ARIADNE

D15.1: Report on Thesauri and Taxonomies

Authors:

Douglas Tudhope, USW
Ceri Binding, USW

Ariadne is funded by the European Commission's 7th Framework Programme.

Version: 1.5 *(final*) **July 2016**

Authors: **Douglas Tudhope and Ceri Binding (USW)**

Contributing partners: **Holly Wright (ADS),**

Florence Laino (AIAC, L-P Archaeology),

Philipp Gerth, Francesco Mambrini (DAI),

Federico Nurra, Emmanuelle Bryas, Nouvel Blandine, Evelyne Sinigaglia (INRAP, FRANTIQ),

Paul Boon, Hella Hollander, Peter Brewer (KNAW-DANS, University of Arizona),

Evie Monaghan, Louise Kennedy, Anthony Corns (Discovery),

Sara Di Giorgio, Tiziana Scarselli (MIBACT-ICCU),

Esther Jansma (RCE),

with additional contributions from all partners

Quality review **Holly Wright (ADS - Archaeology Data Service)**

Table of contents

Executive Summary **4**

1 Introduction **5**

1.1 Controlled vocabularies 5

1.2 ARIADNE partner vocabularies 6

2 Mapping between thesauri **7**

2.1 Brief description of thesaurus mapping 7

2.2 Mappings in ARIADNE to support cross search 8

2.3 Getty Art and Architecture Thesaurus 9

2.4 Prototype experiment with AAT as hub vocabulary 9

2.5 Prototype experiment with AAT hierarchical expansion in Elasticsearch 11

3 Creating mappings for ARIADNE **20**

3.1 Overview of mappings 21

3.2 Description and reflections on mapping exercise 22

4 Mappings in the ARIADNE infrastructure **31**

4.1 Mapping enrichment process 31

4.2 Mappings within the ARIADNE portal 32

5 Conclusion **34**

6 References **35**

7 Appendix A **37**

8 Appendix B **41**

9 Appendix C **42**

Executive Summary

This deliverable reports on the work of ARIADNE WP15, Task 1: *SKOS thesauri and taxonomies*. This includes vocabularies, such as thesauri and term lists in different languages used by partners for subject indexing.

When searching free text with uncontrolled terms, significant differences can arise from trivial variations in search statements and from differing conceptualisations of a search by users. Different people use different words for the same concept, or employ slightly different concepts. As such, this was a key issue to be addressed within the ARIADNE project, and is a key focus of this report. The issues posed for interoperability and cross search by ARIADNE's multilingual collection of datasets and reports are discussed, along with the use of a controlled vocabulary to reduce ambiguity between terms by various features. The vocabularies most relevant for ARIADNE are also listed and described.

Mapping between vocabularies is a key aspect of semantic interoperability in heterogeneous environments. Mapping between native partner vocabularies can provide a useful mediation platform for ARIADNE cross search, particularly as subject metadata are in different languages. However the creation of links directly between the items from different vocabularies can quickly become unmanageable as the number of vocabularies increases. Therefore, a hub architecture was adopted, using an intermediate structure onto which the concepts from local vocabularies were mapped. The work on producing mappings is described, together with the incorporation of mappings in the ARIADNE infrastructure, and their use to date in the emerging ARIADNE Portal.

The Getty Art and Architecture Thesaurus (AAT) was chosen as an appropriate hub vocabulary, following a prototype mapping and retrieval exercise involving five ARIADNE vocabularies in three different languages. In another prototype experiment, the implementation of hierarchical expansion techniques was investigated using the Elasticsearch infrastructure adopted for the ARIADNE Portal.

A large scale pilot exercise with one ARIADNE partner was conducted, in order to allow for refinement of the methodology and mapping guidelines after reviewing the results. The first complete mapping exercise was successfully performed by ADS, using a custom linked data vocabulary matching tool developed for the ARIADNE project. Analysis of results from this pilot mapping informed an iteration of the mapping guidelines and the matching tool user interface. Following the review of the pilot mapping exercise, an additional, basic spreadsheet based utility was developed for recording mappings made manually in situations where the source vocabularies were not available as Linked Data. Mappings were conducted by the various content partners from their native vocabularies to the AAT. A summary of mappings with statistics on the SKOS match types employed by the various content partners is discussed. This shows that in almost all cases mappings were successfully established to the AAT. About half were exactMatch, with the other half mostly closeMatch and broadMatch. As expected only a small number were narrower matches – most partner vocabularies were considered to be reasonably congruent or were more specialized than the AAT. Reflections by partners on the mapping exercise are discussed.

The output from the partner mappings from their source vocabularies to the AAT is transformed to the required format for further processing by the relevant MoRe enrichment services used by the ARIADNE Registry. The enrichment process augments the data imported to the Registry with mapped AAT concepts. These derived subjects in turn make possible concept based search and browsing in the ARIADNE Portal. While the Portal is still evolving at the time of writing, a query on the Portal illustrates how the mappings make possible concept based search across subject metadata in different languages.

1 Introduction

This document is a deliverable (D15.1) of the project ARIADNE - Advanced Research Infrastructure for Archaeological Dataset Networking in Europe that has been funded under the European Community's Seventh Framework Programme. This deliverable reports on the work of ARIADNE WP15, Task 1: *SKOS thesauri and taxonomies*. This includes thesauri and term lists in different languages used by partners for subject indexing. Following on from the survey of vocabularies described in D3.3, those most relevant for ARIADNE are identified and augmented by a small number of additional vocabularies. The issues posed for interoperability and cross search by ARIADNE's multilingual collection of datasets and reports are discussed. Linking between vocabularies, following standard mapping relationships is considered the best practice approach towards multilingual functionality. The work on producing mappings is described, together with the incorporation of mappings in the ARIADNE infrastructure and their use to date in the emerging ARIADNE Portal.

1.1 Controlled vocabularies

Vocabularies are used for control of subject metadata. Other types of metadata can also benefit from vocabulary control, including place names, time periods and personal names. Vocabulary control aims to reduce the ambiguity of natural language (free text) when indexing and retrieving items while searching for information (Svenonius 2000; Tudhope et al. 2006).

Controlled vocabularies consist of terms, that is, words from natural language selected for retrieval purposes. A term can consist of one or more words. In a controlled vocabulary, such as a thesaurus, a term is used to represent a concept (which can have several terms associated with it).

Two features (synonyms and ambiguity) in natural language pose potential problems for retrieval:

a) Different terms (synonyms) can represent the same concept.

b) The same term (homographs) can represent different concepts. This can be a major problem in a mono-lingual system and becomes a significant problem in a multi-lingual collection, such as ARIADNE.

A controlled vocabulary can attempt to reduce ambiguity between terms by various features:

- Defining the scope of terms - how they are to be used within a particular vocabulary.
- Providing a set of synonyms (or effective synonyms for retrieval purposes) for each concept
- Restricting scope so that terms only have one meaning (and relate to only one concept).

Not all vocabularies provide all three features above. Some are just simple lists of authorised terms (term lists). Controlled vocabularies also provide vocabulary for Knowledge Organization Systems (KOS), which additionally structure their concepts via different types of semantic relationship (such as broader and narrower concepts).

Controlled vocabularies are sometimes contrasted with free text searching, assisted by statistical techniques in automatic indexing and ranking. These are not however exclusive options and different combinations of the two approaches are possible. Controlled vocabularies can be used to augment free text search.

When searching free text with uncontrolled terms, significant differences can arise from trivial variations in search statements and from differing conceptualisations of a search by searchers. Different people use different words for the same concept or employ slightly different concepts. This may not be a problem in casual search. However, in systematic research on a specialized topic, it is undesirable to miss relevant resources.

At the simplest level, a controlled list of terms ensures consistency in searching and indexing, helping to reduce problems arising from synonym and homograph mismatches. At a more complex level, the presentation of concepts in hierarchies and other semantic structures helps the indexer and searcher choose the most appropriate concept for their purposes. Browse-based user interfaces become possible.

A KOS can assist both precision (by allowing specific searching) and recall (by retrieving items described by related concepts or equivalent terms). It also provides potential pathways (for human and machine) that connect a searcher and indexer's choice of terminology. The more formal specification of logical semantic relationships within an ontology can assist applications where rules are specified about the relationships and logic-based inferencing is appropriate.

The information retrieval thesaurus is designed for retrieval purposes and has a restricted set of relationships (Tudhope and Binding 2016). These relationships are Equivalence (connects a concept to terms that act as effective synonyms), Hierarchical (broader / narrower concepts) and Associative (more loosely related, 'see also' concepts). These are defined by an international standard (the recently approved ISO 25964). The equivalence relationship connects a concept with a set of equivalent terms, treated as synonyms for the retrieval situations envisaged by the designers. Either mono or poly hierarchical structures may be employed. Thesauri are usually employed for descriptive indexing purposes and the corresponding search systems. Thesauri can also be used as a query expansion resource or as the basis for auto-complete suggestions in a search user interface, as in the ARIADNE Portal.

1.2 ARIADNE partner vocabularies

The vocabularies themselves vary from a small number of keywords in a picklist for a particular dataset to standard national vocabularies with a large number of concepts. ARIADNE Deliverable 3.1 (Initial report on standards and on the project registry) listed some archaeology-related subject vocabularies (terminology resources) and more details can be found there.

These included elements of the following vocabularies, considered particularly relevant for WP15 purposes:

- Art and Architecture Thesaurus (Getty Research Institute) – a thesaurus used for describing items of art, architecture and material culture
- Pactols Thesaurus (Frantiq) – six multilingual thesauri for describing items on antiquity and archaeology
- Thesaurus of Monument Types (FISH) – thesaurus of monument types by function
- Archaeological Objects Thesaurus (FISH) – thesaurus for recording of archaeological objects in Britain and Ireland over all archaeological periods
- Building Materials Thesaurus (FISH) – thesaurus of materials used in archaeological monuments
- PICO (MiBACT) – cultural heritage thesaurus covering Who/What/Where/When for use in Culturaitalia portal
- ICCD (MiBACT) – pictorial thesaurus for describing archaeological finds
- Referentienetwerk Erfgoed / ABR (RCE - Cultural Heritage Agency of the Netherlands) – contains the structured set of concepts of cultural heritage in the Netherlands
- ARKAS term list (ZRC-SASU) – a list of terms for the definition of archaeological sites in Slovenia
- FEDOLG-R term list (MNM-NÖK) – a list of terms for describing archaeological finds in Hungary
- Museums vocabularies (DAI) - a group of vocabularies for describing museum objects and concepts
- Archaeological Dictionary (DAI) – a multilingual dictionary for archaeological concepts under development

And additionally considered for WP15

- FASTI term list (AIAC) – set of terms for describing monument types in FASTI Online
- Irish Monuments Vocabulary (NMS) - for describing monument types in Ireland

- Archaeological term list (SND) – a set of terms for describing archaeological objects and monument types in Sweden drawing on national standards

Some of these vocabularies are available online or published as Linked Data in SKOS representation. This allows programmatic access to the vocabulary elements and the use of vocabularies as linking hubs in the web of data. This is further described in the forthcoming D15.2.

2 Mapping between thesauri

2.1 Brief description of thesaurus mapping

Mapping between vocabularies is a key aspect of semantic interoperability in heterogeneous environments, and is particularly important to multi-lingual collections (Tudhope et al. 2006). It can improve both recall (in different languages) and precision (false results may arise from literal string search).

Significant effort is required, however, for useful results; detailed mapping work at the concept level is necessary, requiring a combination of intellectual work and automated assistance. Zeng and Chan (2004) review different methodological approaches to mapping:

a) Derivation / Modeling of a specialised or simpler vocabulary from an existing vocabulary.

b) Translation / Adaptation from an existing vocabulary in a different language.

c) Satellite and Leaf Node Linking of a specialised thesaurus to a large, general thesaurus.

d) Direct Mapping between concepts in different controlled vocabularies, usually with an intellectual review.

e) Co-occurrence mapping between two vocabularies based on their mutual occurrences within the indexing of items within a collection. Co-occurrence mappings are considered looser than direct mapping made by experts.

f) Switching language used as an intermediary. It can be a new system created for the purpose or an existing system.

