This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

GRANT AGREEMENT NUMBER:	**693548**
PROJECT ACRONYM:	**ArchAIDE**
PROJECT TITLE:	**Archaeological Automatic Interpretation and Documentatic of cEramics**
FUNDING SCHEME:	**H2020-REFLECTIVE-6-2015**
PRINCIPAL INVESTIGATOR	**Prof Maria Letizia Gualandi, UNIPI**
PROJECT COORDINATOR	**Dr Gabriele Gattiglia, UNIPI**
TEL:	**+39 05022 15817**
E-MAIL:	**maria.letizia.gualandi@unipi.it** **gabriele.gattiglia@for.unipi.it**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

D 10.4 Web APIs for REST Services

Version: 1.0

Work Package:	10 Archiving and Dissemination
Lead Author (Org):	Tim Evans (UoY)
Contributing Author(s) (Org):	Holly Wright (UoY)
Due Date:	M32
Date:	31 January, 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Description
0.1	24/01/2019	Tim Evans	Full draft
0.9	25/01/2019	Holly Wright	Draft edit
1.0	30/01/2019	Gabriele Gattiglia	Final revision

Disclaimer

Table of Contents

Abbreviations

WP: Work package

M: Month

UNIPI: Università di Pisa

UoY: University of York

UB: Universitat de Barcelona

UCO: Universitaet zu Koeln

TAU: Tel Aviv University

CNR: Consiglio Nazionale delle Ricerche

INERA: Inera srl

BARAKA: Baraka Arqueologos S.L.

Elements: Elements centro de gestio i difusio de patrimoni cultural

Executive summary

Deliverable 10.4 documents completion of the ArchAIDE Semantic Web database for use by Web APIs and REST services as described in Task 10.2 and developed by UoY with support of UB and UCO and INERA.

Following a Linked Data approach, the project has successfully uploaded the multilingual vocabularies to the ADS (UoY) graph database, creating a resource, which can be accessed, integrated and manipulated in a variety of Web APIs and REST services. The data is aligned with existing Linked Data vocabularies, allowing live cross-searching of ArchAIDE data in comparison to other datasets which are using the same semantic standards. It is hoped that this work will meant that the ArchAIDE vocabularies will not only be found and used by other projects recording archaeological ceramics, but also provide an extension to the growing corpus of Linked Open Data that looks to establish simple technical and semantic interoperability between diverse and distributed datasets and gazetteers.

1 Introduction

In Work Packages two and three, the ArchAIDE project partners successfully developed a single database instance that contains the definitions of pottery types and supporting documentation used by the public facing application. From the outset of the project, it was clear that a significant component of this application would be the facility to contain multilingual vocabularies to enable linguistic interoperability within the dataset being assembled. However, early in WP2 discussion between project partners quickly identified the difficulty of semantics when faced with the task of describing either a fragment of, or the whole vessel. For example, the difference between what may be referred to as a "dish", "plate" or "platter". Furthermore, there arose a need to establish a scheme of recording that was not only consistent across languages, but also archaeological traditions. Whereby people working in different countries may not only use different words but also varying hierarchies or classifications for how they categorise an object. As a hypothetical example, one tradition may have four different words to categorise what kind of candlestick they are describing, another tradition may simply have one.

The solution was the creation of a discrete set of vocabularies to account for, and reconcile these differences. These have been used as the data that forms the backbone of the ArchAIDE reference database, and will also allow different catalogues - and therein different classification schemes - to be cross-searched and then presented by the ArchAIDE application. This basic functionality is however only the first part of what has been created. Following the precedent set by the EU-funded ARIADNE project, the vocabularies have been built with an eye on creating a resource with maximum interoperability with both current initiatives and with future projects. It has achieved this through taking a so-called Linked Data approach (ARIADNE 2016). The initial mapping exercise has established a concordance between ArchAIDE and a neutral Linked Open Data (LOD) vocabulary, and a later phase has successfully created the projects own LOD resource through hosting by the ADS; the work which forms the focus of this report. The creation of LOD for ArchAIDE allows project data to be easily cross-searched as part of the wider Semantic Web, and a format that intuitively lends itself to use through a range of methods and processes that use established web protocols to interact with, and retrieve data. These are often very broadly referred to as Application Programming Interfaces (API)

The following report covers the methods, decisions and processes involved in creating of the vocabularies with an overview of the Linked Data approach taken within that phase. The report then covers the creation of the ArchAIDE LOD, with examples of its practical implementation as well as possibilities for future innovations and re-use.

2 Multilingual vocabularies

The methods and results of the package of work creating the vocabularies, and the implementation within the ArchAIDE reference database, is covered in extensive detail in D3.1. Key elements are presented below to provide a simple introduction to process.

