D15.3 Report on Semantic Annotation and Linking

Version 1.0

January 2017

Grant Agreement number:	313193
Project acronym:	ARIADNE
Project title:	Advanced Research Infrastructure for Archaeological Dataset Networking in Europe
Funding Scheme:	FP7-INFRASTRUCTURES-2012-1
Project co-ordinator:	Franco Niccolucci, PIN Scrl - Polo Universitario "Città di Prato"
Tel:	+39 0574 602578
E-mail:	franco.niccolucci@gmail.com
Project website address:	http://www.ariadne-infrastructure.eu

ARIADNE is a project funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193. The views and opinions expressed in this report are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

About this document

This document is a contractual deliverable of the ARIADNE project. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7-INFRASTRUCTURES-2012-1) under grant agreement n° 313193. But the content of this document cannot be considered to reflect the views of the European Commission.

Partner in charge of the deliverable:	University of South Wales (USW)
Authors:	Douglas Tudhope (USW) Ceri Binding (USW)
Contributors:	Dimitris Gavrilis (ATHENA RC) Paul Boon, Hella Hollander (KNAW-DANS) Jeremy Azzopardi (SND) Andreas Vlachidis (USW) Guntram Geser (SRFG)
Quality review:	Franco Niccolucci and Paola Ronzino (PIN)

Table of content

1 Executive Summary

Brief summary

The review and development work described in this report focuses on the aspects of semantic linking and annotation particularly relevant to ARIADNE. Semantic linking within ARIADNE is considered within the spatial, temporal and subject dimensions. The subject dimension is considered in depth, starting with a review of linking tools considered relevant to ARIADNE followed by a discussion of the ARIADNE approach and the vocabulary mapping tools used within ARIADNE. The Getty AAT proved an appropriate vocabulary mapping hub that afforded a multilingual search capability in the ARIADNE Portal via the semantic enrichment of partner subject metadata with derived AAT concepts.

A case study conducted an exploratory investigation of the semantic integration of extracts from archaeological datasets with information extracted via NLP across different languages. The investigation followed a broad theme relating to wooden material including shipwrecks, with a focus on types of wooden material, samples taken, wooden objects with dating from dendrochronological analysis, etc. The Demonstrator is available for general use. The user is shielded from the complexity of the underlying semantic framework (based on the CIDOC CRM and Getty AAT) by the Web application user interface. The Demonstrator highlights the potential for archaeological research that can interrogate grey literature reports in conjunction with datasets. Queries concern wooden objects (e.g. samples of beech wood keels), optionally from a given date range, with automatic expansion over hierarchies of wood types.

Lessons learned

- The spatial, temporal and subject dimensions are key to archaeology and the main vehicle for semantic linking. Cleansing and normalisation via semantic enrichment is necessary in each dimension.
- The Getty AAT proved an appropriate vocabulary mapping hub that afforded a multilingual search capability in the ARIADNE Portal.
- The Wood/Dendrochronology case study demonstrated the feasibility of connecting information (at a detailed level) extracted from datasets and grey literature reports in different languages and semantic cross-searching of the integrated information. The case study suggests that the semantic linking of textual reports and datasets opens up possibilities for integrative archaeological research across diverse resources.
- The Wood/Dendrochronology demonstrator shows that a Web application can hide the complexity of the underlying semantic framework from the user and allow querying and browsing the information without expertise in SPARQL; it illustrates that more intuitive user interfaces are possible for searching RDF datasets than the usual SPARQL endpoint or browsing hyperlinks.

2 Semantic annotation and linking within ARIADNE

Semantic annotation is covered in part by the ARIADNE Natural Language Processing (NLP) work reported under WP16. For more information on automatic text annotation (and text-mining) within archaeology, including machine learning outcomes from the ADS ArchaeoTools Project and the rule-based OPTIMA toolkit from the USW STAR Project, see Richards *et al.* (2015) and the deliverable ARIADNE D16.4 - Final Report on Natural Language Processing, which also reports on NLP outcomes from ARIADNE. One strand of these outcomes contributed to the Wood/Dendrochronology data integration case study and Demonstrator, which combines semantic links between datasets and (via NLP) archaeological reports. This is discussed below.