A switching language is one of the most frequently used approaches. This is the approach adopted by ARIADNE, as described below, where the switching language is described as a “hub” for the ARIADNE metadata connections. See also the discussion in the recent thesaurus standard, ISO25964-2:2013 section 6 “Structural models for mapping across vocabularies”.

There are also variants and combinations of these mapping approaches in practice. Effective mapping requires some degree of overlap and congruence of purpose in the vocabularies being mapped. Some prominent examples of mapping work are mentioned briefly. OCLC, providers of the Dewey Decimal Classification (DDC), developed various mappings between major vocabularies (both intellectual and statistical co-occurrence mappings) making them available as terminology web services (Vizine-Goetz et al. 2003). The OAI protocol was used to provide access to a vocabulary with mappings, via a browser to human users and through the OAI-PMH web service mechanisms to machines. Both direct mappings and co-occurrence mappings were provided, depending on the situation. The DDC was employed as a switching language in the Renardus FP5 project to support a cross-browsing service for a European academic subject gateway service (Koch et al. 2003).

More recently, the United Nation’s Food and Agriculture Organization (FAO) has devoted considerable resources to its AGROVOC thesaurus, which is a significant element of the VocBench collaborative vocabulary editing and publishing platform and the associated AIMS (Agricultural Information Management

Standards) portal. This has been expressed as Linked Data and there is an extensive mapping programme with (SKOS) mappings established for 13 vocabularies including LCSH (Library of Congress Subject Headings), GEMET (General Multilingual Environmental Thesaurus) and STW (Standard Thesaurus for Economics / Standard Thesaurus für Wirtschaft) (Caracciolo et al., 2013). Mapping services have been a longstanding focus of the German bilingual STW Thesaurus, a structured vocabulary for subject indexing and retrieval of economics literature. This is now based upon a Linked Data architecture Linked Data (Kempf and Neubert, 2016).

Tim Berners-Lee, creator of the World Wide Web and the concept of Linked Data has proposed a five star deployment scheme for grading Linked Open Data, which stresses linking to external Linked Open Data resources to achieve full potential. In the context described here, these links take the form of machine readable mappings to a common reference vocabulary.

★	Data made *openly* available on the web in any format
★★	As above, but in a machine readable structured data format (e.g. Excel)
★★★	As above, but in a non-proprietary structured data format (e.g. XML)
★★★★	As above, but using W3C open standards (e.g. URIs, RDF & SPARQL)
★★★★★	As above, and also **linking out** to other external LOD

Figure 1: The 5 star deployment scheme for Linked Open Data

Part 2 of the International Thesaurus Standard (ISO25964-II) aims to facilitate high quality information retrieval across networked resources indexed with different types of vocabularies. It explains how to set up mappings between the concepts in such vocabularies and includes a discussion of the impact of mapping on retrieval. This is an important consideration, particularly when no exact equivalent concept exists, and it is necessary to map to a broader or narrower concept, a partially overlapping concept, or to a (Boolean) combination of concepts. Section 14 of ISO25964-II discusses techniques for identifying candidate mappings.

Mapping between native partner vocabularies could provide a useful mediation platform for ARIADNE cross search, particularly as subject metadata are in different languages. However the creation of links directly between the items from different vocabularies can quickly become unmanageable as the number of vocabularies increases. Mapping between more than three vocabularies would be more efficient and scalable using the hub architecture (i.e. switching language), using an intermediate structure onto which the concepts from each local vocabulary may be mapped. A search on a concept originating from one vocabulary can then utilise this mediating structure to route through to concepts originating from other vocabularies, possibly expressed in other languages.

2.2 Mappings in ARIADNE to support cross search

For subject access, the ACDM *ArchaeologicalResource* class has two kinds of subject property. The property, *native-subject,* associates the resource with one or more items from a controlled vocabulary used by the data provider to index the data. However, there are a large number of partner vocabularies in several different languages. Cross search and semantic interoperability is rendered difficult, as there are no semantic links or mappings between the various local vocabularies. Standard ontologies for metadata schemas, such as the CIDOC-CRM, do not cover particular subject vocabularies but expect the ontology to be complemented with the terminology contained in the relevant subject vocabularies for an application domain. Spelling variations or different synonyms for the same concept can result in failure to find relevant results. This problem is exacerbated when subject metadata may be in different languages, which is clearly the case when providing an infrastructure for European archaeology. Not only may useful resources be

missed when searching in a different language from the subject metadata but there is also the problem of false results arising from homographs where the same term has different meanings in different languages. For example, "*vessel*" has different archaeological meanings in the English language, while "*coin*" is French for *corner*, "*boot*" is German for *boat* and "*monster*" is Dutch for *sample* (very different from the English language meanings of these words).

2.3 Getty Art and Architecture Thesaurus

The Getty Art and Architecture Thesaurus (AAT) is an influential and longstanding, multi-lingual thesaurus used world-wide, with over 40,000 concepts and over 350,000 terms (Harpring, 2016). The AAT has 7 facets (and 33 hierarchies as subdivisions): *Associated concepts, Physical attributes, Styles and periods, Agents, Activities, Materials, Objects* and optional facets for time and place. The AAT's scope is broader than archaeology, encompassing fine art, built works, decorative arts, other material culture, visual surrogates, archival materials, archaeology, and conservation. However it contains much useful, high level archaeological content, particularly in the Built Environment, Materials and Objects hierarchies.

The AAT has a faceted poly-hierarchical structure, containing generic concepts, with labels in multiple languages. It appears to have a good breadth of archaeological coverage to map local vocabularies to, together with clear scope notes defining the scope of usage for each concept. The AAT has recently been made available as Linked Open Data by the Getty Research Institute (Getty Research Institute, 2016b), which fits well with ARIADNE's strategy for semantic interoperability.

2.4 Prototype experiment with AAT as hub vocabulary

The AAT was chosen as an appropriate hub vocabulary, following a prototype mapping and retrieval exercise involving five ARIADNE vocabularies in three different languages. This is discussed in more detail in Binding and Tudhope (2015). Briefly, a small extract from the published AAT linked data was used as a hub, together with a set of intellectual mappings via consulting the Getty Vocabularies search facility (http://vocab.getty.edu/). For this exercise, the *skos:closeMatch* relationship was used rather than *skos:exactMatch.* Mappings were created manually (by USW) for the set of concepts employed in the pilot study. In some cases, partner vocabularies contained more specialised concepts than contained in the AAT. However, it was considered that the *skos:broadMatch* relationship should be appropriate in these situations, since the use case was cross-search in the ARIADNE Portal, rather than fine grained semantic processing.

In addition, the possibility of query expansion based upon the AAT's hierarchical structure (semantic expansion over the thesaurus hierarchical relationships) was noted. This would open up the possibility in retrieval of matching on terms associated with narrower concepts when querying at a more general level. This would have the potential of improving recall without loss of precision. As part of the pilot, a freely available desktop RDF search facility (SparqlGui, 2016) was employed to query the extract of AAT concepts, combined with the mappings produced for the pilot exercise. Using the query tool, a SPARQL 1.1 query on the AIAC concept fasti:cemetery (see Figure 2) returns results from five different vocabularies with terms in different languages via the AAT semantic structure (see Table 2). This search makes use of the mappings and also the hierarchical query expansion. The results from the pilot exercise were presented and discussed at the ARIADNE session in the Research Infrastructures on Cultural Heritage conference, co-organized in Rome by the ARIADNE project and the Italian Ministry of Culture (MIBAC) in November 2014 (and published in an accompanying ARIADNE booklet). It was decided that they held sufficient promise to proceed with a full mapping exercise, in order to deliver some degree of multilingual capability for the ARIADNE search system in the forthcoming Portal.

```
# SPARQL 1.1 to locate concepts related via AAT to FASTI "cemetery" concept
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX aat: <http://vocab.getty.edu/aat/>
PREFIX fasti: <http://fastionline.org/monumenttype/>
PREFIX iccd: <http://www.iccd.beniculturali.it/monuments/>
PREFIX tmt: <http://purl.org/heritagedata/schemes/eh_tmt2/concepts/>
PREFIX dans: <http://www.rnaproject.org/data/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dai: <http://archwort.dainst.org/thesaurus/de/vocab/?tema=>

SELECT DISTINCT ?concept ?label WHERE {
fasti:cemetery (skos:exactMatch | skos:broadMatch | skos:closeMatch) ?aatconcept
.
 ?aatdescendant gvp:broader+ ?aatconcept .
 {
  {?concept (skos:exactMatch | skos:broadMatch | skos:closeMatch)
?aatdescendant}
  UNION
  {?concept (skos:exactMatch | skos:broadMatch | skos:closeMatch) ?aatconcept}
}
 OPTIONAL {?concept skos:prefLabel ?label}
}
```

Figure 2: SPARQL 1.1 query on the semantic framework of AAT plus local vocabulary mappings.

Concept identifier	**Concept label**
iccd:catacomba	catacomba
tmt:91386	catacomb (funerary)
fasti:catacomb	Catacomb
iccd:colombario	colombario
fasti:columbarium	Columbarium
dai:3736	Kolumbarium
dans:6a7482e5-2fd5-48fb-baf4-66ad3d4ed95e	kerkhof
dai:1947	Gräberfeld
iccd:necropoli	necropoli
dai:2485	Nekropole
tmt:70053	cemetery
tmt:70053	necropolis

dans:be95a643-da30-40b9-b509-eadfb00610c4	christelijk/joodse begraafplaats
dans:b935f9a9-7456-4669-91d0-2e9c0ff7d664	vlakgrafveld
iccd:cimitero	cimitero
dans:abb41cf1-30dc-4d55-8c18-d599ebba1bc2	rijengrafveld

Table 1: Sample extract of the results from the query in Figure 2

2.5 Prototype experiment with AAT hierarchical expansion in Elasticsearch

Following an ARIADNE Joint Technical Meeting, it was decided to investigate further how to implement the hierarchical expansion techniques at scale in the context of the Elasticsearch infrastructure adopted for the ARIADNE Portal. Therefore a second prototype experiment with the AAT was conducted using the Elasticsearch platform.

Hierarchical semantic expansion makes use of broader generic ("IS-A") relationships between concepts in a hierarchically structured knowledge organization system, allowing a search on a particular subject indexing concept to also retrieve any items indexed using concepts that are positioned *below* that concept within the hierarchical structure.

- aat:300264092 *Objects Facet*
  - aat:300264551 *Furnishings and Equipment (hierarchy name)*
    - aat:300036743 *Weapons and Ammunition (hierarchy name)*
      - aat:300036926 *weapons*
        - aat:300036973 *edged weapons*
          - aat:300036982 *axes (weapons)*
            - aat:300036983 *battle axes*

Figure 3: full hierarchical ancestry of AAT concept ID 300036983 (battle axes)

Figure 3 illustrates the full hierarchical ancestry for an example AAT concept aat:300036983 (*battle axes)*. Using hierarchical semantic expansion a query on concept aat:300036926 (*weapons)* should therefore also retrieve items indexed as *edged weapons*, *axes (weapons)*, *battle axes* etc.

