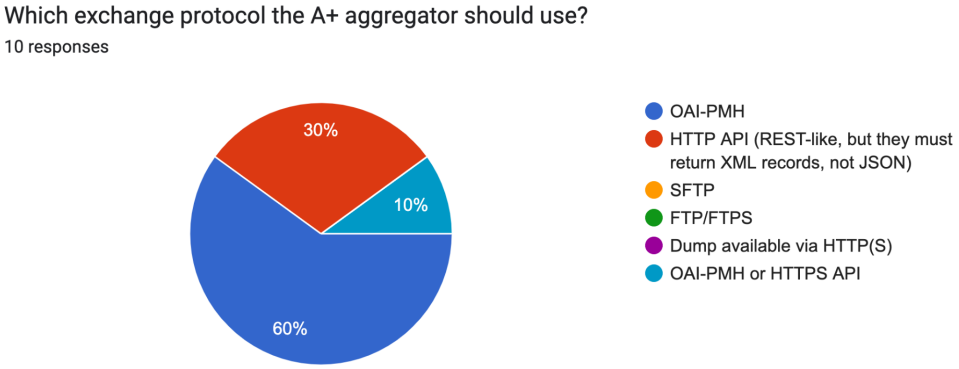

Figure 4 Replies of the providers on the exchange protocols supported by their online services

Providers without an online service have uploaded their metadata as dumps in different formats (Figure 5). The majority (10 out of a total of 21, 47,6%) uploaded one single XML file containing all the metadata records. One provider followed a similar approach, but each single XML file groups records belonging to the same collection. 23,8% uploaded one XML file per record. One provider had a mixed approach (set of files, each with several records) to avoid the creation of huge files in terms of size. Five providers (23,8%) decided to zip their dump to limit the problem of the size. Two providers had to convert their metadata record into XML because they were originally in JSON or CSV. Finally, two providers contributed to ARIADNEplus directly with triples compliant to AO-Cat: their records were directly added to GraphDB and only the part of the aggregation workflow for the enrichment was executed.

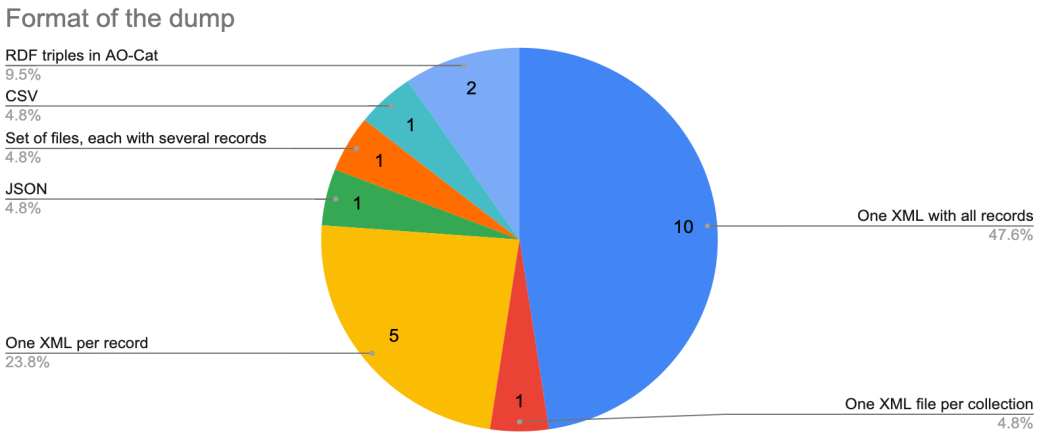

Figure 5 Replies of the providers on the format of their dumps

3.2.2 Deviation from work plan

No deviation from the work plan.

3.2.3 Plans for the next period

By the end of the project, new workflows will be defined for the publishing of aggregated and enriched records directly in the production environment in order to support the automatic update of the ARIADNEplus Knowledge Base and the portal based on the updates made available by the providers. Each dataset will be assigned a workflow that can be scheduled to run autonomously. The scheduling will be set up according to the indications of the data providers, summarized in Figure 6.

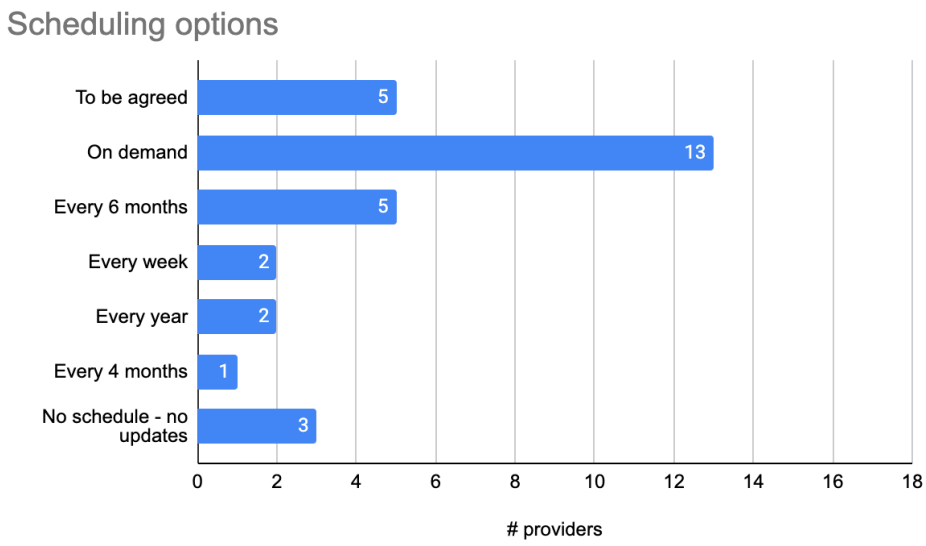

Figure 6 Replies from providers about their preference on the schedule updates

3.3 T12.3 – JRA 1.3 Implementing the ARIADNEplus Portal

This task comprises the user-centred design, implementation and testing of a web application making the services and resources exposed by the previous ARIADNE Portal available through a new enriched ARIADNEplus Portal. First, the Portal developed by this task relies on the ARIADNEplus Data Infrastructure instead of the former ARIADNE Catalogue, the former being a large expansion of the latter with all the resources collected and enriched by the ARIADNEplus project. Second, the discovery service of the new Portal significantly extends the Data Infrastructure search of the former ARIADNE Portal by adding the functionality developed by Task 12.5. This task has designed and implemented the GUI of the discovery service as part of the Portal. The GUI offers a browsing facility, allowing the visual exploration of the ARIADNEplus Knowledge Base according to semantic categories defined in the ARIADNEplus Ontology. Finally, the new Portal will also expose the ARIADNEplus pilots and make the related data and knowledge accessible to the public. Therefore, a special section for the different pilots will be added to the services section of the application.