2.1.1 Methodology

Each of the project partners (UCO, UoY, UNIPI and UB) contributing data to the reference database, worked through their catalogues and identified the descriptive terms used within that refer to, or classify significant elements of a ceramic form. Learning from the methodology and using the tools developed for the (European

Commission-funded) ARIADNE project by the Hypermedia Research Group at the University of South Wales, a neutral spine to which partners could map these terms was deemed a preferable solution to establishing a whole new ontology bespoke to this project. As with the ARIADNE project, use of Getty Institutes Art and Archaeology Thesaurus (AAT)[1] proved a suitable candidate, with the added value of interoperability with ARIADNE and *any other data* also mapped to the AAT, thus giving the mapping work done by ArchAIDE a strong sustainable base. In the initial phase of the project all partners agreed on a subset of AAT terms that would be used for this neutral spine, describing the following methods of recording pottery from archaeological excavations:

- The sherd type
- The vessel form
- The decoration type
- The decoration colour

Later in the project, partners established a need for controlled terminologies to describe the type or characteristics of specific parts of a ceramic vessel. At the time of writing the AAT does not include this level of detail. As an alternative, and as a concerted effort to ensure that ArchAIDE data was interoperable with past and future projects, the project used the concepts of recording established by the original creators of the Roman Amphorae database (Simon Keay and David Williams). The concepts defined by Keay and Williams have since been converted to SKOS by a separate research project and then made available as Linked Open Data via the ADS (UoY) triplestore[2]. The following vocabularies were thus used as the spine to be followed by the ArchAIDE project:

- The characteristics of the rim
- The characteristics of the neck
- The characteristics of the shoulder
- The characteristics of the body
- The characteristics of the base
- The characteristics of the handle

As noted above, it is important to highlight the benefit of aligning the terms used by ArchAIDE with existing vocabularies, in which they offer short-term benefits of concordance and definition, but also longer-term by allowing the archive to be re-used with an inherent and unambiguous understanding of the terminologies used within. The mapping of 'native' to neutral terms was undertaken by each project partner using the mapping tool created by the ARIADNE project. The template uses a simple iteration of SKOS which maps the native label (source) to the ARCHAIDE term (target), and defines the level of match. In the illustrative example below from the Italian mapping, one can see the benefit of this approach by allowing the mapping of *unguentario* and *balsamario* to *ointment vessel*, but also setting the preciseness of that match in conceptual terms. Thus for example *balsamario* has no close or exact match, being a very particular way of describing a small container used to store balsams, but hierarchically has a broad match to what the Getty calls an ointment vessel.

[1] http://www.getty.edu/research/tools/vocabularies/aat/

[2] http://data.archaeologydataservice.ac.uk/page/romanamphorae/roman_amphorae_characteristics

olla	skos:closeMatch	*pot*	http://vocab.getty.edu/aat/300045697
bacino	skos:exactMatch	*basins*	http://vocab.getty.edu/aat/300045614
incensiere	skos:exactMatch	*incense burner*	http://vocab.getty.edu/aat/300198817
unguentario	skos:closeMatch	*ointment vessel*	http://vocab.getty.edu/aat/300197583
balsamario	skos:broadMatch	*ointment vessel*	http://vocab.getty.edu/aat/300197583
lucerna	skos:broadMatch	*lamp*	http://vocab.getty.edu/aat/300037592
candeliere	skos:closeMatch	*candlestick*	http://wvocab.getty.edu/aat/300037588
bugia	skos:broadMatch	*candlestick*	http://wvocab.getty.edu/aat/300037588
colino	skos:closeMatch	*colander*	http://vocab.getty.edu/aat/300200469
scaldino	skos:closeMatch	*bed warmer*	http://vocab.getty.edu/aat/300375242
scaldavivande	skos:closeMatch	*dish warmers*	http://vocab.getty.edu/aat/300209583
brocca	skos:closeMatch	*pitchers (vessels)*	http://vocab.getty.edu/aat/300194765
boccale	skos:closeMatch	*pitchers (vessels)*	http://vocab.getty.edu/aat/300194765

Fig. 1: Screenshot of the SKOS matching tool being used for Italian descriptions of ceramic form types

The final mappings were reviewed, to reflect and then incorporate any misunderstandings or inconsistencies. At a later date a contribution to ArchAIDE blog[3] drew attention to previous work in this sphere by Caroline Sourzat, and subsequently a copy of her Master's thesis was used to add French terms, and enhance the German terminologies. A further contribution of an historic journal article from an attendee of the December 2017 multiplier event, provided a basis for a mapping in Dutch, which was edited and refined by Leontien Talboom, a Digital Archivist at the ADS (UoY), and a native Dutch speaker. In late 2018, a mapping in Portuguese has also recently been contributed by Guilherme DAndrea Curra. At the end of this phase, a total of 1338 mappings in seven European languages had been completed.

2.1.2 Use of the multilingual vocabularies in next phase of development

The mapped vocabularies have been uploaded to the ArchAIDE reference database as a reference resource and are being used for the manual and automated (via text recognition) recording of paper and digital ceramic catalogues. The vocabularies will also be the spine which will allow users of the public application to cross-search the reference database in their own language, and as noted to reconcile differences in understanding or classification. As demonstrated below in a simplified representation (Figure 2), this allows a user with only a very broad understanding or certainty of what they may be looking at (in this case a "jar"), to be able to cross-search any typological form that has been exact, close or broadly mapped to this concept.