The prototype experiment demonstrated hierarchical semantic expansion using SPARQL against RDF resources. The Elasticsearch infrastructure used in ARIADNE has functionality referred to as *genre expansion* (Gormley and Tong, 2015) which should be able to achieve similar results to the SPARQL prototype described in section 2.4. The object of this exercise was therefore to again use the existing poly-hierarchical structure of the AAT, this time to produce configuration data in the format required to implement Elasticsearch genre expansion. We first extracted the AAT broader generic relationships by running the SPARQL query in Figure 4 against the Getty Vocabulary Program SPARQL endpoint (Getty Research Institute, 2016c).

```
# Extract the poly-hierarchical structure of the AAT
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX aat: <http://vocab.getty.edu/aat/>

CONSTRUCT { ?s gvp:broaderGeneric ?o }
WHERE { ?s skos:inScheme aat: ; gvp:broaderGeneric ?o }
```

Figure 4: SPARQL query to extract the poly-hierarchical structure of the AAT

The results of this query were downloaded in N-Triple RDF format to produce a local file containing 45,443 RDF triples. The configuration of Elasticsearch genre expansion requires the full ancestry chain of identifiers for each concept to be expressed as textual "rules" containing a comma separated list of identifiers, formatted as shown in Figure 5 (note the full AAT concept URIs have been shortened for illustration purposes):

```
aat:300264551 => aat:300264551, aat:300264092
aat:300036743 => aat:300036743, aat:300264551, aat:300264092
aat:300036926 => aat:300036926, aat:300036743, aat:300264551, aat:300264092
aat:300036973 => aat:300036973, aat:300036926, aat:300036743, aat:300264551, aat:300264092
 (etc.)
```

Figure 5: Elasticsearch genre expansion rules expressed

The extracted RDF data file resulting from the query in Figure 3 was imported to SparqlGui (SparqlGui, 2016) — a desktop tool for performing experimental SPARQL queries on RDF data. The SPARQL query shown in Figure 6 then retrieved the expansion rules data in the format shown in Figure 5, producing a 17 MB file, consisting of 41,866 lines of text.

```
# Produce the ancestry chains required for Elasticsearch genre expansion
PREFIX gvp: <http://vocab.getty.edu/ontology#>
SELECT (concat(str(?uri), " => ", str(?uri), ", ", group_concat(?broader; separator=", ")) AS ?ancestry)
WHERE {
 ?uri gvp:broaderGeneric+ ?broader .
}
GROUP BY ?uri
```

Figure 6: SPARQL query to produce the ancestry chains required for Elasticsearch genre expansion rules

Note: This process was split (i.e. extracting a subset of AAT data then querying the extract) only to alleviate potential performance issues, as this is a fairly demanding query. In practice it was found that the Getty SPARQL endpoint does actually support running the Figure 6 query directly - so in hindsight this would simplify the overall process.

The next stage was to incorporate the extracted and formatted data into Elasticsearch and test the genre expansion functionality. A local desktop copy of Elasticsearch was used in conjunction with the "Marvel – Sense" dashboard used for configuring and populating indexes and running experimental queries. The file of AAT genre expansion rules was copied to the */config* folder of the Elasticsearch installation, and was then referenced in a *synonym filter* for a *custom analyzer* when specifying the settings for initially creating an index, as illustrated in Figure 7.

```
#recreate the index specifying settings for the genre expansion analyzer
PUT /ariadnedata
{
  "settings": {
    "analysis": {
      "filter": {
        "aat_genre_expansion_filter": {
          "type": "synonym",
          "synonyms_path": "AAT-genre-expansion.txt"
        }
      },
      "analyzer": {
        "aat_genre_expansion_analyzer": {
          "tokenizer": "keyword",
          "filter": "aat_genre_expansion_filter"
        },
        "aat_uri_analyzer": {
          "tokenizer": "keyword"
        }
      }
    }
  }
}
```

Figure 7: specifying settings for the AAT genre expansion analyzer and synonym filter

A *mapping* was then created specifying how to handle values in the *dct:subject* subject indexing field (note: this was for demonstration and testing purposes; the actual naming of this field would have to be in accordance with the ARIADNE Elasticsearch index structure, as implemented). Note that genre expansion was configured during initial creation of the index and not at query time (see the *index_analyzer / search_analyser* configuration settings in Figure 8); otherwise the expansion would run in both broader and narrower directions - leading to incorrect and potentially misleading results. This means that (by default) genre expansion of AAT concept identifiers would always be enabled in search, though possibly some method could be devised to override it within the search parameters and the associated user interface, if that was deemed necessary.

```
# Add a mapping for the "dct:subject" field.
# Note we incorporate the genre expansion synonyms at index time,
# not at search time otherwise it could lead to some odd results
PUT /ariadnedata/item/_mapping
{
  "properties":
  {
    "dct:subject":
    {
      "type": "string",
      "index_analyzer": "aat_genre_expansion_analyzer",
      "search_analyzer": "aat_uri_analyzer"
    }
  }
}
```

Figure 8: Adding a mapping specifying how to handle the subject indexing field

Some sample data items indexed using the *dct:subject* field (with various AAT URI identifiers from the example in Figure 2) were created for testing purposes and added to the experimental index, using the commands shown in Figure 9.

```
# Add some sample records, dct:subject indexed with AAT identifiers

# item indexed with "battle axes"
PUT /ariadnedata/item/10
{ "dct:subject": "http://vocab.getty.edu/aat/300036983" }

# item indexed with "axes (weapons)"
PUT /ariadnedata/item/11
{ "dct:subject": "http://vocab.getty.edu/aat/300036982" }

# item indexed with "edged weapons"
PUT /ariadnedata/item/12
{ "dct:subject": "http://vocab.getty.edu/aat/300036973" }

# item indexed with "weapons"
PUT /ariadnedata/item/13
{ "dct:subject": "http://vocab.getty.edu/aat/300036926" }

# item indexed with "weapons" (another)
PUT /ariadnedata/item/14
{ "dct:subject":"http://vocab.getty.edu/aat/300036926" }
```

Figure 9: Adding some sample items to the index for testing

Testing the item index

Testing was achieved by querying for the items indexed using specific AAT concept URIs. The example query shown in Figure 10 is searching for items indexed using a *dct:subject* field value of aat:300036926 *(weapons).* A number of queries were run using different *dct:subject* values.

```
# perform search incorporating AAT genre expansion
GET /ariadnedata/item/_search
{
  "query":{
  "match":{
    "dct:subject": "http://vocab.getty.edu/aat/300036926"
  }}
}
```

Figure 10: Testing the genre expansion by querying for items indexed using specific AAT concepts

The results shown in Table 3 illustrate the effects of genre expansion. A search on aat:300036983 *(battle axes)* retrieved only the single item indexed using that concept identifier, but a search on aat:300036973 *(edged weapons)* retrieved items indexed using that concept AND items indexed using any of the descendant concepts, in accordance with the AAT hierarchical structure example in Figure 3.

dct:subject **search on AAT concept identifier**	**ID(s) of the items retrieved**
aat:300036983 *battle axes*	10
aat:300036982 *axes (weapons)*	10 & 11
aat:300036973 *edged weapons*	10, 11 & 12
aat:300036926 *weapons*	10, 11, 12, 13 & 14

Table 2: Results of searching for specific dct:subject values

Use of vocabulary resources

The previous documentation discusses genre expansion directly applied to registry items. A similar approach can therefore be taken to indexing and expanding the ARIADNE vocabulary concept resources themselves. Using the same test index as previously (*ariadnedata)* and the same analysers, some sample vocabulary concept resources were indexed. First, a new mapping specifying how to handle the concept metadata fields was added (Figure 11).

```
# create mappings for concept metadata fields
PUT /ariadnedata/concept/_mapping
{
  "properties":
  {
    "dct:identifier":
    {
      "type": "string",
      "index_analyzer": "aat_genre_expansion_analyzer",
      "search_analyzer": "aat_uri_analyzer"
    },
    "skos:inScheme":
    {
      "type": "string",
      "search_analyzer": "aat_uri_analyzer"
    },
    "skos:prefLabel":
    {
      "type": "string"
    }
  }
}
```

Figure 11: Adding a mapping specifying how to handle the concept metadata fields

Some sample concept metadata was then created for testing purposes and manually added to the experimental index, using the commands shown in Figure 12. A bulk import process would have to be adopted for importing the actual Getty AAT concept metadata, as it is a large dataset.

```
#  add some vocabulary concept metadata
PUT /ariadnedata/concept/http%3A%2F%2Fvocab.getty.edu%2Faat%2F300036743
{
  "dct:identifier": "http://vocab.getty.edu/aat/300036743",
  "skos:inScheme": "http://vocab.getty.edu/aat/",
  "skos:prefLabel": "Weapons and Ammunition"
}
PUT /ariadnedata/concept/http%3A%2F%2Fvocab.getty.edu%2Faat%2F300036926
{
  "dct:identifier": "http://vocab.getty.edu/aat/300036926",
  "skos:inScheme": "http://vocab.getty.edu/aat/",
  "skos:prefLabel": "weapons"
}
PUT /ariadnedata/concept/http%3A%2F%2Fvocab.getty.edu%2Faat%2F300036973
{
  "dct:identifier": "http://vocab.getty.edu/aat/300036973",
  "skos:inScheme": "http://vocab.getty.edu/aat/",
  "skos:prefLabel": "edged weapons"
}
PUT /ariadnedata/concept/http%3A%2F%2Fvocab.getty.edu%2Faat%2F300036982
{
  "dct:identifier": "http://vocab.getty.edu/aat/300036982",
  "skos:inScheme": "http://vocab.getty.edu/aat/",
  "skos:prefLabel": "axes (weapons)"
}
PUT /ariadnedata/concept/http%3A%2F%2Fvocab.getty.edu%2Faat%2F300036983
{
  "dct:identifier": "http://vocab.getty.edu/aat/300036983",
  "skos:inScheme": "http://vocab.getty.edu/aat/",
  "skos:prefLabel": "battle axes"
}
```

Figure 12: Inserting the metadata for some example concepts

Testing the concept index

As the genre expansion analyzer had already been previously created and configured, we could now perform semantic genre expansion queries directly on the vocabulary concept resources themselves. Note how the query shown in Figure 13 is quite similar to that shown in Figure 10, but this time we are searching the resources under */ariadnedata/concept* for a specified *dct:identifier* value – which in this case is the AAT concept representing "weapons" (see Figure 3).

```
# perform vocabulary search incorporating AAT genre expansion
GET /ariadnedata/concept/_search
{
  "query":{
  "match":{
    "dct:identifier": "http://vocab.getty.edu/aat/300036926"
  }}
}
```

Figure 13: query to perform genre expansion on AAT concept 300036926 ("weapons")

The results of this query are shown in Figure 14. The results include the specified concept AND all hierarchically descendant concepts in accordance with the AAT hierarchical structure (from Figure 3).

```
"hits": {
  "total": 4,
  "max_score": 1.5108256,
  "hits": [
    {
      "_index": "ariadnedata",
      "_type": "concept",
      "_id": "http://vocab.getty.edu/aat/300036926",
      "_score": 1.5108256,
      "_source": {
        "dct:identifier": "http://vocab.getty.edu/aat/300036926",
        "skos:inScheme": "http://vocab.getty.edu/aat/",
        "skos:prefLabel": "weapons"
      }
    },
    {
      "_index": "ariadnedata",
      "_type": "concept",
      "_id": "http://vocab.getty.edu/aat/300036983",
      "_score": 1.5108256,
      "_source": {
        "dct:identifier": "http://vocab.getty.edu/aat/300036983",
        "skos:inScheme": "http://vocab.getty.edu/aat/",
        "skos:prefLabel": "battle axes"
      }
    },
    {
      "_index": "ariadnedata",
      "_type": "concept",
      "_id": "http://vocab.getty.edu/aat/300036982",
      "_score": 1.4054651,
      "_source": {
        "dct:identifier": "http://vocab.getty.edu/aat/300036982",
        "skos:inScheme": "http://vocab.getty.edu/aat/",
        "skos:prefLabel": "axes (weapons)"
      }
    },
    {
      "_index": "ariadnedata",
      "_type": "concept",
      "_id": "http://vocab.getty.edu/aat/300036973",
      "_score": 1,
      "_source": {
        "dct:identifier": "http://vocab.getty.edu/aat/300036973",
        "skos:inScheme": "http://vocab.getty.edu/aat/",
        "skos:prefLabel": "edged weapons"
      }
    }
  ]
}
```


Figure 14: results of Elasticsearch genre expansion query on AAT concept 300036926 ("weapons")

This demonstrates one possible method of implementing the hierarchical semantic expansion of AAT concepts in Elasticsearch. The technique can improve the recall measure of query results without sacrificing precision. The full AAT "expansion rules" data file as produced could be reused in other projects, and the same approach can be easily adapted to other hierarchically structured knowledge organization resources, such as the Getty Thesaurus of Geographic Names.