The portal is accessible at https://portal.ariadne-infrastructure.eu/.

3.3.1 Work performed

Portal Working Group

Development of the portal followed a sprint-based methodology, where the functionality to be developed was discussed, evaluated, prioritized and approved by a set of experts from various partner organizations. To support this effort, a Portal Working Group was formed consisting of technical and non-technical members from SND, PIN, USW, CNR, ADS, and SRFG. The group met regularly on a monthly or bi-monthly basis, as well as regular meetings between individuals as necessary in-between group meetings, to discuss the upcoming development plan and to sign off on the developed functionality.

Staging and production environment

To automate the deployment of new functionalities for the portal, all source code has been contained within a Docker[13] image and uploaded to Docker Hub[14]. Upon image updates, the image is then automatically downloaded from Docker Hub and deployed in a cluster of Docker engines, called a swarm[15]. This swarm is running on the D4Science platform[16], which is used to run all VREs (Virtual Research Environments) produced in the ARIADNEplus project. This automatization means deployment of new functionalities for the staging and the production environments has been done effortlessly, with minimal involvement of personnel resources.

The production environment of the portal, available to the public, is also run as a Docker image deployed in the Docker swarm. For development and showcase purposes, a staging environment has

[13] https://www.docker.com/
[14] https://hub.docker.com/
[15] https://docs.docker.com/engine/swarm/
[16] https://www.d4science.org/

also been deployed on the Docker swarm. This has given the developers the opportunity to show and test new functionalities with a smaller number of users before launching the new feature in production. The staging portal has also been used by partners to quality check the ingested resources before publishing these in the production portal.

New technology stack

The old ARIADNE portal was developed using technology chosen with the intention of having a search portal and a backend feature where partners could login to curate their metadata. A backend framework was selected with this feature in mind that was very fit for that purpose. The backend curation functionality saw little use, and for the ARIADNEplus portal this feature has been removed. Therefore, the login functionality is no longer required. As the code base of the old framework was several major releases behind and in need of security updates, as well as additional code maintenance, a choice was made to migrate the existing functionality of the portal to a new code base, not based on a framework. This choice has also given enough flexibility to the developers in developing the new functionalities of the portal, while using a framework would have added too much unneeded overhead.

Elasticsearch[17] was initially used for persistence of the metadata of the resources. Elasticsearch was a fit-for-purpose database with a built-in search engine where all metadata is stored as JSON documents. Elasticsearch has been developed as a scalable database especially focused on searching large amounts of data in near real-time. As Elasticsearch was changed from a free to a proprietary software during the project, a decision was made to migrate from Elasticsearch to an open-source form of the same software called OpenSearch[18]. This version provided all the functionality needed by the Portal, while retaining the free license. Migration from Elasticsearch to OpenSearch turned out fairly unproblematic.

A mapping document[19] , which contains rules of the structure of the documents, has been developed to ensure that all documents in OpenSearch have the same structure and are interpreted in the same way. The mapping document also provides various structures that facilitate the various search options offered by the Portal, described below.

The backend core of the Portal is developed using PHP[20], which is a common, general purpose, programming language for web development. The core is responsible for communicating with and performing searches on the resources in the OpenSearch software and delivering the results of these queries to the front end. The official PHP Elasticsearch API[21] is used for the communication with OpenSearch.

The frontend of the ARIADNEplus portal is developed using the Vue.js[22] Javascript framework. Vue.js is a framework focusing on the view layer of an application and is used mainly to build user interfaces.

[17] https://www.elastic.co/products/elasticsearch
[18] https://opensearch.org/
[19] https://github.com/ARIADNE-Infrastructure/Index
[20] https://www.php.net/
[21] https://www.elastic.co/guide/en/elasticsearch/client/php-api/current/index.html
[22] https://vuejs.org/

In addition, Typescript[23] is used to provide type checking of the Javascript produced. Vuex[24] is used for the state handling of the application, to manage the various states a component of the frontend application can have at any given time. Tailwind[25] is used as a CSS framework to build the look-and-feel of the application and Font Awesome[26] is used to provide scalable fonts to the application.

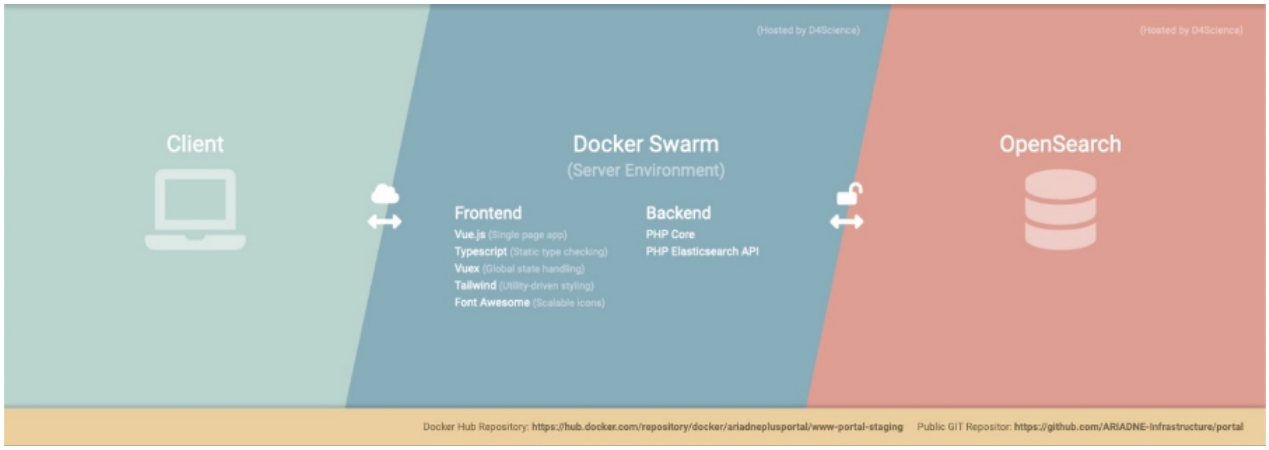

Figure 7 Technology stack of the portal

Portal functionality

The portal is developed with three keywords in mind, in addition to the standard free-text and faceted search options. These keywords, when – where – what, revolve around the concepts of temporal coverage, spatial coverage, and topical coverage. These keywords are focal points in the different findability options a user has when using the portal.