In archaeology, *semantic linking* generally involves connections (usually semi-automatic with intellectual review) between archaeological resources via an intermediary knowledge organization system covering relevant aspects of the space, time and subject dimensions. Semantic linking in the subject dimension is discussed below as regards mapping between subject thesauri and controlled vocabularies, while linking via the CIDOC CRM core ontology, is discussed in ARIADNE D14.2 - Pilot Deployment Experiments.

2.1 Semantic linking in the spatial dimension

For ARIADNE, a key component of the strategy for the spatial domain has been the adoption of a standard format for spatial coordinates, WGS84. Different partners employ different local coordinate systems which are normalised to WGS84. Normalised spatial coordinates constitute the main spatial access method in the ARIADNE Portal. Relevant gazetteers of place names, including GeoNames and TGN, are discussed in ARIADNE D3.1 - Initial report on standards and on the project registry. The work of the Pelagios project has been a significant contributor to the Digital Humanities arena of spatial linked data, including various workshops in the Linked Pasts series. Pelagios makes use of the Pleiades gazetteer[1] (and its URIs) to connect online resources that refer to places in the ancient world via Linked Open Data. Pelagios does not attempt to define a complex data model, rather it seeks to offer a uniform way to build links between different gazetteers via the Open Annotation Ontology. Pelagios metadata aims to support interoperability while imposing minimal overheads on data providers. Pelagios is developing various tools (beta at time of writing), including Recogito[2] , an annotation platform for annotating images, creating maps, linking research data to the wider web of data, the Pelagios Map Tiles mapping application, the Peripleo search service and API[3] . Various other efforts are ongoing in the second phase of the Pelagios project, which aims to encourage community participation in the enterprise: “Pelagios Commons provides online resources and a community forum for using open data methods to link and explore historical places”[4].

2.2 Semantic linking in the temporal dimension

In the temporal domain, ARIADNE made the decision to adopt the recently emerged PeriodO gazetteer as a central hub for expressing standard period definitions, in order to link and visualize

[1] Pleiades, https://pleiades.stoa.org

[2] Recogito, http://recogito.pelagios.org

[3] Peripleo, https://github.com/pelagios/peripleo#peripleo-api

[4] Pelagios, http://commons.pelagios.org

time period data. "PeriodO is a gazetteer of scholarly definitions of historical, art-historical, and archaeological periods. It eases the task of linking among datasets that define periods differently. It also helps scholars and students see where period definitions overlap or diverge" (PeriodO)[5]. A 'period' is of course a complex archaeological concept, which can encompass timespans, spatial extents and sometimes cultural forms (Niccolucci & Hermon 2015). PeriodO defines a data model that includes a name for the period, some temporal bounds, an association with a geographical region and which has some literary warrant (Shaw 2015; Shaw *et al*. 2015). The PeriodO Client[6] allows users to browse the PeriodO dataset of periods, compare different period definitions, extract the URI for a period, extract semantically structured data (JSON-LD), submit new periods or update existing ones the user has previously entered. ARIADNE partners expressed temporal metadata for archaeological periods using local vocabularies, with start and end dates for each term. The unified list of ARIADNE period vocabularies was made available to PeriodO in a collaboration with that project. This resulted in the ARIADNE collection of period definitions in PeriodO[7], where URIs identify each period and distinguish the meaning of a period name in different places.

2.3 Semantic linking in the subject dimension

Much of the effort of WP15 has concerned semantic linking in the subject domain. Mapping between subject thesauri and other controlled vocabularies is an important contributor to interoperability and cross search in disciplines with diverse datasets, such as archaeology, and is particularly important for multilingual capability (see the review in Zeng & Chan 2004). ARIADNE D15.1 gives an overview of vocabulary mapping drawing on the ISO thesaurus standard (ISO 25964-2:2013), discussing the rationale of vocabulary mapping within ARIADNE and the choice of the Getty Art and Architecture Thesaurus (AAT)[8] as a mapping hub. The AAT is available as Linked Open Data in SKOS, published under the Open Data Commons Attribution License (ODC-By) 1.0[9]. The AAT contains over 40,000 concepts and over 350,000 terms, organised in seven facets (and 33 hierarchies as subdivisions): Associated concepts, Physical attributes, Styles and periods, Agents, Activities, Materials, Objects and optional facets for time and place (Harpring 2016). The AAT's scope is broader than archaeology, encompassing visual art, architecture, other material heritage, archaeology, conservation, archival materials, etc., but contains many useful high level archaeological concepts, particularly in the Built Environment, Materials and Objects hierarchies.