The two prototype experiments also show the potential of working with the URI identifiers of AAT concepts rather than the ambiguous strings of term labels. Using the URI identifier for the concept avoids the problem

of ambiguity, common in multilingual datasets, of terms that are homographs in different languages. Working at the concept level also makes possible hierarchical semantic expansion, making use of the broader generic ("IS-A") relationships between concepts in a hierarchically structured knowledge organization system, such as the AAT. Thus a search expressed at a general level can (if desired) return results indexed at a more specific level. For example, a search on *settlements* might also return *monastic centres*.

3 Creating mappings for ARIADNE

Following the prototype experiment, the next step was to produce the mappings from the subject vocabularies employed to index the various datasets selected for the ARIADNE Catalogue. It was decided to proceed with a large scale pilot exercise with one ARIADNE partner, in order to allow for refinement of the methodology and mapping guidelines after reviewing the results.

The first complete mapping exercise was performed by ADS on SKOSified national heritage vocabularies for England, Scotland and Wales, using a custom linked data vocabulary matching tool developed by USW for the ARIADNE project. For details of the mapping exercise and the tool, see Binding and Tudhope (2015) and the forthcoming D15.3 will discuss tools in more detail. Analysis of results from this pilot mapping informed an iteration of the mapping guidelines and the matching tool user interface. For example, it was decided that mapping to AAT *Guide Terms* (not normally used for indexing) was undesirable for ARIADNE purposes. Also, multiple mappings from the same source concept were only considered useful in certain circumstances. A complete set of mappings was then produced for the subject metadata used in the ADS data imported by the ARIADNE Registry. Examples of mappings from the ADS mapping exercise are shown in Table 4. These were reviewed by a senior archaeologist and the final mappings (after minor fine tuning) were communicated to the ATHENA DCU Registry team as RDF/JSON statements (see section 4). This exercise, together with the guidelines, was reviewed by the USW team. Revisions to the mapping guidelines included recommendations on the appropriate SKOS mapping relationship to employ in different contexts, and when appropriate, to specify more than one mapping for a given concept.

Source concept	**matchURI**	**Target concept**
DITCHED ENCLOSURE http://purl.org/heritagedata/schemes/eh_tmt2/concepts/70361	skos:broadMatch	*agricultural settlements* http://vocab.getty.edu/aat/300008420
CROFT http://purl.org/heritagedata/schemes/eh_tmt2/concepts/68617	skos:closeMatch	*small holdings* http://vocab.getty.edu/aat/300000211

Table 3: Examples from the ADS mapping exercise

The revised guidelines were employed in the mappings of vocabularies from the other partners (and see Appendix C). Following the review of the pilot mapping exercise, an additional, basic spreadsheet based utility was developed for recording mappings made manually in situations where the source vocabularies were not available as Linked Data (see D15.3, forthcoming).

3.1 Overview of mappings

Partner	Vocabularies mapped to AAT	Match type						Totals
		No match	skos:exactMatch	skos:closeMatch	skos:broadMatch	skos:narrowMatch	skos:relatedMatch	
ADS	Archaeological Objects (subset)	0	197	96	118	0	0	**411**
ADS	Building Materials (subset)	0	8	4	0	0	0	**12**
ADS	Monument Types (subset)	0	141	107	141	0	1	**390**
ADS	Components Thesaurus (subset)	0	7	1	1	0	0	**9**
ADS	Maritime Craft (subset)	0	13	8	3	0	0	**24**
DAI	ARACHNE - books	0	13	4	0	0	0	**17**
DAI	ARACHNE - collections	0	8	2	1	0	0	**11**
DAI	ARACHNE - inscriptions	0	18	1	0	0	0	**19**
DAI	ARACHNE - buildings and structures	0	81	37	44	14	0	**176**
DAI	ARACHNE - multi-part monument	0	51	35	22	0	0	**108**
DAI	ARACHNE - topographic objects	0	46	7	2	0	0	**55**
DANS	DCCD vocabulary	0	245	9	82	0	0	**336**
DANS	EASY - Complextypen	3	3	59	34	0	15	**114**
Discovery	Irish Monument Types	0	168	69	249	0	0	**486**
IACA	FASTI Monument Types	2	23	80	24	0	0	**129**
ICCU	ICCD RA and PICO thesauri (subset)	0	642	94	310	258	0	**1304**
INRAP	PACTOLS thesaurus (subset)	0	1161	121	346	6	0	**1634**
MNM-NOK	site types	0	0	34	7	0	0	**41**
OEAW	DFMROE DB	0	4	0	0	3	0	**7**
OEAW	Franzhausen Kokoern DB	0	5	2	2	1	0	**10**
OEAW	UK Material Pool DB	0	7	0	4	5	0	**16**
OEAW	UK Thunau DB	0	3	1	0	0	0	**4**
SND	combined terms list	5	71	156	144	11	0	**387**
ZRC-SAZU	ZBIVA vocabulary	0	0	25	5	0	0	**30**
ZRC-SAZU	ARKAS vocabulary	0	0	76	17	0	0	**93**
ADS	**5**	**0**	**366**	**216**	**263**	**0**	**1**	**846**
DAI	**6**	**0**	**217**	**86**	**69**	**14**	**0**	**386**
DANS	**2**	**3**	**248**	**68**	**116**	**0**	**15**	**450**
Discovery	**1**	**0**	**168**	**69**	**249**	**0**	**0**	**486**
IACA	**1**	**2**	**23**	**80**	**24**	**0**	**0**	**129**
ICCU	**1**	**0**	**642**	**94**	**310**	**258**	**0**	**1304**
INRAP	**1**	**0**	**1161**	**121**	**346**	**6**	**0**	**1634**
MNM-NOK	**1**	**0**	**0**	**34**	**7**	**0**	**0**	**41**
OEAW	**4**	**0**	**19**	**3**	**6**	**9**	**0**	**37**
SND	**1**	**5**	**71**	**156**	**144**	**11**	**0**	**387**
ZRC-SAZU	**2**	**0**	**0**	**101**	**22**	**0**	**0**	**123**
Totals:	**25**	**10**	**2915**	**1028**	**1556**	**298**	**16**	**5823**
	Proportion:	**0.17%**	**50.06%**	**17.65%**	**26.72%**	**5.12%**	**0.27%**	**100%**

Table 4: Summary of mappings with statistics on match type (as of June 2016). Note – for ADS, ICCU and INRAP the mappings are based on a subset of the source thesaurus terms

Table 5 gives a summary of mappings completed at time of writing. The vocabularies are described in Section 1 and reflections by partners on the mapping exercise are given in Section 3.2.

From the overall statistics we can see that in almost all cases mappings were established to the AAT. Some 50% were *skos:exactMatch*, with 18% *skos:closeMatch* and 27% *skos:broadMatch*. As expected only a small number (5%) were narrower matches – most partner vocabularies were considered to be reasonably congruent or were more specialized than the AAT. However, there were a few exceptions where the AAT was more specialized. The mapping guidelines steered partners away from using *skos:relatedMatch* but that was found useful by DANS in a very small number of cases when it was considered appropriate to make more than one mapping, perhaps additionally to a related activity (see discussion in Section 3.2). Considering the mapping choices made by individual partners, we can see some difference in the mapping relationships chosen, e.g. a higher proportion of *skos:closeMatch* in mappings for ADS, DANS (EASY), MNM-NOK, SND, ZRC-SASU. This could variously reflect the nature of the vocabularies involved or the style of the person doing the mapping (e.g. when thought appropriate to assert an exact match). Another factor could be the amount of contextual information available in the form of scope notes etc. – if no information other than the preferred term label is available, then that might be considered a reason to assert a *skos:closeMatch* relationship rather than *skos:exactMatch*.

3.2 Description and reflections on mapping exercise

A selection of example reflections on their respective mapping exercises are given below by ARIADNE content provider partners.

ADS

ADS carried out initial evaluations regarding the suitability of the AAT to describe archaeological subjects, to determine whether it was an appropriate thesaurus for the multi-lingual mappings necessary for ARIADNE. Results were very positive during tests using the UK national vocabularies, and it was felt that the AAT was sufficient, although there were some odd areas of extreme detail (i.e. knives) and other areas where there was nothing directly comparable (i.e. human or animal remains). However, there was felt to be a sufficient range of SKOS mapping types available to handle these situations. There was also understood to be a certain amount of subjectivity in mapping choices, even for domain experts, and it was deemed a good practice future idea to have mappings done by multiple people (essentially creating an authoritative mapping by attribution, or “expert crowdsourcing”).

ADS also carried out the initial mapping exercise to test the matching tool developed by USW and create the mapping to the UK thesauri, and provide an exemplar for other partners. It was determined to be impractical to do complete mappings of every term in the UK thesauri, so all the distinct terms in use by the ADS were mapped instead. This still represented around 1000 terms to be mapped, the majority of which were derived from the English Monument and Type thesaurus. ADS was able to achieve comprehensive coverage of their distinct terms mapped to the AAT. Inevitably there were some broad matches in cases where the granularity of the AAT does not match the more fine-grained detail of the archaeology domain, but it was confirmed that the AAT does give sufficient breadth and depth of domain coverage for some very good matches on all the terms used, despite being quite diverse – including maritime craft, organic and inorganic materials, objects and monument types. The mapping exercise also clearly showed that purely automated string matching would indeed have been insufficient, and that expert input was necessary (e.g. Alan Williams Turret => field fortifications, lynchet => agricultural land, etc.)

AIAC

Some 130 mappings from the Fasti monument thesaurus to the AAT were provided by USW to Fasti. These were imported via a script, and an interface developed to edit them on the Fasti Admin page. Several more mappings were added with this interface and some minor corrections were made to the mappings from USW. These are now available on the Fasti website as part of the published Fasti concepts via http://www.fastionline.org/concept/attributetype/monument.

By providing a URI it is possible to refer to these thesaurus items in a controlled way, with an explicit reference to the AAT and to the translations to many languages that are available in Fasti. The concepts use both an English 'human readable' URI and a numeric URI using identifiers from the Fasti database, to create language independent identifiers in a manner reflecting the AAT URIs. These mappings were used to make sure that the terms in the OAI-PMH XML match the terms used in the ARIADNE Portal for ingestion. It is planned that before the end of the ARIADNE project, these mappings will be made available to the public throughout the Fasti interface so that the concepts are usefully defined. The process of issuing URIs for the concepts used in Fasti will add meaning to the presented data, by linking to existing enriched thesauri.

DAI

The IT infrastructure of the German Archaeological Institute (DAI) contains many different subject specific information systems, e.g. for excavations and surveys (iDAI.field), objects and for publication of data (Arachne), bibliographical information (Zenon) and digitized books (iDAI.bookbrowser). While the places are already centrally structured within the iDAI.gazetteer (http://gazetteer.dainst.org/) and all information systems refer to the gazetteer, each of the systems has their own vocabulary for describing the stored objects. At the moment work is ongoing to harmonize the different DAI thesauri to one common standard in iDAI.vocab (http://archwort.dainst.org/).

For the mapping activities in ARIADNE, the relevant vocabulary categories of the object database Arachne were chosen, as Arachne contains, in contrast to iDAI.field, with more than 3.6 million datasets, a large amount of which is openly available. The vocabulary of the following categories was mapped to Getty AAT:

- **Topographie** (eng. *Topography*, http://arachne.dainst.org/category/?c=topographie): Arachne's most granular object unit, which is the superior context for all related classes, which includes landscapes, sites, and part of sites. It is mapped to the ACDM class "sites and monuments" and contains 55 values mapped to Getty AAT from two different value lists.
- **Bauwerke** (eng. *Buildings*): This class comprises buildings and monuments, which forms a context for single object records and could be part of a larger site. It is mapped to the ACDM class "*sites and monuments*" and contains 176 values mapped to the Getty AAT from four different value lists.
- **Mehrteilige Denkmäler** (eng. *Multipart monuments*): All kinds of groups, which are not buildings or topographic units, are subsumed into multipart monuments, e.g. groups of statues, graveyards, hoards. This class is mapped to the ACDM classes "*sites and monuments*" or "*burials*", depending on the object type, and contains 108 values mapped to the Getty AAT from six different value lists.
- **Sammlungen** (eng. *Collections*): Private and museum collections belong to this class. It is mapped to the ACDM class "*diverse*" and contains 11 values mapped to the Getty AAT from two different value lists.
- **Bücher** (eng. *Books*): Digital reproduction, characterization and context of classical study prints from the 16th to 19th century. It is mapped to the ACDM class "*textual documents*" and contains 17 values mapped to the Getty AAT from three different value lists.
- **Inschriften** (eng. *Inscriptions*): This class contains inscriptions and epigraphs depicted on objects. It is mapped to the ACDM class "*textual documents*" and contains 19 values mapped to Getty AAT from one value list.