The front page of the portal has been refined to a more streamlined and simplified experience. Here the user can directly carry out a free text search in all resources, with an option to narrow down the search to a specific field. The capabilities of the Getty Arts and Architecture Thesaurus (Getty AAT) have been used to enable a multilingual search with suggestions, when selecting Getty AAT Subject as the search parameter. This feature enables the user to search for resources using any language where Getty AAT has a translation of the term.

[23] https://www.typescriptlang.org/
[24] https://vuex.vuejs.org/
[25] https://tailwindcss.com/
[26] https://fontawesome.com/

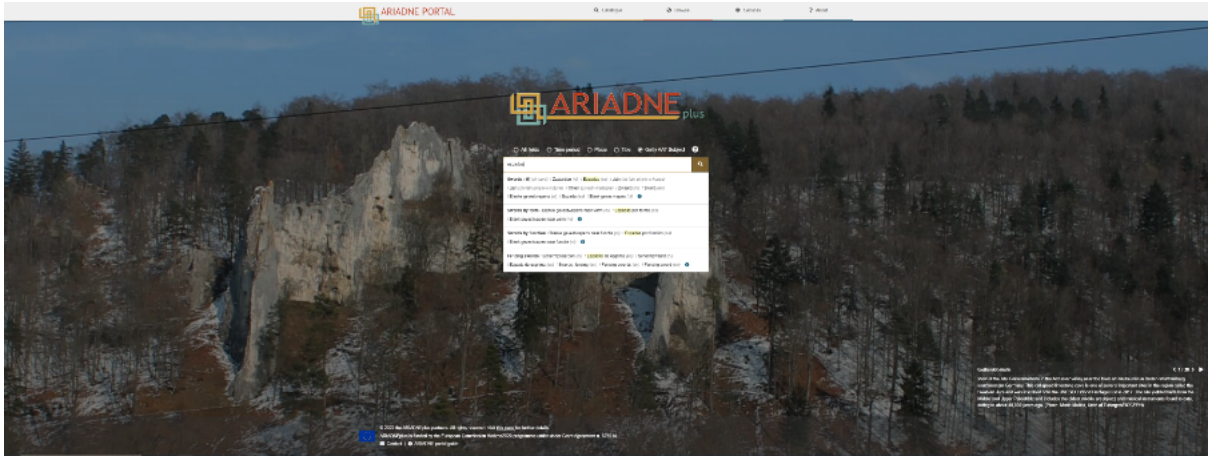

Figure 8 Front page of the portal featuring multilingual suggestions from Getty AAT

The "when" keyword represents the time period covered by a resource and it is enabled by a Timeline search option where the user can choose a year span for browsing resources covering a specific period. In addition to the time span, resources can be filtered using various facets to further narrow down the search results.

Figure 9 Timeline search

The "where" keyword represents the physical location of a resource and it is enabled by a Map based search option, where the user can choose an area on a world-wide map to browse resources within that specific area. Several visualization options are available to help the user narrow down the search results. On a high zoom level, clustered resources are shown using a "heat map"[27], which allows the

[27] A "heat map" graphically represents the number of resources available in one point, with colors ranging from green (few) to red (many).

user to see areas with a high density of resources. On lower zoom levels, marker clusters or individual markers are shown where the density of resources for each map tile is lower. Additional visualization options are available using various map layers, where the user can choose a suitable option for their preferences.

A new feature available In the ARIADNEplus portal is the possibility to explore geo shapes (polygons, lines, etc.). These resources are shown on the map using a different marker indicating the centroid of the geo shape. The complete geo shape is shown on the resource landing page.

For resources where exposing the exact position would provide a risk, e.g. due to plundering, a different indicator is shown on the map indicating this to be an approximate position.

As with the timeline search, additional filtering is possible using facets. Moreover, the user can select a map layer suitable for their preference. Examples of these map layers are topological or satellite views.