[3] http://www.archaide.eu/blog/-/blogs/248039?groupId=20181&pk=248039&userId=23312

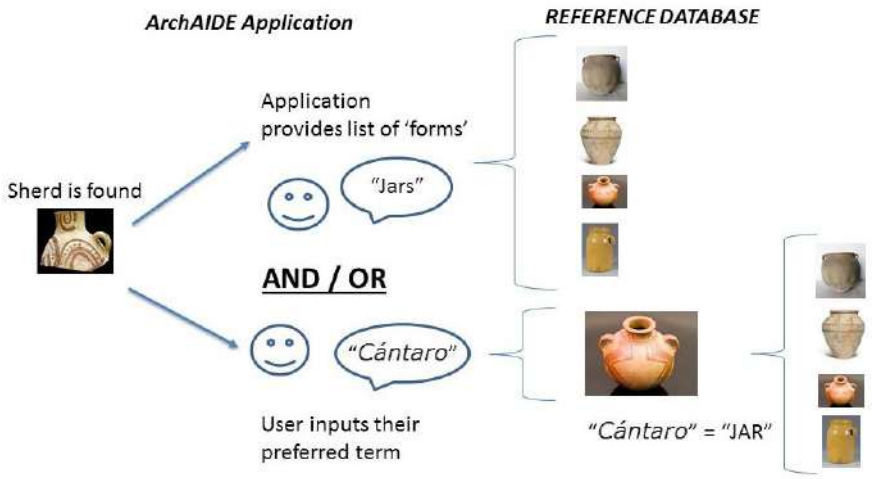

Figure 2: Simplified illustration of reconciling a specialised term ("Cántaro") to the ArchAIDE concept of "Jar".

3 The Linked Data approach

As a separate phase of work to supporting the reference database and ArchAIDE application, UoY have created a sustainable application to allow the searching and integration of the multilingual vocabularies by a new generation of web-based Application Programming Interfaces (APIs). This was done using a Linked Data approach: a style of publishing data on the world wide web that makes it easy to interlink, discover and consume data. This approach as part of a wider vision of a *Semantic Web* has been a growing for nearly 20 years, as web pioneers have pushed for a move away from siloed data and a web of documents to a fully integrated, machine-readable web of data (Berners-Lee, Hendler, and Lassila 2001). A variety of archaeological data resources already make use of Semantic Web principles and technologies, and it is notably a key feature of ongoing collaborative European initiatives to reconcile diverse datasets are exploring and incorporating these methodologies[4].

One of the best examples is Pelagios Commons[5]. Pelagios Commons is a research community working to develop infrastructures for the humanities based on ancient places. This includes the Recogito[6] annotation platform where researchers can manage and share their work with other resources connected to an ancient place, and the Peripleo[7] search service, which allows researchers to undertake map-based searches across a wide variety of interoperable datasets associated with ancient places. Pelagios is also an example of best practice because it uses Linked Open Data from many outside sources—making them interoperable and searchable through a sophisticated Web interface, while also making the data they hold freely available to other Semantic Web-based applications via APIs.

Considering the success of such projects, and the ongoing sustainability they provide in comparison to a single bespoke API that would need support, maintenance and possible migration after the end of the project, it was decided to follow the lead of ARIADNE and Pelagios and create a resource that was itself part of a wider system (the Semantic Web) that effectively acted as a API. As with Pelagios, by creating the resource that is intuitively understood by the new generation of applications and programs, the data will have a wider

[4] http://www.ariadne-infrastructure.eu/Community/Special-Interest-Groups/Linked-Data
[5] http://commons.pelagios.org/
[6] http://recogito.pelagios.org/
[7] http://pelagios.org/peripleo/map

re-use and lifespan than a single 'call and response' application. To achieve this the following tasks have been undertaken.

3.1 Semantics

As previously discussed, ArchAIDE has achieved this from the outset by using the ARIADNE tools to get disparate groups of archaeologists to agree to use not only the same neutral vocabularies (the AAT and Roman Amphora), but also the same appreciation of *how* to record using the SKOS to simply organise how accurately their terms match the spine.

3.2 Identifiers

Each of the 1338 ArchAIDE terms have then been assigned a unique uniform resource identifier (URI) by the ADS. This URI has followed the generic syntax, for example:

http://data.archaeologydataservice.ac.uk/archaide/dutch/type_form/platte

Refers to:

- The ADS Linked Data domain.
- The unique ArchAIDE repository.
- The language.
- The scheme (e.g. type_form).
- The concept identifier (e.g. platte)

On the advice of a Linked Data expert, URIs were created with Unicode characters only. Although modern web-browsers and web-based/desk-based applications *should* be able to resolve URIs with non-ASCII characters, this is not always a certainty. It was thus decided to remove them for practical reasons, although it should be noted that all native characters are retained in the original data.