DANS

DANS translated the ABR terms into English as a first step towards mapping the DANS EASY Complextypen to the AAT. As DANS discovered, translating a term and understanding the concept it stands for, go hand in hand. Associated with this work, DANS translated the terms not only to English, but also to German, French, Italian, Spanish and Czech, with the help of colleagues and volunteers, many of whom had no archaeological background. The process of trying to find translations in different languages helped in better understanding and "pinning down" the concept and thus finding an optimal AAT mapping for it. Besides the websites of the AAT (Getty and the Dutch RKD) and the site of the ABRplus (RCE) DANS also used Wikipedia. Even if a term was not a Wikipedia lemma, DANS could sometimes find it mentioned in a description of an evidently related lemma. Most of the matches found were either *skos:closeMatch* or *skos:broadMatch.* Finding the mappings was far from easy however. Firstly it was difficult to understand the archaeological concept behind the ABR term when only the term and the hierarchical context were available (without scope notes). Secondly, it was sometimes difficult to understand AAT concepts when they reflected a perspective not specifically archaeological. For example, in some cases, the DANS (EASY) ABR essentially captured the notion of a place where an activity occurred and this had no exact match in the AAT. In these situations, a *skos:broadMatch* was sometimes generated plus an additional *skos:relatedMatch* to a corresponding activity, material or object. Future work will consider steps for making use of the Dutch transactions in the ARIADNE Portal and in archaeological terminology resources more generally.

The Tree Ring Data Standard (TRiDaS)

TRiDaS (Jansma et al, 2010) was designed collaboratively by dendrochronologists and computer scientists to accurately describe the wealth of data and metadata used in dendrochronological research. The standard supports information produced by all sub-disciplines of dendrochronology, not just archaeological and historical research facilitating the exchange of data within and between sub-disciplines. Controlled vocabularies within TRiDaS are a key aspect enabling this exchange of data.

TRiDaS provides for two mechanisms for describing vocabulary entries. For concepts with limited (<20 terms), relatively static vocabularies there are 'normalTridas' term lists defined within the TRiDaS schema. Examples of this include: dating type; timber shape; measurement method; and measurement variable (see Table 5). These simple lists of terms were devised during the design of the standard itself with the potential to extend them if necessary when the standard is revised.

normalTridas vocabulary	Description
Dating type	Typically dating in dendrochronology is absolute, however, there are circumstances where this isn't the case. The dating type allows the user to define if the dating is relative or dated with uncertainty, typically using radiocarbon.
Location type	The type of location recorded for dendrochronological samples can be extremely important when interpreting results. For example dendrochronological data can be used for palaeoenvironmental reconstructions, but for these analyses to be valid the growth location of the tree is required. Samples can be taken from trees in their growth location, from items (such as ships) that are inherently mobile, or from items (such as buildings) that are static.
Measuring method	There are a number of methods used for recording dendrochronological measurements depending on the circumstances, each with their pros and cons.
Remark	Observations about individual tree rings can be an extremely useful indicator of environmental change. The TRiDaS remark vocabulary standardises the most common features such as: false rings; missing rings; and frost damage.
Shape	This vocabulary standardises the description of the shape of timbers.
Unit	The unit vocabulary standardises the units for both ring-width measurements and measurements of timbers.
Variable	The typical measurement variable in dendrochronology is the ring-width; however researchers may also record sub-annual measurements (early/late wood), various density metrics, and vessel size. This vocabulary is likely to be revised as novel approaches are developed.

Table 5: Summary of the 'normalTridas' vocabularies used in TRiDaS. These short, simple term lists are defined within the TRiDaS schema and are relatively static

The second and more typical style of vocabulary in TRiDaS is the 'controlledVoc' datatype. This enables users to define links to external vocabularies with a standardised term and identifier. This mechanism was designed into TRiDaS, recognising the rapid development of standard vocabularies that are suitable for use in dendrochronological research.

While the TRiDaS development team intended for the standard to largely use external vocabularies as they become available, they also acknowledged the short-term needs of the dendrochronological community. As such, a vocabulary was developed for use primarily by members of the Digital Collaboratory for Cultural Dendrochronology (DCCD – Jansma et al, 2012) project describing the object/element types used in dendrochronological research. These range from the obvious "*tree*", to many items found in the archaeological and cultural record e.g. *buildings, barrels, ships, doors, musical instruments, paintings* etc.

The object/element vocabulary was written as a multilingual (English, Dutch, French and German) flat-table containing no hierarchical relationships. Terms in one or more languages with no direct translations caused confusion and overlapping concepts. Many of the terms have exact matches with the AAT. However, substantial proportions are specialist terms (especially nautical terms) that have only very generic matches.

During the course of the ARIADNE project the DCCD object / element vocabulary has been substantially reworked. Using bespoke scripts, the redundancy within the flat table has been removed, and basic hierarchical relationships defined. The simple terms list has been converted to a true concepts-based vocabulary with redundant terms assigned as alternate labels. Links to the AAT have been established for all concepts (either exact or broader relationships) and scope notes added.

The majority of the effort required to rework the vocabulary came from content specialists. Combining the specialist knowledge for all subject areas across four languages was painstaking work. Attempts to locating existing software aimed at content rather than informatics specialists were unsuccessful. Such a tool is sorely needed to fully leverage the knowledge of content specialists.

The enhanced vocabulary is currently in the process of being incorporated back into the DCCD repository. The ambiguous nature of the original term list means work is required to cross-map existing records to the new vocabulary, and in some cases this unfortunately requires consultation with the original data providers.

The second substantial vocabulary used in TRiDaS/DCCD is the species taxonomic dictionary. The basis of this vocabulary is the Species 2000 and ITIS Catalogue of Life (http://catalogueoflife.org/). The Catalogue of Life (CoL) forms the taxonomic backbone for many major projects including the Global Biodiversity Information Facility (GBIF), the Encyclopedia of Life (EoL) and the IUCN red list of endangered plants and animals. Annual editions of the CoL have been produced since 2000 with the most recent edition including over 1.6 million species from 158 contributing databases. While the CoL is an incredible resource, it suffers from the drawback that there is no linkage between concepts in each edition. While efforts are underway to produce a true SKOS mapping, this is not yet available. In the interim, TRiDaS/DCCD is using a static subset with the intention of migrating to the dynamic CoL SKOS once released.

ICCD / RA Thesaurus

The issue of multilingualism is a matter that needs to be taken into account, not only because of the variety of national thesauri that are going to be integrated by the ARIADNE initiative, but also for the future creation of common and transnational terminological tools. Linguistic issues often make the direct mapping of a concept via the *skos:exactMatch* property to the AAT concept difficult. However, other mapping relationships are available. The conceptual mapping between the ICCD RA Thesaurus and AAT has been completed and revised; for this purpose it was decided to manually construct a mapping from the various terms and functions (if any), following in sequences the three main categories of the RA Thesaurus. The work pattern was based on an Excel representation of the thesaurus to which additional columns were added in order to specify:

- The *targetLabel* and the identifier (*targetURI*) of the corresponding concept selected in AAT
- *matchURI* was one of the SKOS mapping properties (*skos:closeMatch*; *skos:exactMatch*; *skos:broadMatch)*
- The name of the institution in charge of the definition of each specific mapping (creator)

Only a subset of the RA Thesaurus was taken into account to demonstrate the feasibility of these operations. The subset includes 1191 terms related to 10 major categories (highlighted in the original source as "*livello_1_categoria*") relating to:

- CLOTHING AND ACCESSORIES
- FURNISHING
- TRANSPORTATION
- CONSTRUCTION INDUSTRY
- PAINTING
- ARCHAEOBOTANICAL FINDINGS

- ARCHAEOZOOLOGICAL FINDINGS
- SCULPTURE
- INSTRUMENTS - TOOLS AND OBJECTS OF USE
- GENERAL TERMS

The analysis for finding the corresponding entries in the AAT thesaurus took into account the information provided by scope notes and images accompanying each concept; extensive web searches were performed to find the most appropriate matching term between Italian and English; and terminological research was carried out using different resources to identify synonyms to make the associated *targetLabel* as unique and as precise as possible.

The mapping work also includes other "113" terms and COINS category (derived from "dc: title" element of XML files uploaded to CulturaItalia and delivered to the ARIADNE Portal). In total, the thesaurus includes 11 categories and 1304 terms.

The mapping work has identified the following SKOS match types:

- 642 skos:exactMatch;
- 94 skos:closeMatch;
- 310 skos:broadMatch;
- 258 skos:narrowMatch.:

The mapping methodology adopted is based on the following three examples of association provided in the table:

Categoria						
livello1	**livello2**	**livello3**	**Livello4 termine**	**targetLabel**	**AAT ID**	**matchLabel**
Mezzi di trasporto	Terrestri	A trazione animale	**cisium**	two-wheeled carriages	300215685	broad match
Strumenti - Utensili e Oggetti d'uso	Armi e Armature	Armi da difesa	**farsetto da armare**	arming doublets	300226824	close match
Scultura			**imago clipeata**	**clipei (portraits)**	300178246	exact match

Table 6: Examples of mappings between ICCD/RA terms and Getty AAT concepts

In reflection, the most significant activity, from the scientific-methodological point of view, has been the review of the whole process. Started as punctual control "1:1" correspondence between the terms of the two terminology tools (thesaurus ICCD / RA and AAT), this review has expanded by realizing the mapping of the terminological categories relating to individual entries with the codes referring to the facet and the

hierarchy AAT. This has made possible:

1. Disambiguating and correction of matches previously selected - and often lexically corrected - but decontextualised from their original domain;
2. Providing the basis for future matching between different categories of multilingual thesauri.

It is worth emphasising that the focus of the mapping work is the concept of individual terms meant as records entered in a complete hierarchical structure of related terms and notes.

Among the results achieved — and which are highlighted though the mapping between classes — are the high level of correspondence between the ICCD / RA thesaurus entries and the AAT thesaurus record types. Out of 1,191 basic records, 1,164 among them are linked to "concept" and only 27 to "guide term". According to the AAT Thesaurus guidelines:

- Concept: Refers to records in the AAT that represent concepts; records for concepts include terms, a note, and bibliography.
- Guide term: Refers to records that serve as place savers to create a level in the hierarchy under which the AAT can collocate related concepts. Guide terms are not used for indexing or cataloguing.

INRAP (FRANTIQ)

DOLIA is the catalogue of the archaeological reports at the French National Institute for Preventive Archaeological Research (Inrap). The DOLIA catalogue was developed with Flora 3.1.0 software, created by Everteam (© Everteam 2015) http://dolia.inrap.fr:8080/flora/jsp/index.jsp. The reports, stored in pdf format, are indexed with native subjects inherited by the Pactols "Sujets / Subjects" thesaurus.

The DOLIA catalogue currently has 1,573 (5,149 occurrences) subject metadata terms in the Pactols thesaurus. The current mapping concerns only the indexed terms from the DOLIA catalogue used in ARIADNE.

The alignment has been done between those terms and the AAT thesaurus by using a source term from Pactols, a source URI, a target term from AAT and a target URI, specifying the SKOS match.