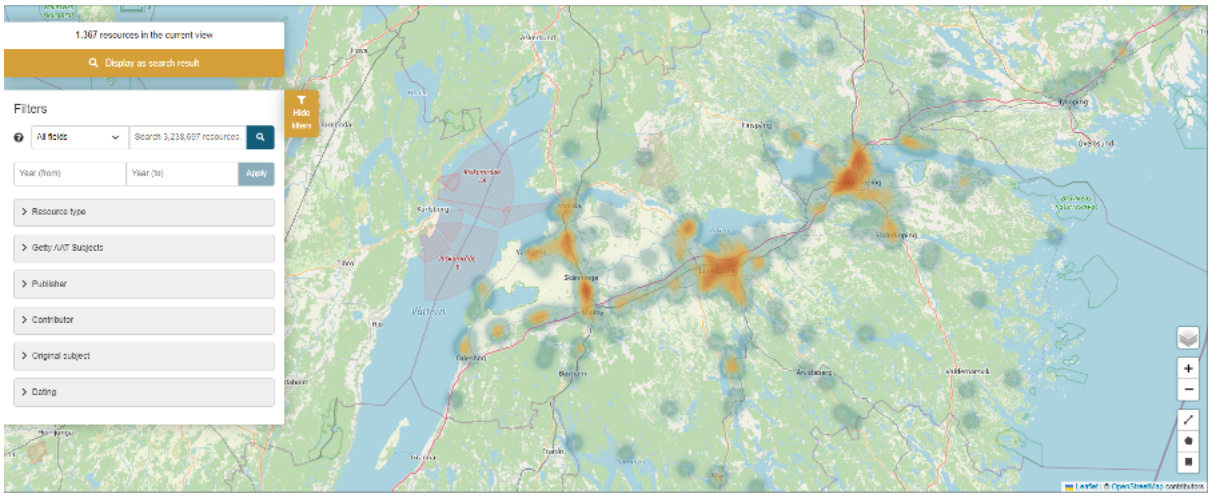

Figure 10 Map search with heatmap

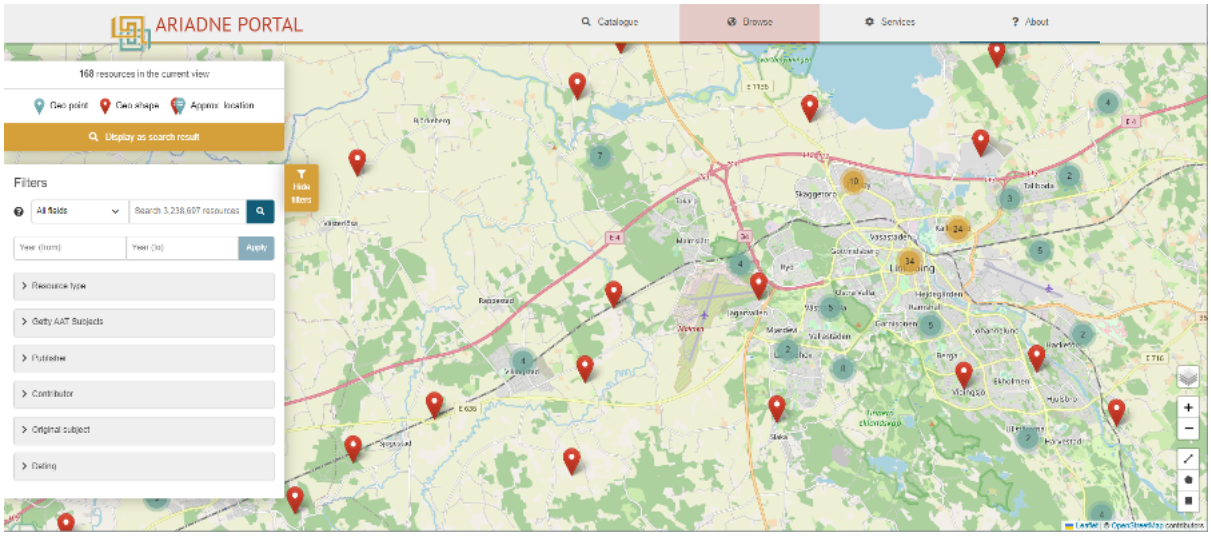

Figure 11 Map search with cluster markings and individual markers

The “what” keyword is represented by a word cloud using terms from the Getty Arts and Architecture Thesaurus (Getty AAT). The topical coverage of the resources is normalized in the ingest phase so that all resources are using the same vocabulary. The user can, using the word cloud, explore resources by clicking on terms of their interest. Size and colours of the terms in the word cloud are calculated based on the number of resources using each respective term. The word cloud has been enhanced with additional filtering capabilities to be able to narrow down the search to e.g. resources from a specific publisher or time period.

Additionally, the “what” keyword has been represented on every location where Getty AAT terms are shown, with a link to a Term landing page, where the user can get information about the term, its translations, mappings from native terms, and a list of resources that reference the term.

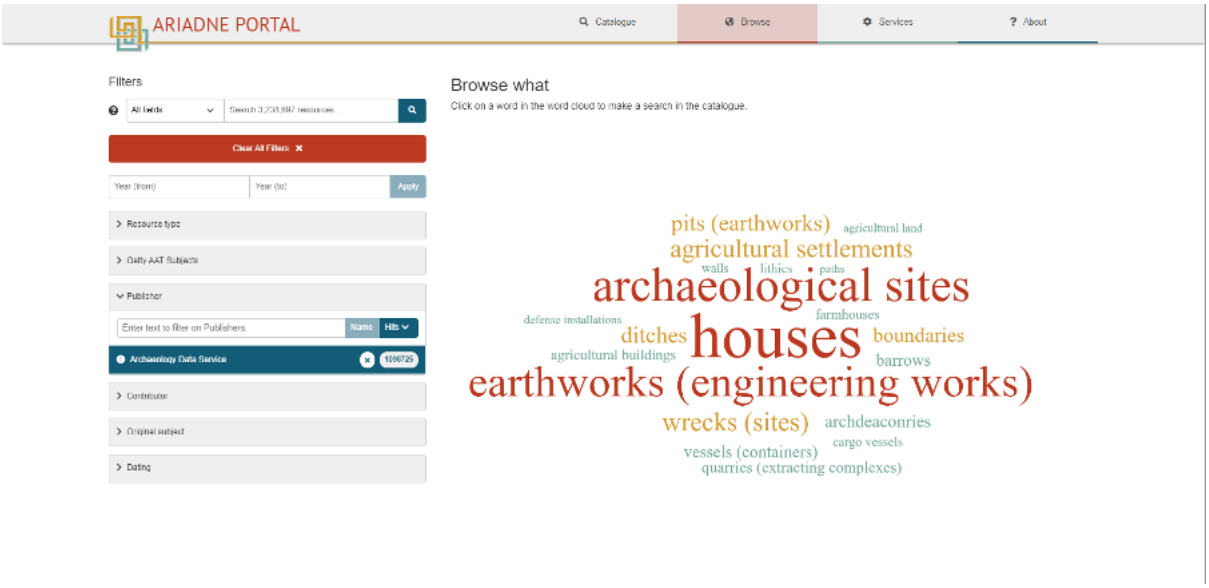

Figure 12 Word cloud featuring filtering on Archaeology Data Service

Resource landing pages have been developed to include as much information as possible about the resources. This includes presenting all useful metadata available about the resource. Links to the original resource located at the provider site are readily accessible, and a map showing the resource is provided if the metadata contains spatial information. The map has been enhanced by showing polygons and other geographic shapes in those cases where such information exists.

The user is also given the opportunity to find similar resources, either by using the map to find nearby resources, or by using the “Thematically similar” feature, where similarity is calculated on a set of different metadata.

Another feature of the landing pages is the possibility of displaying sample images in those cases where they exist at the provider. These sample images provide the user with an additional layer of qualitative metadata as a supplement to the regular metadata.