3.3 Conversion to a Linked Data format

The complete dataset (with SKOS mappings and URIs) was converted to Comma Separated Values, with individual files for each language. This was then converted to the N-Triples format, a plain text serialisation format for RDF (Resource Description Framework) graphs, commonly used in the Linked Data approach. Graph data are organised into a three-part relationship of *subject*, *predicate*, and *object*, also referred to as a triple. Thus, graph databases are often referred to as triplestores. The conversion was acheived using an XSLT, openely available as part of the tools developed by the ARIADNE project. Within the XSLT, the user is able to select the native language of the file to be converted using ISO language identifiers (e.g. nl for dutch), which would ensure that any labels would be accompanied by this information, for example: "platte (nl)". The resulting N-Triples were also outputed using UTF-8 encoding to ensure that non-ascii characters were retained. Due to the detailed nature of the recording undertaken by project partners and collaborators, 2560 N-Triples were created. An example of this output is included below.

```
<http://data.archaeologydataservice.ac.uk/archaide/german/type_form/
trichter> <http://www.w3.org/2004/02/skos/core#broadMatch> <http://
vocab.getty.edu/aat/300200490> .
<http://data.archaeologydataservice.ac.uk/archaide/german/type_form/
trichter> <https://www.w3.org/2000/01/rdf-schema#label>
"trichter"@ger .
```

Figure 3: The N-Triples for the German concept of "trichter". The triples define the concept by stating the URI, and then how it related to another concept (in this case the AAT concept of a funnel), and that this ArchAIDE concept is described as "trichter", which the triple marks as a German word.

3.4 Creation of formal Linked Open Data

As noted above, the graph data created by ArchAIDE does not itself become Linked Open Data. To acheive this next step the N-Triples were uploaded to the ADS triplestore (Allegrograph 4.6.1). Upon upload they are stored in the RDF/XML format

```
<Description rdf:about="http://data.archaeologydataservice.ac.uk/
archaide/italian/sherd_type/fondo_ad_anello">
  <ns0:label xmlns:ns0="http://www.w3.org/2000/01/rdf-schema#"
  xml:lang="it">fondo ad anello</ns0:label>
  <ns0:broadMatch xmlns:ns0="http://www.w3.org/2004/02/skos/core#
  " rdf:resource="http://vocab.getty.edu/aat/300001656"/>
</Description>
```

Figure 4: RDF for the ArchAIDE concept of "fondo ad anello".

At this time of writing the ArchAIDE data has their own repository within the triplestore (to keep the data discretely packaged), which has been copied into the publicly accessible repository which the ADS use for dissemination of all their triples. The ADS use Pubby, an Open Source software (https://github.com/cygri/pubby) for linking this data to a SPARQL endpoint. SPARQL is a Declarative Query Language (like SQL) for performing data manipulation and data definition operations on Data represented in a Linked Data format. SPARQL endpoints can be accessed only by SPARQL client applications that use the SPARQL protocol. It cannot be accessed by the growing variety of Linked Data clients. Pubby is designed to provide a Linked Data interface for use by the types of APIs or REST services using RDF data sources. In this instance the ADS interface (http://data.archaeologydataservice.ac.uk/query/) offers a very simple web-based interface to allow users to generate their own SPARQL, which in turn generates an XML output. For ArchAIDE, a simple query to find all concepts mapped to "funnel" would be:

```
SELECT ?subject ?object ?predicate
WHERE
    { ?subject
      ?predicate
      <http://vocab.getty.edu/aat/300200490>
    }
```

Figure 5: SPARQL to find all ArchAIDE terms mapped to AAT concept for a funnel

```
▼<results>
  ▼<result>
    ▼<binding name="subject">
      ▼<uri>
          http://data.archaeologydataservice.ac.uk/archaide/german/type_form/trichter
        </uri>
      </binding>
    ▼<binding name="predicate">
        <uri>http://www.w3.org/2004/02/skos/core#closeMatch</uri>
      </binding>
    </result>
  ▼<result>
    ▼<binding name="subject">
      ▼<uri>
          http://data.archaeologydataservice.ac.uk/archaide/italian/type_form/imbuto
        </uri>
      </binding>
    ▼<binding name="predicate">
        <uri>http://www.w3.org/2004/02/skos/core#closeMatch</uri>
      </binding>
    </result>
  ▼<result>
    ▼<binding name="subject">
      ▼<uri>
          http://data.archaeologydataservice.ac.uk/archaide/portuguese/type_form/funis
        </uri>
      </binding>
    ▼<binding name="predicate">
        <uri>http://www.w3.org/2004/02/skos/core#exactMatch</uri>
      </binding>
    </result>
```


Figure 6: Example of XML outpur from SPARQL query of ArchAIDE Linked Data

At the time of writing it is more common for users of Linked Open Data to use their own desk-based software of web applications to cross-search multiple triplestores and SPARQL endpoints. Common applications such as *Gruff*, *Gephi*, *Blazegraph* or *LoDlive* allow a user to connect to simply point their application at whatever sources exist. Alternatively, it is now common for programmers to design bespoke services or widgets to interact with triplestores in a more easily understood fashion, effectively acting as an intermediary to provide a lookup of a term against a LOD vocabulary. For example, the ADS use their triplestore to align metadata terms from collections against external services from the Library of Congress and UK heritage agencies (see Figure 7). Such an approach could be developed by those wishing to align their terms against ArchAIDE, with a request generating both a human understandable string and a URI.