E.g.
Pactols: Archéologie
http://ark.frantiq.fr/ark:/26678/pcrty05M9SVnLu
skos:exactMatch
AAT: archaeology
http://vocab.getty.edu/aat/300054328

or
Pactols: amphore gauloise
http://ark.frantiq.fr/ark:/26678/pcrtiUhJYvi7PG
skos:broadMatch
AAT: amphorae (storage vessels)
http://vocab.getty.edu/aat/300148696

Match type	Mappings	Proportion
skos:exactMatch	1161	71%
skos:closeMatch	121	7%
skos:broadMatch	346	21%
skos:narrowMatch	6	

Table 7: result of the alignment

A complete mapping of the Pactols Subjects is planned in the next few months.

Irish Monuments Vocabulary

The most detailed classification system available for Irish Monument types is the class list developed by the National Monuments Service (NMS). This is a flat / simple hierarchical list which was used in the classification of sites and monuments that formed part of the Archaeological Survey of Ireland, which was established to compile an inventory of the known archaeological monuments in the State. The information is stored on a database and in a series of paper files that collectively form the ASI Sites and Monuments Record (SMR). Each site / monument has a unique SMR number which greatly facilitates the creation of Linked Data, and each site / monument is given a classification based on the NMS class list. The development of the list was an organic and evolving process and the list is subject to review with amendments being made on an on-going basis.

Irish Monuments Mapping was undertaken by the Discovery Programme in order to map the subject classifications in the NMS list to the Getty AAT. This was done for each term by comparing the scope notes of the NMS class list to the notes field of the AAT Online. This automatically introduces a level of subjectivity which was countered by using an appropriate SKOS mapping property when linking to the target vocabulary (AAT). Where there was any ambiguity about the term, broader mapping properties were always used.

In certain cases where mappings were difficult and could be more closely related to the UK FISH Thesaurus of Monument Types, the Vocabulary Matching Tool developed by USW was first used to identify matching terms, which was in turn mapped to the AAT (i.e. a two stage mapping process).

The nature of the classification list of the NMS presented occasional difficulties:

- Some classifications contained highly detailed elements e.g. object terms were refined at term level by their present location [*Cist (present location)*] or were developed in order to classify idiosyncratic sites [*turf stand; watchman's hut-burial ground*].
- There was greater congruence between the FISH Monument Type vocabulary and the Irish subject terms enabling greater possibilities to find an exact or close match. In some cases terms had clearly been based on the FISH vocabulary. This was to be expected due to geographical / historical contiguity. For example *bullaun stone*, for which there are over 1000 currently documented in the ASI, relates more closely to a *'cup-marked stone'* in FISH but can only be satisfactorily mapped using two (or more) terms in the Getty AAT [*ceremonial objects; mortaria*].
- Some terms are not clearly defined in the NMS class list [e.g. *settlement platform*: '*A raised area, often surrounded by waterlogged or boggy land, which has evidence of former human habitation*'] which made mapping, even at a high level, difficult.
- Subject definitions often included broad period classifications within the scope note; it was decided not to take this into consideration as period terms could be covered by the Irish Periods Vocabulary. Occasionally terms contained period terms in their term name (e.g. *House-16th century; House-16th/17th century*) as well as a refining subject element (e.g. *House-fortified house*) This necessitates both the use of the Irish Periods Vocabulary and/or additional terms from the AAT.

- Some classifications were subdivided (but not hierarchically) into more specific elements (e.g. *Ringfort-cashel*; *Ringfort-rath*; *Ringfort-unclassified*). The granularity of the terms was conserved by using the appropriate mapping property, in some cases by mapping terms to multiple terms in the target vocabulary e.g.
    - *Ringfort-cashel -> [skos:broadMatch] -> raths*
    - *Ringfort-cashel -> [skos:broadMatch] -> dry walls (masonry)*

The mapping process attempted to balance the pressing need to implement Linked Data with the reality that the available vocabulary was rich in detail, but lacked a structure that was easily reconciled with standard concepts of controlled vocabularies and indexing. This was largely achieved by multiple mappings to the target vocabulary, as well as by utilising an intermediate vocabulary which more closely reflected the particular nuances of Irish monument types.

4 Mappings in the ARIADNE infrastructure

The ARIADNE Catalogue Data Model (ACDM) specifies the metadata schema that underpins the ARIADNE infrastructure (see D12.2: *Infrastructure Design*). The ACDM is based on the DCAT vocabulary, adding classes and properties needed for describing ARIADNE assets. The ARIADNE Catalogue aggregates metadata, such as descriptions for datasets, metadata schemas, vocabularies, etc. provided by the project partners through metadata file uploads, or the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[1] The metadata and object repository aggregator (MORe)[2] (Isaac et al., 2013) has been customized for ARIADNE purposes and is driven by the ACDM. MORe includes a set of micro-services, including various metadata enrichment services. For ARIADNE purposes, a bespoke derived AAT subject enrichment service has been developed by ATHENA DCU that applies the partner vocabulary mappings (in JSON format) to the partner subject metadata and derives an AAT concept (both preferred label and URI) to augment the subject metadata, both in the Registry and also supplied to the ARIADNE Portal.

For subject access, the ACDM *ArchaeologicalResource* class has two kinds of subject property. The property, native-subject, associates the resource with one or more items from a controlled vocabulary used by the data provider to index the data. However as discussed in Section 2.2, there are a large number of partner vocabularies in several different languages, and cross search is rendered difficult, as there are no semantic links or mappings between the various local vocabularies. The established solution to this problem is to employ mapping between the concepts in the different vocabularies. However, as discussed above, the creation of links directly between the items from different vocabularies can quickly become unmanageable as the number of vocabularies increases. A scalable solution to this mapping problem is to employ the hub architecture, an intermediate structure where concepts from the ARIADNE data provider source vocabularies can be mapped (ISO 2013). In the portal, retrieval based on a concept from one vocabulary (in a search or browsing operation) can use the hub to connect to subject metadata from other vocabularies, possibly expressed in other languages. In the ACDM, *ariadne-subject* is used for shared concepts from the hub vocabulary (the AAT), which have been derived via the various mappings from source vocabularies. This underpins the MORe enrichment services augmenting the data imported to the Registry with mapped hub concepts. These derived subjects in turn make possible concept based search and browsing in the ARIADNE Portal. It is thus anticipated that the mappings can form one of the stepping stones towards a multilingual capability in the Portal.

4.1 Mapping enrichment process

The AAT Linked Open Data that forms the basis of the ARIADNE mapping hub vocabulary is expressed in a combination of ontological models including SKOS. The appropriate representation for the mappings is via SKOS mapping properties (see SKOS Mapping Properties). The output from the mapping tools of the partner mappings from their source vocabularies to the AAT is transformed to the required JSON format by USW for communication to the Registry team at DCU, where it is processed by the relevant MoRe enrichment services. A brief example of this JSON format is given in Appendix B.

The information from the mapping a tool is passed to MORe which associates it with the provider of the vocabulary. It updates the property *derived-subject* using the AAT mappings and enriches an ACDM record (see Figure 15), adding a broader term, or a *skos:altLabel* to correlate a term using the 'use for' relationship, or adds multilingual labels (*skos:prefLabel* and *skos:altLabel*) in order to facilitate multilingual search.

[1] http://www.openarchives.org/pmh/

[2] http://more.dcu.gr/

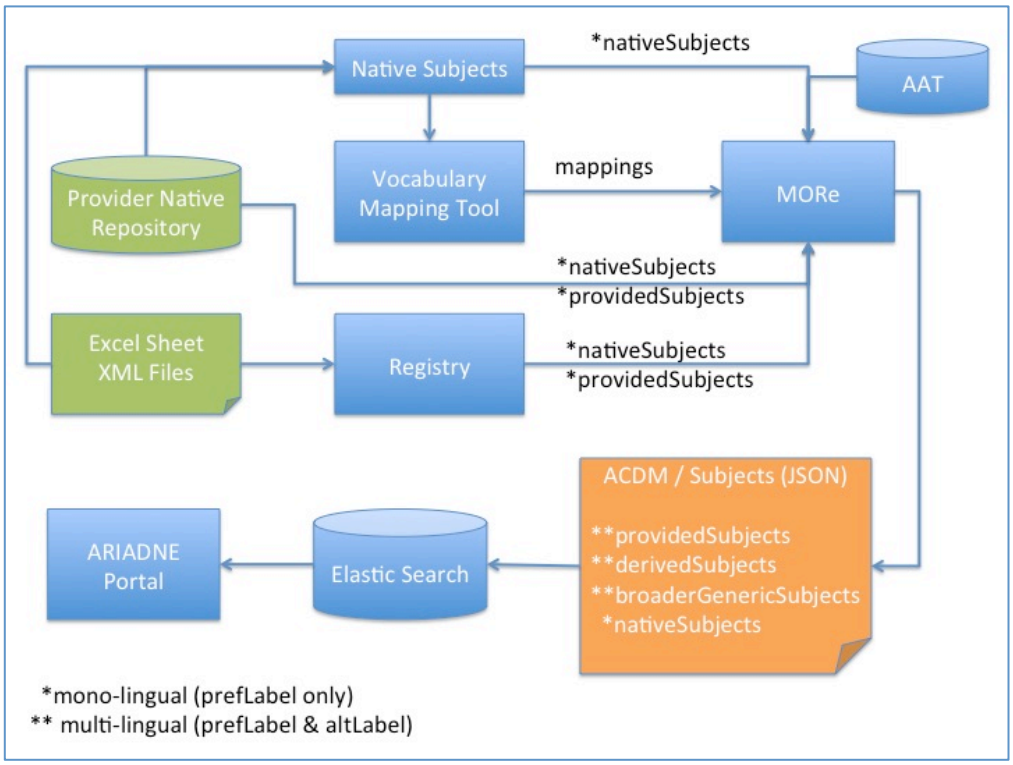

Figure 15: MORe enrichment

4.2 Mappings within the ARIADNE portal

At the time of writing, development of the ARIADNE Portal and the search functionality is still ongoing with mappings still being imported from some partners. However, it is possible to have a preview at this stage. Figure 16 shows a query on the Portal making use of the mappings. On the main Results screen, a set of filters is available for refining a search following the faceted search paradigm. One filter, currently named *Derived Subject*, is populated by the MORe enrichment process described in section 4.1; effectively the Derived Subjects are AAT concepts, which have been mapped to the native vocabulary concepts that form the subject metadata of the data resources in the Portal. Figure 16 shows that a simple query on the single AAT (mapped) concept, *churches (buildings*), is able to retrieve results in multiple languages from AIAC (Fasti), DAI and DANS ARIADNE content providers. Results from ADS are also returned though not shown in this screen dump, which only shows a small number of the overall total results.