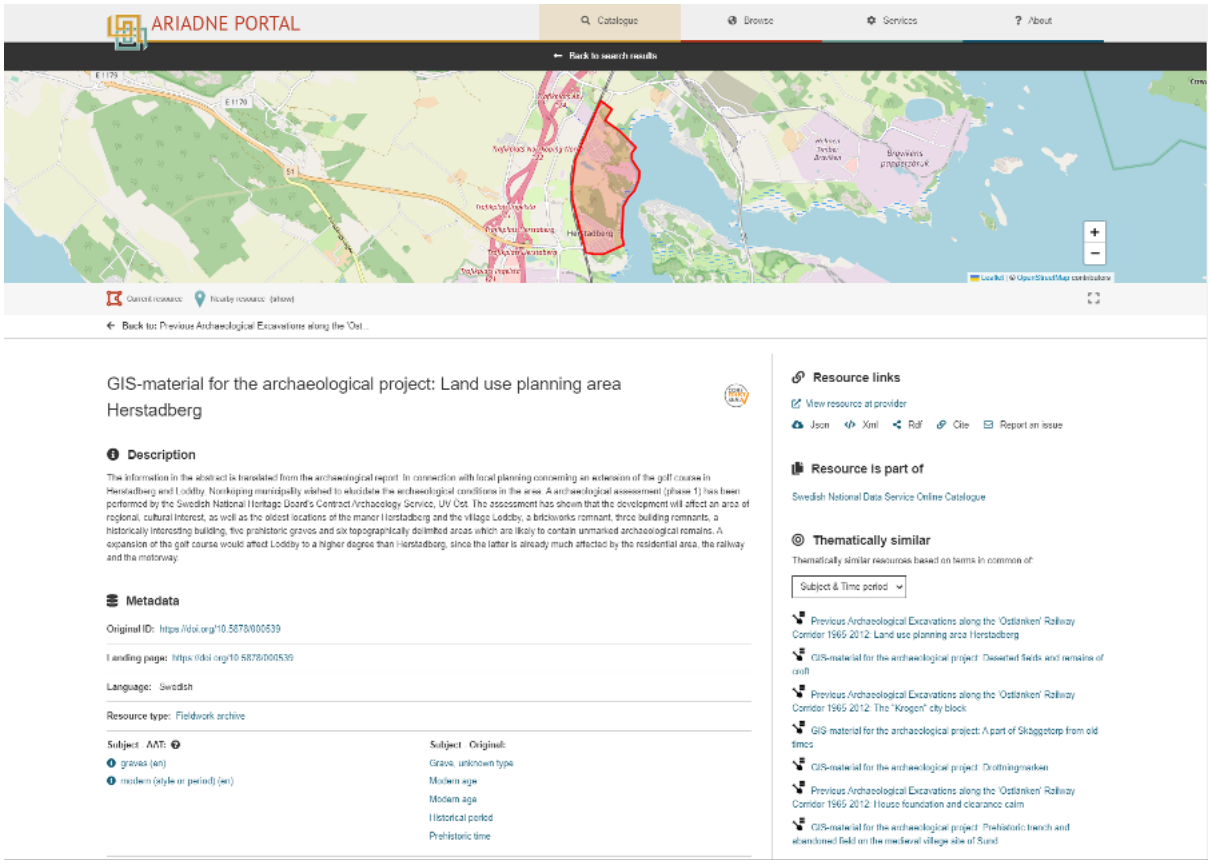

Figure 13 Resource landing page featuring polygon display

Source code and licensing

All source code is released on the well-known source-code hosting platform Github[1] under the organization ARIADNE-infrastructure[28].

Source code is distributed using the European Union Public License v1.2[29].

3.3.2 Deviation from the work plan

The launch of the new portal, initially planned in M12 (December 2019) was postponed due to the major corrections and updates required, as described above. A first beta version has been available to the consortium members since M15 (March 2020). The new version was launched in the first half of 2021, with several additional releases provided since then.

3.3.3 Plans for the next period

Development of the ARIADNEplus portal will be ongoing for the remainder of the project, although focus in the last months of the project will be on documentation and bug fixes to leave the software

[28] https://github.com/ARIADNE-Infrastructure
[29] https://joinup.ec.europa.eu/collection/eupl/eupl-text-eupl-12.

as stable and well-documented as possible. There are two features listed below which need to be finalized and go through quality control before release and thus are not yet in the production portal. If these features don't pass quality control, they will be left out of the final production portal.

Hierarchical genre expansion

This feature will enhance the search functionality when filtering using Getty AAT by enabling a hierarchical expansion of the search query which will include all descendant terms of a found term in the thesaurus. Resources in the portal are tagged with terms from the thesaurus but lack information about terms lower in the hierarchical tree. This will currently result in a poorer search result where e.g. a search for weapon will only return resources directly tagged with the term "weapon" and not those tagged with terms lower in the hierarchy as "bows" or "clubs".

Temporal search enhancements

Currently, the temporal search function is a timeline based on a start year and an end year, which the resource covers. Considering the fact that archaeological time periods vary enormously from region to region, an additional option has been developed, using the PeriodO time periods to give the researcher the opportunity to find and compare resources belonging to the same time period of specific regions. There have been concerns about the quality of the results given by this feature, as the metadata of many resources lack the correct URI of the corresponding PeriodO time period, which is needed in order to be able to link resources to the selected time period. This feature will go through quality control in the staging portal to investigate the feasibility of releasing it in the production portal.

3.4 T12.4 – JRA1.4 Item-level data integration

The overall objective of ARIADNEplus is to provide a deeper (than AO-Cat can achieve) integration of data resources of the Archaeological Research Communities to enable integrated searches. These deeply integrated datasets may be exploited via specialized Virtual Research Environments (VREs) that would be specifically designed to support specific searches of interest to the relevant research community/group. There is no limit to the extent and scope of deeply integrating resources.

The work is unfolding in three directions:

1. Identifying research questions where relevant data from multiple sources are available.
2. Developing the application profiles, thus adding the expressive power needed to represent these data to the ARIADNEplus ontology.
3. Building the mappings from the original schemas to the extended ARIADNEplus ontology, using the relevant application profiles, and transforming the data into Linked Open Data (LOD) to be accessed via a SPARQL endpoint or a specialized VRE.

As explained below, three application domains were chosen and formed the core activity of this Task during the second half of the project: mortuary archaeology, numismatics, and epigraphies.

3.4.1 Work performed

The work performed in this task is closely related to the work of WP4 and WP14. The major effort of the project concentrated towards dataset aggregation, as described in WP4, the mappings to AO-Cat, the development of the application profiles (see Deliverables D4.3 "Final report on dataset integration" and D14.2 "Final report on the ARIADNEplus knowledge management system") and the presentation of the aggregated resources through the ARIADNEplus portal (see Section 3.3).

Following the first mappings, and after validating the suitability of AO-Cat to describe the provided resources, we decided to focus on three domains mature enough and suitable for deeper integration: the domains of mortuary archaeology, numismatics and epigraphies.

Mortuary archaeology

Mortuary archaeology is the study of human remains in their archaeological context. As described in the Mortuary Data Application Profile [9], it "consists of a series of research activities and analyses carried out either directly on mortuary evidence (archaeological evidence containing human remains or contexts that are interpreted to relate to the disposal of the dead), and/or on documentation and finds (human remains, objects, samples) from such contexts."

Important information for the interpretation of mortuary records comes from human osteology and other science-based analyses of human remains (e.g. ancient DNA analyses). Hence, mortuary archaeology is closely related to other scientific fields and integrates their results for synthesis. In the deliverable D14.2 "Final report on the ARIADNEplus knowledge management system" we describe the alignment of the Mortuary Data Application Profile with the Ancient DNA Application Profile and the Heritage Science Application Profile.