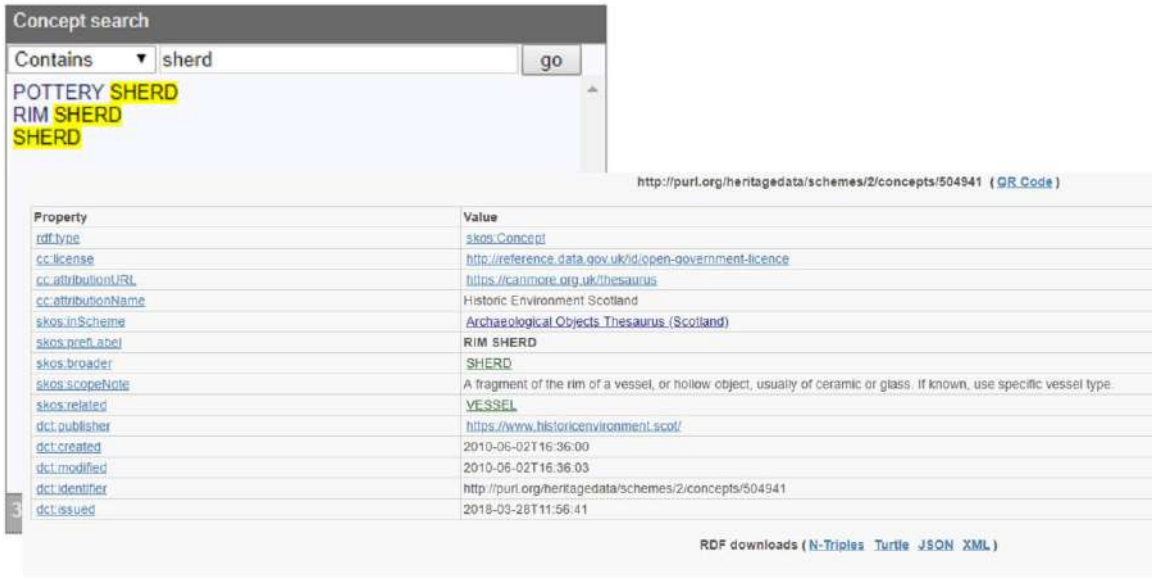

Figure 7: Example of an API returning Linked Data string and URI (from https://heritagedata.org/)

In addition to this simple functionality, there are now a growing number of powerful tools and APIs for visualising a single dataset, and the relationships between the concepts. For example (Figure 8), the user connects to the ADS triplestore (under the terms of a CC 0 licence) using a tool with a built-in API for working with LOD datasets (LodLive). Within this they see the node for the ArchAIDE vocabularies, with concepts defining what they are, what they contain, licence, creators and so on.