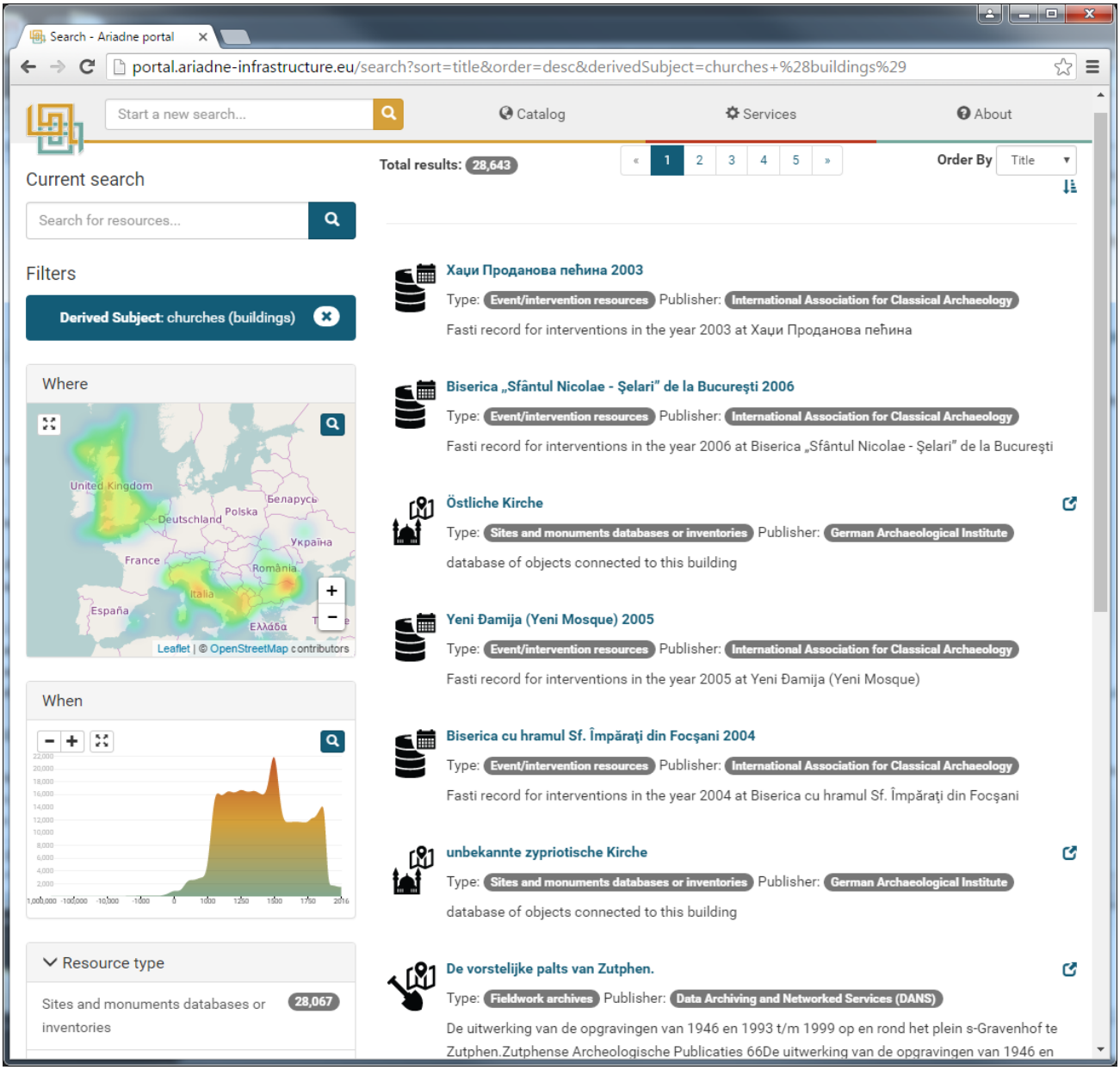

Figure 16: Portal Query on AAT mapped subject: churches (buildings) showing results from AIAC (Fasti), DAI and DANS, with multiple languages (June 2016)

In future work, making the mappings (and mapping services) fully available as outcomes in their own right, with appropriate metadata for the mappings would be desirable, as more than one mapping may be produced for large vocabularies. The mappings may also serve to underpin a multilingual capability in an initial string search, by augmenting the language coverage of the AAT.

Full technical documentation about the mappings presented in this report is available for download from the ARIADNE web site.

5 Conclusion

This report has reviewed the key vocabularies considered relevant to the ARIADNE project. Mapping between vocabularies has been shown to be a key aspect for concept based search, avoiding the ambiguities posed by literal string search and making possible a multi-lingual search capability. The Getty AAT was selected as a mapping hub vocabulary and partner native vocabularies have been mapped to it using SKOS mapping relationships and bespoke mapping utilities. The mappings have been incorporated into the Registry enrichment process so that partner subject metadata has been augmented by AAT concepts.

The two prototype experiments also show the potential of working with the URI identifiers of AAT concepts rather than the ambiguous strings of term labels. Using the URI identifier for the concept avoids the problem of ambiguity, common in multilingual datasets, for terms that are homographs in different languages. Working at the concept level also makes possible hierarchical semantic expansion, making use of the broader generic ("IS-A") relationships between concepts in a hierarchically structured knowledge organization system, such as the AAT. Thus a search expressed at a general level can (if desired) return results indexed at a more specific level. For example, a search on *settlements* might also return *monastic centres*.

An example from the ARIADNE Portal has illustrated the potential for the mappings to assist a query in retrieving results in multiple languages. The mappings have potential to underpin various options in the search functionality and user interface, offering a cost effective route towards different forms of multilingual functionality.

6 References

Aitchison, J., Gilchrist, A., Bawden, D. (2000). Thesaurus construction and use: a practical manual (4th edition). ASLIB, London.

ARIADNE project (2016). Available at: http://www.ariadne-infrastructure.eu/ [Accessed 15 Jun. 2016].

ARIADNE Catalog Data Model (ACDM)

Art and Architecture Thesaurus. J. Paul Getty Trust. http://www.getty.edu/research/conducting_research/vocabularies/aat/index.html [Accessed 15 Jun. 2016]

Berners-Lee, T. Linked Data. Available at: http://www.w3.org/DesignIssues/LinkedData.html

Binding C., Tudhope D. (2016). Improving Interoperability using Vocabulary Linked Data. International Journal on Digital Libraries, 17(1), 5-21. Springer.

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3): 1–22.

Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Jaques, Y. and Keizer, J. (2012). Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. International Journal of Metadata, Semantics and Ontologies, 7(1): 65-75. Inderscience.

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbahndari, S., Jaques, Y. and Keizer, J. (2013). The AGROVOC Linked Dataset. Semantic Web, 4(3): 341-348. IOS Press

Charles, V., Devarenne, C. (2014). Europeana enriches its data with the AAT. EDM case study. Available at: http://pro.europeana.eu/page/europeana-aat [accessed 30/11/2015]

Data Catalog Vocabulary (DCAT) Available at: http://www.w3.org/TR/vocab/dcat/

Getty Research Institute (2016a). Getty Vocabularies [online] Available at: http://www.getty.edu/research/tools/vocabularies/ [Accessed 15 Jun. 2016].

Getty Research Institute (2016b). Getty Vocabularies as Linked Open Data. [online] Available at: http://www.getty.edu/research/tools/vocabularies/lod/ [Accessed 15 Jun. 2016].

Getty Research Institute (2016c). Getty Vocabularies SPARQL endpoint. [online] Available at: http://vocab.getty.edu/sparql/ [Accessed 15 Jun. 2016].

Gormley, C., Tong, Z. (2015). Elasticsearch – The Definitive Guide. Genre Expansion [online] Available at: https://www.elastic.co/guide/en/elasticsearch/guide/current/synonyms-expand-or-contract.html#synonyms-genres

Harpring, P. (2016). Art and Architecture Thesaurus: Introduction and Overview. http://www.getty.edu/research/tools/vocabularies/aat_in_depth.pdf [Accessed 15 Jun. 2016].

Heritagedata.org. (2016). Linked Data Vocabularies for Cultural Heritage [online] Available at: http://www.heritagedata.org/ [Accessed 15 Jun. 2016].

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

ISO25964-1:2011. Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. Available at: http://www.niso.org/schemas/iso25964/#part1 [accessed 30/11/2015]

ISO 25964-2:2013. Information and documentation - Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies. Available at: http://www.niso.org/schemas/iso25964/#part2 [accessed 30/11/2015]

Jansma, E., Brewer, P. and Zandhuis, I. (2010). TRiDaS 1.1: The tree-ring data standard. Dendrochronologia, 28(2), pp.99-130. Available at: http://dx.doi.org/10.1016/j.dendro.2009.06.009 [accessed 15/06/2016]

Jansma, E., van Lanen, R., Brewer, P. and Kramer, R. (2012). The DCCD: A digital data infrastructure for tree-ring research. Dendrochronologia, 30(4), pp.249-251. Available at: http://dx.doi.org/10.1016/j.dendro.2011.12.002 [accessed 15/06/2016]

Kempf, A., Neubert, J. (2016). The Role of Thesauri in an Open Web: A Case Study of the STW Thesaurus for Economics. Knowledge Organization, 43(3), 160-173. Ergon Verlag.

Koch, T., Neuroth, H. and Day, M. (2003). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In: McIlwaine, I.C. (ed.) Subject retrieval in a networked world: proceedings of the IFLA Satellite Meeting held in Dublin, OH, 14-16 August 2001. (UBCIM Publications, New Series, Vol. 25). München: K.G. Saur, 25-33.

SKOS Mapping Properties. Available at: http://www.w3.org/TR/skos-reference/#L4138

SparqlGui (2016) - desktop RDF querying tool. Available at: https://bitbucket.org/dotnetrdf/dotnetrdf/wiki/UserGuide/Tools/SparqlGui

STW Thesaurus for Economics and associated web services. Leibniz Information Centre for Economics. Available at: http://zbw.eu/stw/ [accessed 30/11/2015]

Tudhope D., Koch T., Heery R. (2006). Terminology Services and Technology: JISC state of the art review. Available at: http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf [accessed 15/06/2016]

Tudhope, D., Binding, C. (2016). Still Quite Popular After all Those Years - The Continued Relevance of the Information Retrieval Thesaurus. Knowledge Organization, 43(3), 174-179. Ergon Verlag.

Vizine-Goetz, D., Hickey, C., Houghton, A., Thompson, R. (2003). Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4), Article No. 272, 2004-03-11. Available at: https://journals.tdl.org/jodi/index.php/jodi/article/view/114/113 [accessed 15/06/2016]

Zeng, M., Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5): 377-395. Wiley.

7 Appendix A

Concept mappings used for the prototype mapping exercise (Turtle RDF format):

```
# namespace prefixes
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix aat: <http://vocab.getty.edu/aat/> .
@prefix fasti: <http://fastionline.org/monumenttype/> .
@prefix iccd: <http://www.iccd.beniculturali.it/monuments/> .
@prefix dans: <http://www.rnaproject.org/data/> .
@prefix tmt: <http://purl.org/heritagedata/schemes/eh_tmt2/concepts/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix gvp: <http://vocab.getty.edu/ontology#> .
@prefix dai: <http://archwort.dainst.org/thesaurus/de/vocab/?tema=> .

# ICCD concepts
iccd:catacomba skos:prefLabel "catacomba"@it .
iccd:cenotafio skos:prefLabel "cenotafio"@it .
iccd:cimitero skos:prefLabel "cimitero"@it .
iccd:colombario skos:prefLabel "colombario"@it .
iccd:dolmen skos:prefLabel "dolmen"@it .
iccd:mausoleo skos:prefLabel "mausoleo"@it.
iccd:menhir skos:prefLabel "menhir"@it.
iccd:monumento-funerario skos:prefLabel "monumento funerario"@it .
iccd:necropoli skos:prefLabel "necropoli"@it .
iccd:sepolcreto-rupestre skos:prefLabel "sepolcreto rupestre"@it .
iccd:tomba skos:prefLabel "tomba"@it .

# ICCD->AAT mappings
iccd:catacomba skos:closeMatch aat:300000367 .
iccd:cenotafio skos:closeMatch aat:300007027 .
iccd:cimitero skos:closeMatch aat:300266755 .
iccd:colombario skos:closeMatch aat:300000370 .
iccd:dolmen skos:closeMatch aat:300005934 .
iccd:mausoleo skos:closeMatch aat:300005891 .
```

```
iccd:menhir skos:closeMatch aat:300006985 .
iccd:necropoli skos:closeMatch aat:300000372 .
iccd:sepolcreto-rupestre skos:closeMatch aat:300387008 .
iccd:tomba skos:closeMatch aat:300005926 .

# DANS concepts
dans:8f14ae7e-3d66-4e85-b77c-454a261150e9 skos:prefLabel "begraving"@nl .
dans:e98c8cf0-aa0d-4fcd-99a2-db76cd1d827d skos:prefLabel "begraving, onbepaald"@nl .
dans:87a2f9e9-8e40-4c97-b17b-82275d54c78d skos:prefLabel "brandheuvelveld"@nl .
dans:be95a643-da30-40b9-b509-eadfb00610c4 skos:prefLabel "christelijk/joodse begraafplaats"@nl .
dans:77130cff-58e0-4c6d-b608-33fadc946283 skos:prefLabel "dierengraf"@nl .
dans:df17ef8a-1a58-4c58-ab6f-2e127c90c571 skos:prefLabel "grafheuvel"@nl .
dans:9a729782-ca06-47e1-aa50-87561f36a8ee skos:prefLabel "grafheuvelveld"@nl .
dans:6a7482e5-2fd5-48fb-baf4-66ad3d4ed95e skos:prefLabel "kerkhof"@nl .
dans:e1f67762-c405-42a5-b073-88c13043aab0 skos:prefLabel "megalietgraf"@nl .
dans:abb41cf1-30dc-4d55-8c18-d599ebba1bc2 skos:prefLabel "rijengrafveld"@nl .
dans:74899123-2b00-4e12-83f2-f37bc4f129ff skos:prefLabel "terechtstellingsplaats/galgenberg"@nl .
dans:b98f1315-91c5-411e-b91b-9693e5dfc5c2 skos:prefLabel "urnenveld"@nl .
dans:a156e09c-b40c-45a9-8487-d7b68f8dbae7 skos:prefLabel "vlakgraf"@nl .
dans:b935f9a9-7456-4669-91d0-2e9c0ff7d664 skos:prefLabel "vlakgrafveld"@nl .

# DANS->AAT mappings
dans:8f14ae7e-3d66-4e85-b77c-454a261150e9 skos:closeMatch aat:300387004 .
dans:e98c8cf0-aa0d-4fcd-99a2-db76cd1d827d skos:closeMatch aat:300387004 .
dans:be95a643-da30-40b9-b509-eadfb00610c4 skos:broadMatch aat:300266755 .
dans:6a7482e5-2fd5-48fb-baf4-66ad3d4ed95e skos:closeMatch aat:300000360 .
dans:abb41cf1-30dc-4d55-8c18-d599ebba1bc2 skos:closeMatch aat:300266755 .
dans:b935f9a9-7456-4669-91d0-2e9c0ff7d664 skos:broadMatch aat:300266755 .


# EH-TMT concepts
tmt:70053 skos:prefLabel "cemetery"@en .
tmt:100531 skos:prefLabel "walled cemetery"@en .
tmt:92672 skos:prefLabel "mixed cemetery"@en .
tmt:70060 skos:prefLabel "inhumation cemetery"@en .
```