In several cases, research questions can be answered only if queries can be executed across different datasets such as cemetery databases, excavation databases, scientific analysis databases (aDNA, radiocarbon, environmental data) that have all been integrated via appropriate mappings to the ARIADNEplus research ecosystem using AO-Cat and appropriate Application Profiles. Research questions across different datasets include:

- How mortuary practices have developed over time in relation to gender or age groups?
- How kinship has influenced mortuary treatment or nutrition?
- Are the individuals that are buried as a group in the cemetery area or in very richly furnished graves (archaeology database) biologically related to each other (aDNA Data)?
- Did they have similar nutrition and evidence for mobility (stable isotopes data)?
- Are there differences between male and female individuals (biological anthropology data) in regards to nutrition and mobility (stable isotopes), burial practices (archaeology), health status, evidence for physical labour, injuries (biological anthropology) or evidence for migration (aDNA)?
- How do the burial practices and any other aspects mentioned above change over time (radiocarbon data)?

Two papers, “Integrating data on early medieval graves from different resources” and “The ARIADNE Mortuary Data Application Profile” were submitted to the CHNT Conference on Cultural Heritage and New Technologies (10-12 Nov 2022, Vienna, Austria) reflecting the work in this domain.

Table 1 summarizes the selected mortuary datasets and the relevant mappings for item-level data integration.

Table 1 Mortuary datasets for item-level data integration

Mortuary archaeology	Mapping
In Touch with the Dead: Early Medieval Grave Reopenings in the Low Countries	684
THANADOS, the Anthropological and Archaeological Database of Sepultures.	945
AIS AKB, Archaeological information system for registered archaeological sites in Bulgaria	657
Urnfield Culture Franzhausen-Kokoron Cemetery. Revised mapping performed during ARIADNE 1. A database of 403 cremations from the younger urn field culture (1050-800 BC).	681, 682
aDNA lab projects. Mappings for each step of the scientific analysis workflow of the Allentoft protocol. Details for these mappings are presented in D4.2 Section 6 and D14.1 Section 4.	687-693
Paliambela Kolindros Archaeological Project, Trench 01. The entire dataset provides a data repository of all excavation information recorded during fieldwork or created through subsequent post-excavation study.	668

Numismatics

Numismatics, the study of monetary objects, is represented by nomisma.org[30], a very active community established in 2010. Since the 1967 Congress in Copenhagen, the International Numismatic Council publishes a Survey of Numismatic Research, a volume that provides an

[30] http://nomisma.org

authoritative critical bibliography on the progress of numismatic studies between congresses with a focus on methodology and new discoveries [8].

In ARIADNEplus we did not implement yet another Coin Application Profile, instead we decided to integrate the AO-Cat description of coins using CIDOC CRM core and nomisma.org. An experimental platform based on Research Space[31] (Figure 14) was implemented in order to test specific queries. A typical problem that nomisma.org helps to solve is the diverse terminology used in the different datasets. For example, in the three datasets that we used, the terms *AE*, *bronze* and *copper alloy* are used to represent the material of a coin. By mapping the three terms to the common http://nomisma.org/id/ae term (Figure 15) we guarantee that a search both with native and nomisma terms will succeed.