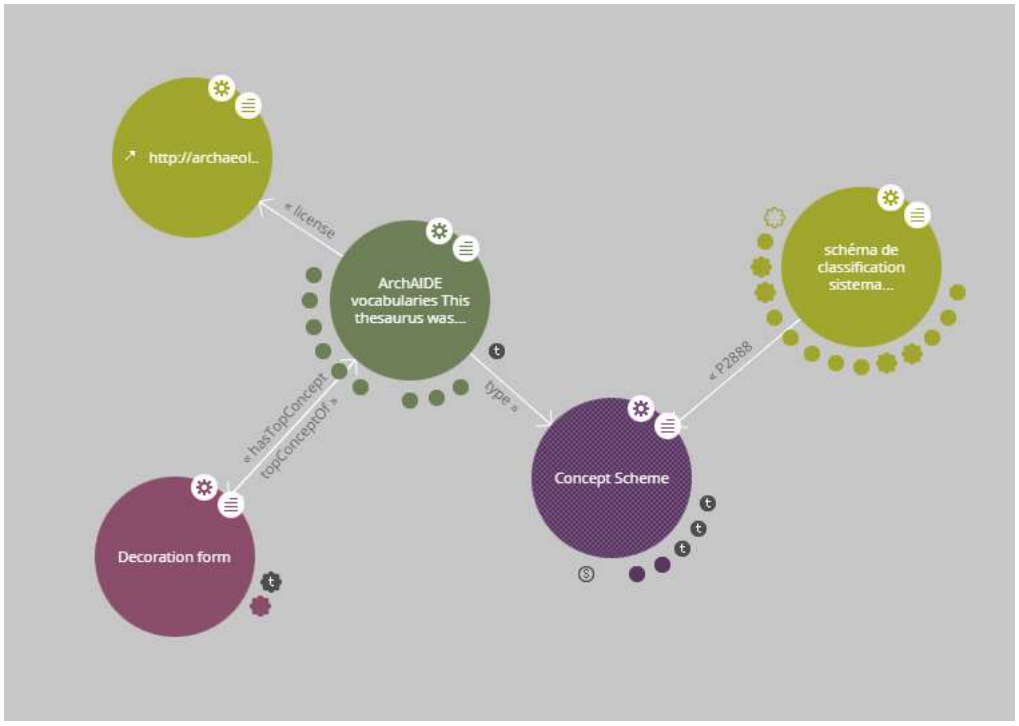

Figure 8: Simple view of the ArchAIDE Linked Open Data, using the LodLive API

The structure of the data allows a user to explore nodes graphically or semantically, and in addition allowing a user to view and query the relationships between whatever other data the application is either incorporating (via other SPARQL endpoints), or linked to by the ArchAIDE data. In the simple example presented below, the ArchAIDE node for form "Form type" is displayed with several nodes expanded to include their link to the AAT Linked Data concepts. The AAT data is effectively live, with the user able to see any other nodes connected to that concept, allowing them to find other data similarly aligned or related. Conversely, a user new to ArchAIDE would be able to see how their data relates to that generated by the project, and through querying establish similarities and new routes of inquiry.

It should be noted that this approach has the advantage of making the data instantly integrated without the need for downloading large amounts of data, and through the ever growing number of platforms designed to use Linked Open Data allowing it to be queried and understood by the non-specialist. This relatively simple approach effectively makes the whole of the Semantic Web – and by inclusion ArchAIDE - a single native API.