tmt:70056 skos:prefLabel "cremation cemetery"@en .

tmt:70055 skos:prefLabel "cairn cemetery"@en .

tmt:70054 skos:prefLabel "barrow cemetery"@en .

tmt:91386 skos:prefLabel "catacomb (funerary)"@en .

tmt:70053 skos:prefLabel "necropolis"@en .

EH-TMT->AAT mappings

tmt:70053 skos:closeMatch aat:300266755 .

tmt:100531 skos:broadMatch aat:300266755 .

tmt:92672 skos:broadMatch aat:300266755 .

tmt:70060 skos:broadMatch aat:300266755 .

tmt:70056 skos:broadMatch aat:300266755 .

tmt:70055 skos:broadMatch aat:300266755 .

tmt:70054 skos:broadMatch aat:300266755 .

tmt:91386 skos:closeMatch aat:300000367 .

tmt:70053 skos:closeMatch aat:300000372 .

FASTI concepts

fasti:burial skos:prefLabel "Burial"@en .

fasti:catacomb skos:prefLabel "Catacomb"@en .

fasti:cemetery skos:prefLabel "Cemetery"@en .

fasti:columbarium skos:prefLabel "Columbarium"@en .

fasti:mausoleum skos:prefLabel "Mausoleum"@en .

FASTI->AAT mappings

fasti:burial skos:closeMatch aat:300387004 .

fasti:catacomb skos:closeMatch aat:300000367 .

fasti:cemetery skos:closeMatch aat:300266755 .

fasti:columbarium skos:closeMatch aat:300000370 .

fasti:mausoleum skos:closeMatch aat:300005891, aat:300263068 .

DAI concepts

dai:1819 skos:prefLabel "Friedhof"@de . #cemetery

dai:1947 skos:prefLabel "Gräberfeld"@de . #graveyard

dai:3736 skos:prefLabel "Kolumbarium"@de . #columbarium

```
dai:2485 skos:prefLabel "Nekropole"@de . #necropolis

# DAI->AAT mappings
dai:1819 skos:closeMatch aat:300266755 .
dai:1947 skos:closeMatch aat:300000360 .
dai:3736 skos:closeMatch aat:300000370 .
dai:2485 skos:closeMatch aat:300000372 .
```

8 Appendix B

Example of the JSON exchange format for communicating the mappings to the ARIADNE Registry team, using the mappings of three FASTI (AIAC) concepts to the AAT


```
[
{
        "created": "2015-11-20T15:27:13.342Z" ,
        "sourceURI": "http://www.fastionline.org/concept/attribute/abbey" ,
        "sourceLabel": "Abbey" ,
        "matchURI": "http://www.w3.org/2004/02/skos/core#closeMatch" ,
        "targetURI": "http://vocab.getty.edu/aat/300000642" ,
        "targetLabel": "abbeys (monasteries)"
} ,
{
        "created": "2015-11-20T15:27:13.342Z" ,
        "sourceURI": "http://www.fastionline.org/concept/attribute/amphitheatre" ,
        "sourceLabel": "Amphitheatre" ,
        "matchURI": "http://www.w3.org/2004/02/skos/core#exactMatch" ,
        "targetURI": "http://vocab.getty.edu/aat/300007128" ,
        "targetLabel": "amphitheaters (built works)"
} ,
{
        "created": "2015-11-20T15:27:13.342Z" ,
        "sourceURI": "http://www.fastionline.org/concept/attribute/ancient_beach" ,
        "sourceLabel": "Ancient beach" ,
        "matchURI": "http://www.w3.org/2004/02/skos/core#broadMatch" ,
        "targetURI": "http://vocab.getty.edu/aat/300008816" ,
        "targetLabel": "beaches"
}
]
```

9 Appendix C

Extract from mapping guidelines

This document should be read in conjunction with *Mapping-Template.xlsx*. This document describes the columns in the spreadsheet template used for mapping partner source vocabularies (thesauri) to the Getty AAT (Art and Architecture Thesaurus), as part of the Subject access strategy for ARIADNE. The mappings will inform cross search for resource discovery in the ARIADNE Portal.

The mapping exercise matches concepts in a Partner vocabulary with concepts in the AAT using SKOS mapping relations (e.g. skos:broadMatch, http://www.w3.org/TR/skos-reference/#mapping). The document also contains guidelines for making the mappings.

The Mapping Template is an alternative to the (USW) Vocabulary Matching Tool, which requires that the source vocabulary be available as Linked Data. The Mapping Template allows mappings to be made by partners' own methods (e.g. using AAT and source vocabulary webpages, or some other tool) and represented in a spreadsheet. A separate spreadsheet should be produced for each partner vocabulary mapped to the AAT. The standard column names in the Mapping Template should be followed. This will allow a subsequent automatic transformation by USW to the RDF statements, employed by the Registry and Portal.

The ***first tab*** in a partner mapping spreadsheet (*Mapping-Template-partner-source.xlsx)* should contain metadata and any necessary description of the mapping exercise. This can inform a subsequent VoID metadata description of the mapping. The metadata should include the following items using the first and second columns (please substitute the Name of Source Vocabulary for XXX):-

dcterms:creator	Name of organisation doing the mapping
dcterms:created	Date of creation (one date representing a complete mapping exercise)
dcterms:modified	Date of last modification
dcterms:title	SKOS Mapping between concepts in source (XXX) and target (AAT) vocabularies using SKOS mapping properties.
void:subjectsTarget	URI of source vocabulary if known (e.g. http://purl.org/heritagedata/schemes/eh_tmt2)
void:objectsTarget	URI of target vocabulary (for ARIADNE this will be http://vocab.getty.edu/aat/)
dcterms:description	An intellectual matching made for ARIADNE from the source vocabulary **XXX** to the Getty AAT for resource discovery cross search purposes. Include here any details of method (hopefully with expert review)
dcterms:license	The Rights appropriate for Partner and ARIADNE, e.g. perhaps CC0 or CC_BY/3.0

The ***second tab*** should hold the mapping using the column names below (one spreadsheet for each different source vocabulary). A different mapping is specified in each row. The following column names **in bold** are mandatory (necessary for expressing the resulting RDF statements).

sourceLabel	*(the preferred term or label for the concept)*
sourceURI	*(use URI if it exists, otherwise unique concept ID, otherwise prefLabel again)*

matchURI *(skos:closeMatch | skos:exactMatch | skos:broadMatch)*

targetLabel *AAT label for concept (e.g. small holdings)*

targetURI *AAT URI for concept (e.g. http://vocab.getty.edu/aat/300000211)*

Additional optional columns may be useful while creating the mappings or for human inspection of the mapping spreadsheet by partners but are not required. Examples of optional columns from partner mapping work to date include:

Source-Hierarchy *(hierarchy or category the source concept belongs to)*

Source-ScopeNote *(scope note or definition of concept - this may be particularly useful)*

Source-En *(an English language translation, or other languages if desired)*

Comment *(if desired, any comment on this mapping, eg a rationale)*

Other-Target-prefLabel *(if useful to partner to also include mappings to other thesauri)*

Other-Target-URI *(if useful to partner to also include mappings to other thesauri)*

Mapping guidelines

The aim of the mapping exercise is to identify subject mappings to AAT for concepts that are likely to be useful to assist browsing and search of the portal (time and space are being handled separately).

If any existing mappings to AAT are known they may be useful to build on. The AAT can also be searched and browsed manually via the Getty website – http://www.getty.edu/research/tools/vocabularies/aat/

Probably the AAT **Objects hierarchy** is the most relevant hierarchy.

If resources are limited, a sensible strategy is to start with the most useful concepts in the first instance for the datasets/reports partners have provided to the Registry. These would probably include the top (say 2) levels of relevant partner thesauri (e.g. Objects and Monument types) and also concepts used to index the data provided to the registry. It will also include controlled keyword lists used by partners to index the data.

Match types for the mapping

If the mapping is approximate then **skos:closeMatch** is probably the best match type. If it is a very good match then **skos:exactMatch** is appropriate. In general, do **not** make use of skos:relatedMatch for ARIADNE purposes (unless perhaps as an additional mapping for a given concept). The idea is to make the most appropriate match for each concept in the Partner vocabulary.

Usually you will just make *one* match (the best one) to AAT for any given concept - there is usually no need to express multiple relationships to AAT concepts as this is provided gratis via the AAT's semantic structure. Thus if you make a match from a given partner concept to an AAT concept then there is no need to also make mappings to narrower AAT concepts for that given partner concept. The only exception is if the partner concept has two genuinely quite different expressions in the AAT (that are not immediate parent or child concepts). In this case one or two additional mappings are possible but that should be very much the exception. Normally you would work through a hierarchy making a mapping for each concept, giving complete coverage of that hierarchy.

If a partner concept is much more specific than any AAT concept then you can make a **skos:broadMatch** to the AAT concept. This is useful for cases when a partner vocabulary has detailed archaeological concepts. It is not expected that you would need to make much use of skos:narrowMatch for ARIADNE vocabularies.

Matches should be made to AAT *concepts* rather than guide-terms (inside <>). If an AAT guide term appears as a match in the tool, consider a narrower or broader concept in the AAT. For example, instead of mapping

to <containers by form>, it is better to map to *containers (receptacles)* even if the mapping relationship needs to be ***skos:broadMatch*.**

Where top level partner concepts are too high level or general (eg perhaps '*society*', '*religion*') to map easily then probably best to consider the next level down. If any partner concepts prove particularly problematic then just set them aside and discuss with USW later.

Optional matching tool

When vocabularies are *already* available as Linked Data via the Registry or via HeritageData then the USW **Vocabulary Matching Tool** may be helpful.

http://heritagedata.org/vocabularyMatchingTool/

When using the Vocabulary Matching Tool, remember to **Save** the data before ending a session (data is saved in JSON format). This allows you to subsequently **Load** the JSON file into the tool and make revisions or further mappings. When sending the final results of the matching exercise, please send us the **JSON** format file.