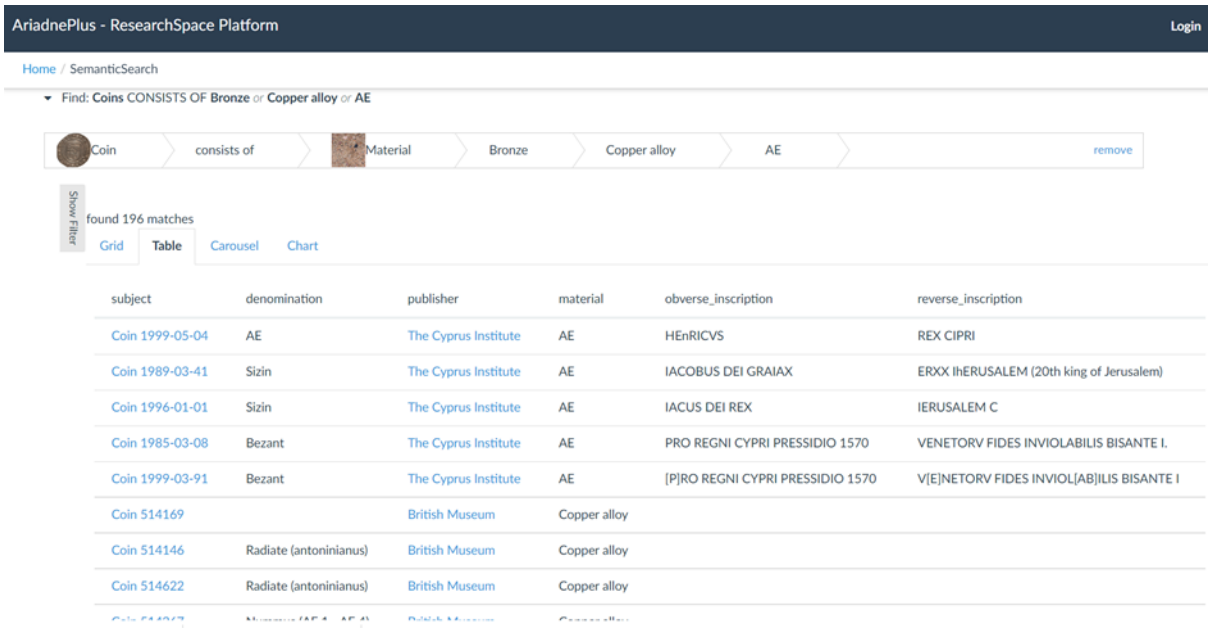

Figure 14 The Research Space platform for Coins' exploration

[31] https://researchspace.org/

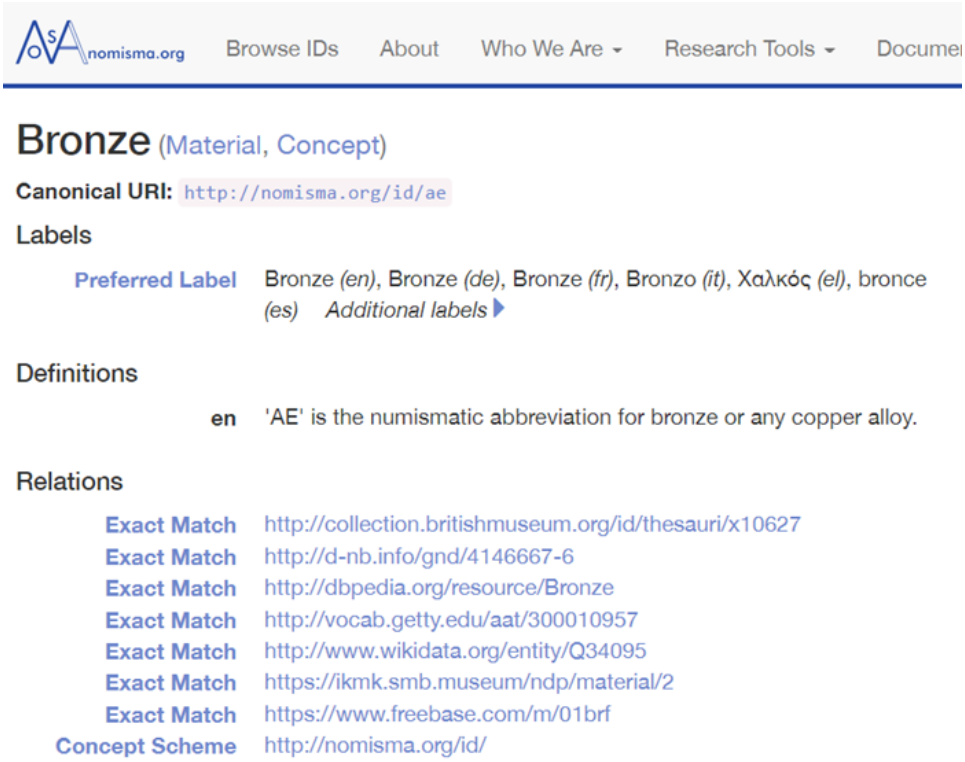

Figure 15 Bronze in nomisma.org

Queries across different datasets include:

- Get a list of mints.
- Get a list of denominations.
- Get the weight and diameter of a coin.
- Get coins with inscriptions (epigraphic data)
- Find mints that issued bronze coins.
- Find coins issued by a specific mint/issuer.
- Get all findspots for all hoards or coins associated with silver.
- Get findspots for coins minted during a specific period.

Table 2 summarizes the selected coin datasets and the relevant mappings for item-level data integration.

Table 2 Coin datasets for item-level data integration

NUMISMATICS	Mapping
British Museum Portable Antiquities Scheme (PAS) Coins, records of the Oxford University Celtic Coin Index (CCI- prefix) and Cardiff University`s Iron Age and Roman coins of Wales project (IARCW- prefix).	*943*
Cypriot Medieval Coins Collection, digital library dedicated to the study, promotion and dissemination of the history of Medieval Cypriot Coinage	*891*
AFE-RGK coin collection, the coin find database of the Roman-Germanic Commission (RGK) of the German Archaeological Institute (DAI)	*954*
DIME collection, find records from the Danish recording portal for archaeological finds produced by members of the public. It consists of over 100.000 records of portable antiquities.	*697*

Epigraphies

Inscriptions, marks and graffiti are peculiar objects, characterized by having a close relationship with the physical objects by which they are carried. This implies that their real meaning cannot be fully understood without the analysis of the object or monument or other archaeological objects on which they appear. Conceptually, an inscription is an element with physical characteristics that are themselves bearers of meaning and of valuable information going far beyond the inherent meaning of the text. For instance, the shape, the materials, the production techniques, the wiring direction and other similar elements can become fundamental not only for their understanding but also for the definition of their nature.

From a conceptual point of view an inscription can be analysed according to three main aspects: the text-bearing object or monument, obviously involving archaeological topics, the text and its obvious correlations with content and linguistic aspects, and the feature engraved on the support in the form of letters or other symbols, which is the central element that characterizes and differentiates an inscription from any other manifestation of written communication.

The specific archaeological aspects (discovery, provenance, archaeological context etc.) relating to the physical support can be documented using AO-Cat and the CRMarchaeo extensions of CIDOC CRM. The specific Application Profile for modelling and expressing knowledge about inscriptions and all their features, is based on CRMtex, an ontological extension of CIDOC CRM created to describe ancient texts and other semiotic features appearing on inscriptions, papyri, manuscripts and other media. The model is also designed to describe in a formal way the

phenomena related to the production, use, conservation, study and interpretation of textual entities.

In modelling inscriptions information, we have verified that integration of CRMtex (Inscriptions, marks and graffiti Application Profile, CRMtex v1.1) with AO-Cat and CIDOC CRM is perfect and allows the formulation of cross-sectional queries such as:

- Get a list of archaeological objects bearing an inscription.
- Find all the inscriptions written by a certain writer (see Figure 16, the result of the search for inscriptions written by Nicocles).
- Get a list of languages and scripts used in the inscriptions.
- Get a list of inscriptions written on stone objects.
- Find all inscriptions that have a transcription.
- Find all writers who have created inscriptions on a certain type of object (e.g. memorial stones, tombstones, coins).
- List all the places where an inscription was found.
- Get a list of scholars who have been involved in the study of inscriptions.
- Get a list of inscriptions having a certain ductus.
- Get a list of inscriptions in a certain language and encoded in different scripts.
- Find all the texts written in a certain language and encoded in a certain script, having a certain ductus (for example, left-handed) and, written on an archaeological support of a certain type, found in a certain place and dated within a certain timespan.
- List all the inscriptions found in a certain place and studied by a given scholar over a certain period of time.
- Find all the transcriptions of inscriptions in a specific language and of a specific era, and the names of the scholars who made them.