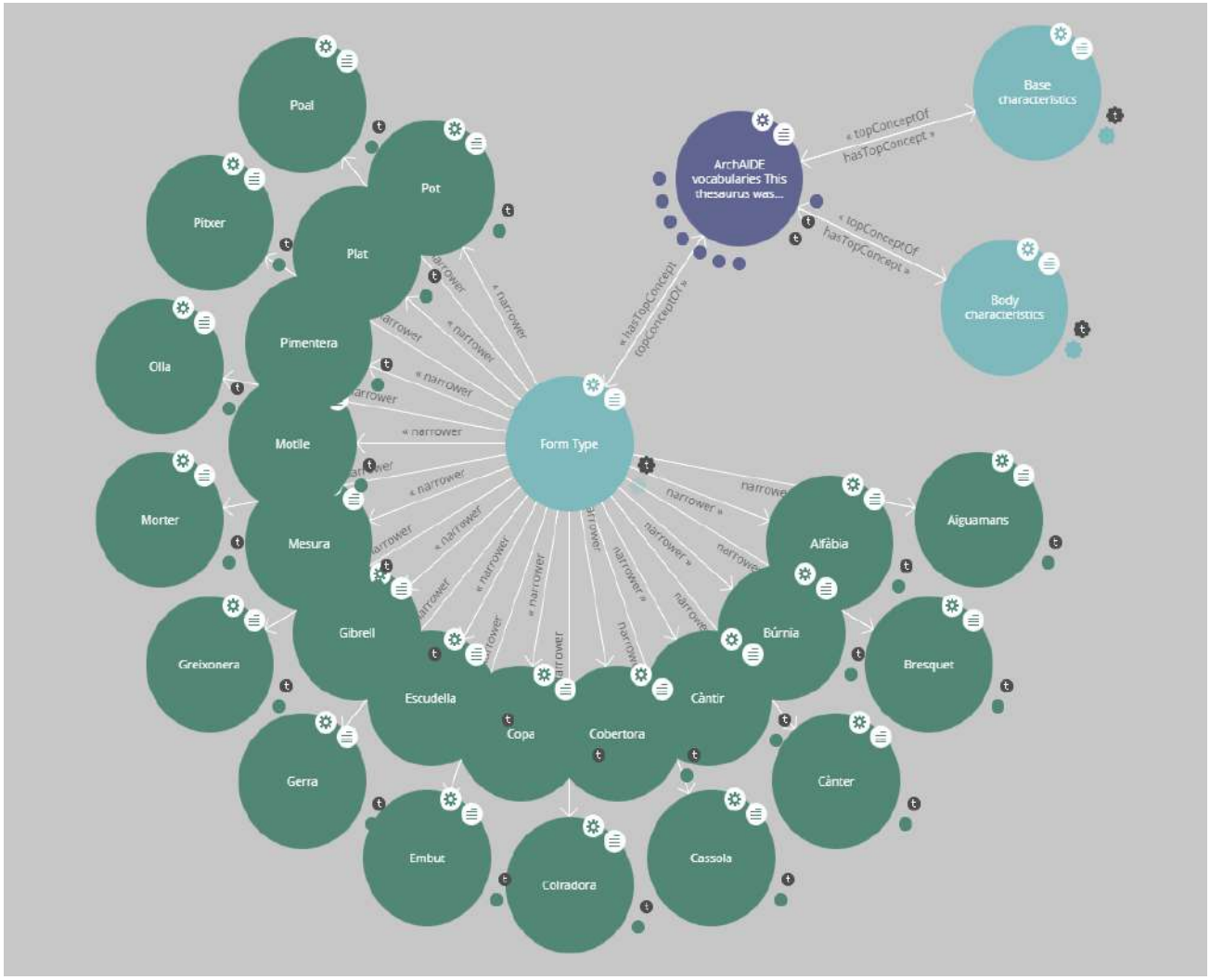

Figure 9: Example of a visualisation of part of the ArchAIDE vocabularies using the LodLive API

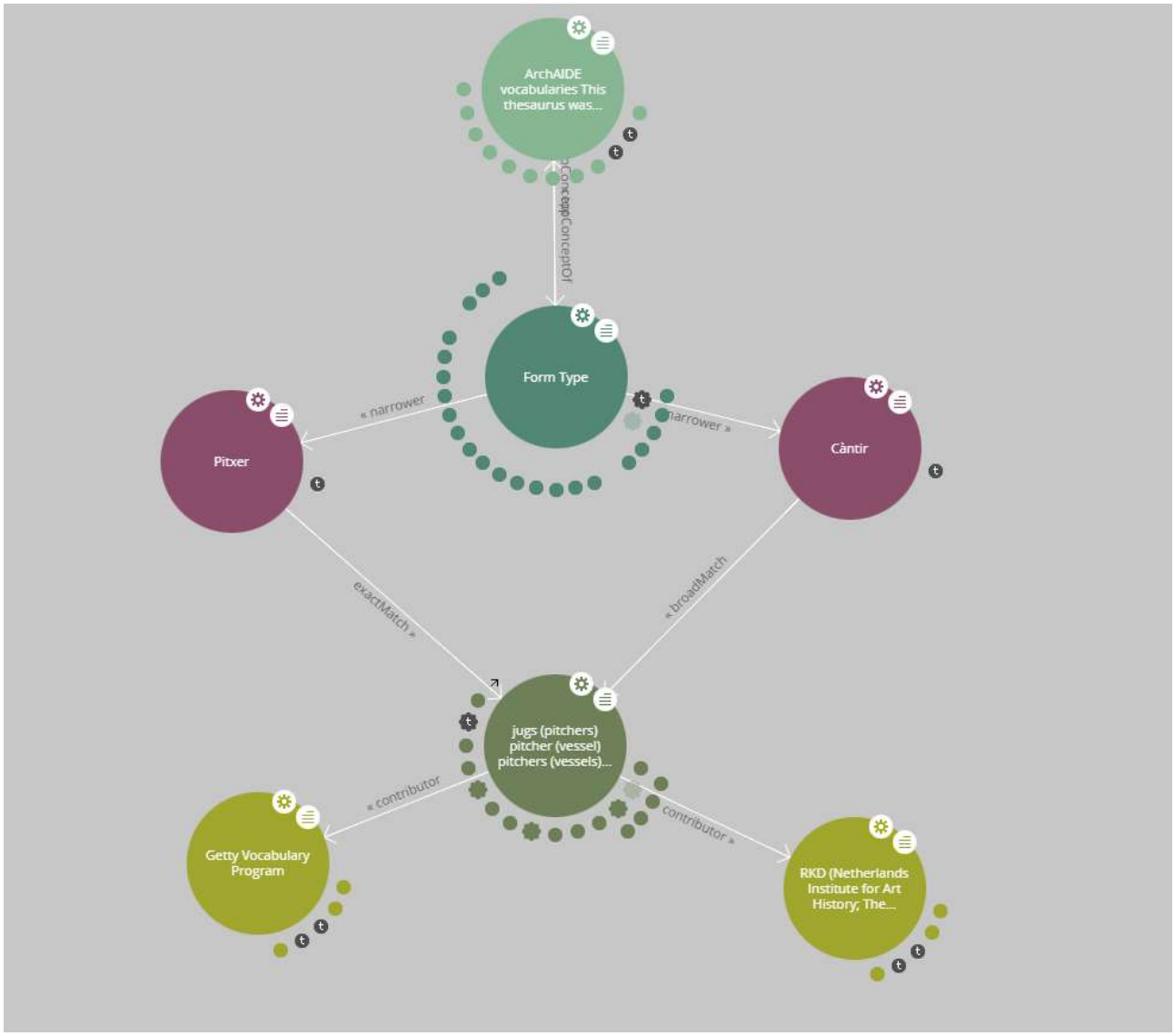

Figure10: Example of ArchAIDE concepts describing the form of a vessel, linked to live AAT data (the concept of a jug or pitcher), with furter nodes and branches from the AAT concept allowing further exploration. LodLive API.