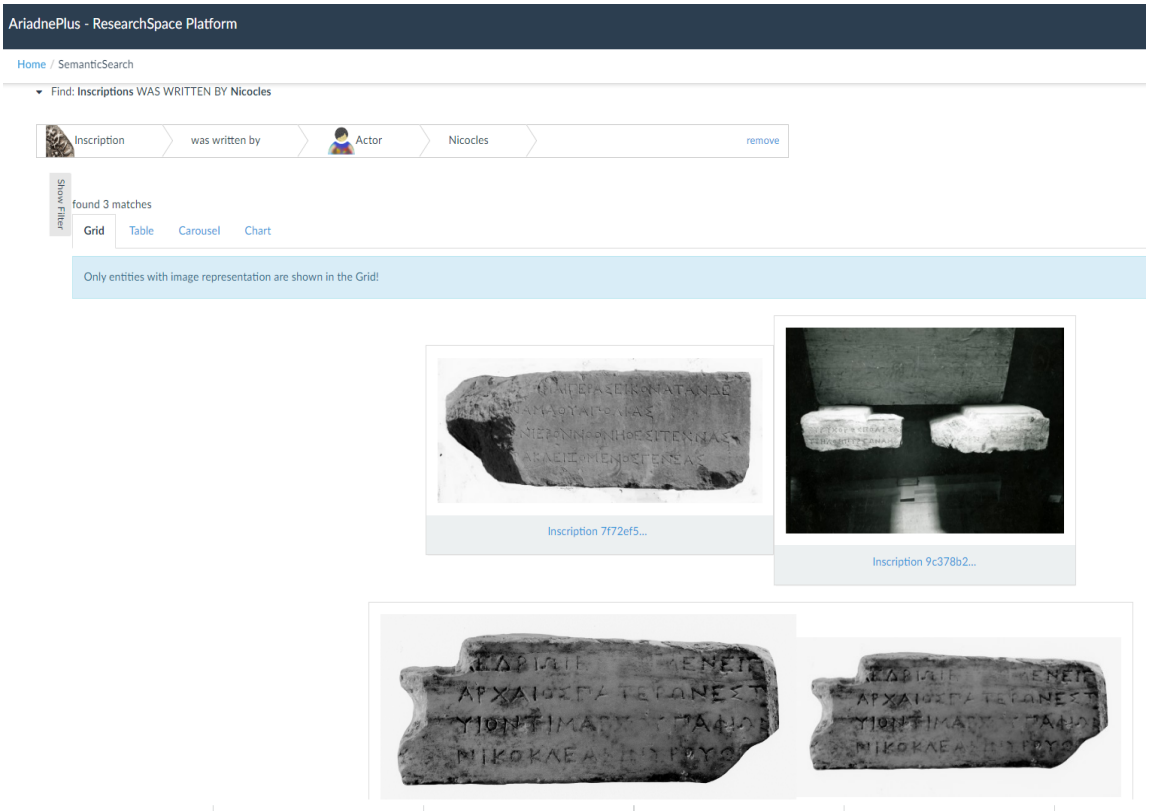

Figure 16 The Research Space platform for exploration of epigraphies

Table 3 summarizes the selected epigraphic datasets and the relevant mappings for item-level data integration.

Table 3 Mappings of epigraphies for item-level data integration

EPIGRAPHIES	
Archaia Kypriaki Grammateia, a corpus of ancient texts that includes a wide range of literary genres such as epic, lyric and dramatic poetry, epigrams inscribed on stone, prose, medical and philosophical texts and covers the ancient Cypriot literary production in a time span of circa thirteen centuries.	*911*
CEIPAC- Corpus de epigrafía anfórica latina, collection of seals, picti titles and graffiti on Roman amphorae,	*971*

3.4.2 Deviations from the work plan

There was no deviation from the work plan. The work in this task started, as expected, during the second year of the project due to its strong dependency to the work performed in WP4 and WP14. Its major activity was during the second half of the project.

3.4.3 Plans for the next period

The activity on item-level integration will continue to run experimental queries on the Research Space platform until the end of the project.

4 Conclusions

Data integration and interoperability call for tools and services capable of addressing different scenarios and requirements. In the first two years of the ARIADNEplus project, WP12 delivered the infrastructure to create, update, enrich and access a knowledge graph containing information collected from the members of the consortium, interlinked with standard ontologies of the cultural heritage domain like PeriodO and Getty AAT.

The infrastructure has been designed and developed according to the requirements of the consortium, giving high importance to data curation and automation of the aggregation, enrichment, and publishing processes. In fact, the infrastructure features two environments: a staging environment where the collected data is aggregated, added to a staging AC and indexed on a staging portal where data experts can manually verify the quality of the data. Upon confirmation from data experts, data is then pushed to the production environment, featuring the public ARIADNEplus portal. The staging environment is only available to the consortium members to check data quality and test the new functionalities of the portal. The tools offered to data experts for mapping the data into the AO-Cat model (3M Editor) and to link them to Getty AAT subject terms (Vocabulary Matching Tool) are fundamental to ensure the quality of aggregated data. Content integration and harmonisation is facilitated by the ARIADNEplus aggregator, built on top of the D-NET software toolkit, that enables the realisation of highly configurable and autonomous aggregation systems. The knowledge base is stored on a GraphDB server and the data is organised in such a way that the graph can be fed automatically by the workflows running on the aggregator.

The knowledge base is accessible via the Graph DB Graphical User Interface and a SPARQL endpoint. User documentation on human and programmatic access was produced and made available to users.

The content of the knowledge base is also exported to an OpenSearch full-text index server that powers the ARIADNEplus portal developed by T12.3.

Task 12.4 addressed the topic of item-level integration, that is a deeper integration of item-level data to support research questions that require information richer than what is available in AO-Cat. It proved the feasibility of integration in three application domains and the need for specialized VREs and vocabularies to support querying and access.

5 References

1. Bardi Alessia, Marberg Johan Fihn, & Theodoridou Maria. (2021). D12.2 – Mid-term report on data integration. Zenodo. https://doi.org/10.5281/zenodo.4922902
2. Assante Massimiliano, Cirillo Roberto, Dell'Amico Andrea, & Pagano Pasquale. (2021). D13.2 – VREs Operation Mid-term Activity Report. Zenodo. https://doi.org/10.5281/zenodo.4926516
3. FAIRsharing.org: PAV; Provenance, Authoring and Versioning Ontology DOI: 10.25504/FAIRsharing.nbfwwv, Last Accessed: Wednesday, October 5th 2022, 17:04
4. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K. et al. PAV ontology: provenance, authoring and versioning. J Biomed Semant 4, 37 (2013). https://doi.org/10.1186/2041-1480-4-37
5. FAIRsharing.org: PROV-O; W3C Provenance Ontology DOI: 10.25504/FAIRsharing.2rm2b3, Last Accessed: Wednesday, October 5th 2022, 17:04
6. Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D. and Pagano, P. (2014), "The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures", Program: electronic library and information systems, Vol. 48 No. 4, pp. 322-354. https://doi.org/10.1108/PROG-08-2013-0045
7. Bardi. (2022). ARIADNEplus questionnaire responses for metadata aggregation (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7221186
8. Survey of Numismatic Research 2014–2020, International Numismatic Congress https://www.inc-cin.org/assets/pdf/survey/Survey%202014-2020.pdf
9. Edeltraud Aspöck, Maria Theodoridou, & Achille Felicetti. (2022). Types of burial data and proposal of a Mortuary Data Application Profile for ARIADNEplus (WP4.4.14) (1.1). Zenodo. https://doi.org/10.5281/zenodo.7243782