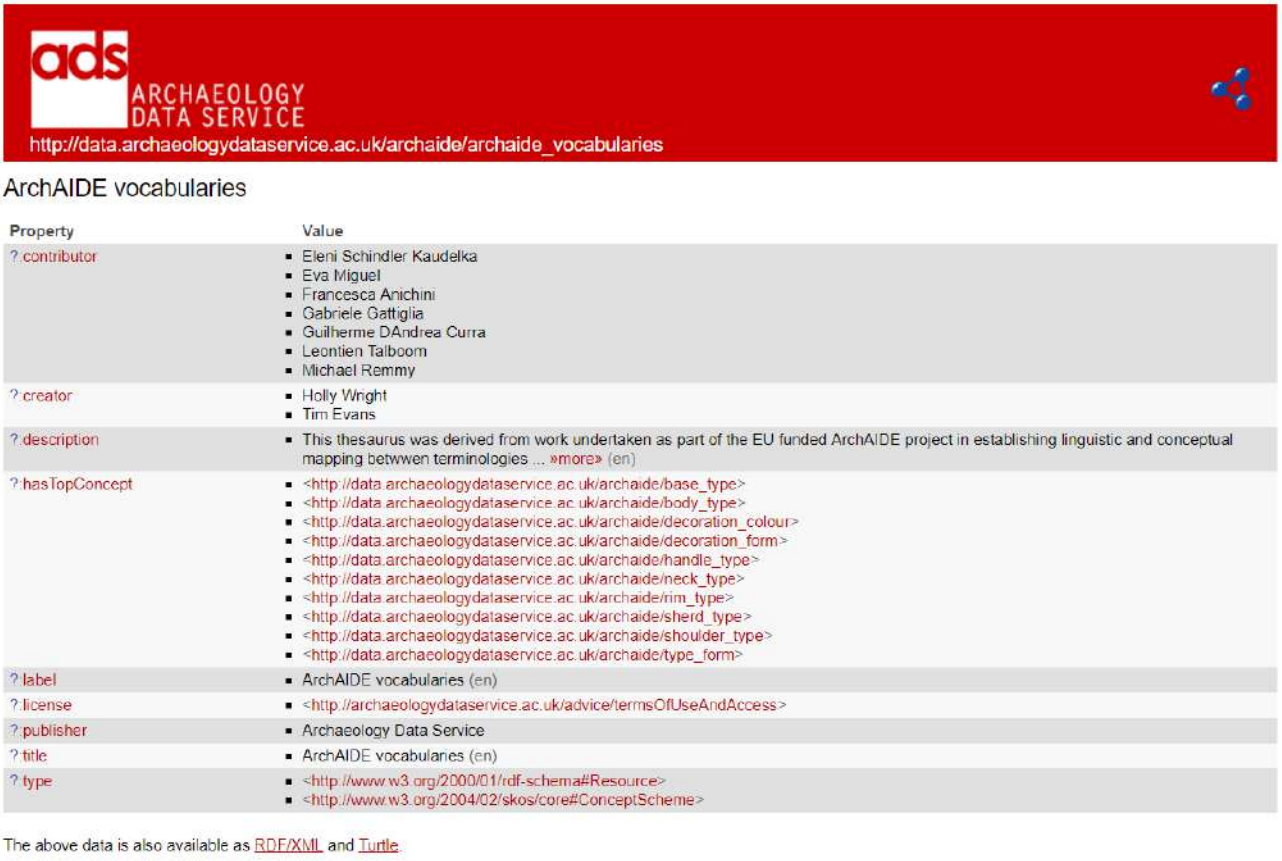

Figure 11: Overview of the ArchAIDE vocabularies, now available live as Linked Open Data from http://data.archaeologydataservice.ac.uk/page/archaide/archaide_vocabularies

4 Conclusion

Task 10.4 has taken a best practice approach. By using existing, standards and technologies to produce a compact and easily understood subset of the ArchAIDE data within a larger living application (the Semantic Web), rather than building a single product with a more limited lifespan. The data held in the ADS graph database (triplestore) now forms both a standalone, sustainable resource, which can be expanded and incorporated into future projects. Either as a small resource for alignment within future catalogues, or for cross-searching against the wider contents of the Semantic Web.

The ADS will continue to curate the triples as part of its ongoing work with developing their capacity for delivery of Linked Open Data. In addition, the original data is being deposited with the ADS in JSON and CSV format as part of the overall package for ensuring the key outputs of ArchAIDE are archived for perpetuity.

5 Bibliography

ARIADNE (2016). Towards a Web of Archaeological Linked Open Data. Available at http://www.ariadne-infrastructure.eu/content/download/8392/49194/version/2/file/ARIADNE_archaeological_LOD_study_10-2016.pdf (Accessed 24th January 2019).

Berners-Lee, T., Hendler, J. and and Lassila, O. (2001). The Semantic Web. *Scientific American* 284 (5): 29–37. https://doi.org/10.1038/scientificamerican0501-34.