Different partner subject vocabularies are mapped to the AAT, which then forms the basis for search across indexing vocabularies in different languages. The major exercise mapping between partner native vocabularies and the AAT as a central hub is described in D15.1, including reflections on the mapping process. The creation of links directly between the items from different vocabularies can become unmanageable as the number of vocabularies increases. A scalable solution is to employ a hub architecture, an intermediate structure on to which concepts from ARIADNE data provider source vocabularies can be mapped. The benefits of vocabulary mapping are clear for collections with multilingual metadata and content. By using the AAT as a central hub, a semantic search on an AAT concept can retrieve results originally indexed by terms from various languages (Binding and

[5] PeriodO, http://perio.do, https://github.com/periodo

[6] PeriodO client, http://n2t.net/ark:/99152/p0

[7] ARIADNE collection of period definitions in PeriodO, http://n2t.net/ark:/99152/p0qhb66

[8] Getty Art and Architecture Thesaurus, http://www.getty.edu/research/tools/vocabularies/aat/

[9] Getty Vocabularies as Linked Open Data, http://www.getty.edu/research/tools/vocabularies/lod/index.html

Tudhope 2016). This can potentially improve recall (indexing in several languages) and precision (in literal string search, false results may arise from homographs).

2.4 Semantic linking tools in the subject dimension

ARIADNE D15.3 is concerned with the tools employed for semantic linking. While some very specific links between specific datasets, or specific cross references between reports, will exist, most semantic linking in the subject dimension is achieved via the core ontology (CIDOC CRM) and the Getty AAT, which have been adopted as ARIADNE linking standards. Before describing the tools employed for mapping vocabularies in ARIADNE, we briefly review other relevant vocabulary mapping tools.

2.4.1 Vocabulary mapping tools relevant to WP15

The VocBench publishing platform of the UN's FAO has seen a major SKOS Linked Data effort in the agricultural domain, mapping the multilingual AGROVOC thesaurus to (at the last count) 13 other thesauri (Caracciolo *et al*. 2012 and 2013), including LCSH (Library of Congress Subject Headings), GEMET (General Multilingual Environmental Thesaurus) and STW (Standard Thesaurus for Economics / Standard Thesaurus für Wirtschaft). This has been a long term project; early work is discussed by Liang and Sini (2006).

In ARIADNE, it is possible to associate AAT concepts directly with partner subject metadata by taking account of mappings from partner vocabularies to the AAT created for ARIADNE. In addition to semantic linking by means of mapping between vocabularies, it is sometimes necessary to associate concepts from controlled vocabularies (via URI Uniform Resource Identifiers) with subject metadata presented only as free text. For example, in the case of a project like Europeana, which sometimes has to ingest subject metadata without knowledge of the original native vocabulary (if any), it is sometimes desirable to associate a controlled vocabulary concept (e.g. from the AAT) with potentially ambiguous, free text metadata, without being able to rely on a previous vocabulary mapping. All that may be available on the source side is the free text subject label without any further context to take into account from a source vocabulary.

In both these situations, the process of associating a controlled vocabulary concept is called *semantic enrichment*. Enrichment can be considered a special case of semantic annotation. For recent developments in Europeana of semi-automatic 'enrichment' of free text metadata elements with AAT URIs, see Charles *et al*. (2014). Stiller *et al*. (2014) report on a Europeana case study of enrichment with discussion of best practice. A Europeana Data Model case study describes an example of the enrichment of Europeana data with AAT (Charles and Devarenne 2014). This can bring benefits for multilingual capability, for example by automatically changing the subject metadata label displayed if the language of the user interface is changed by the user.

A specific MORe enrichment service for deriving AAT concepts from partner vocabularies was developed as part of ARIADNE utilising the AAT vocabulary mappings produced in WP15 (see 'Vocabulary mappings conducted for ARIADNE' below and ARIADNE D15.1 - Report on Thesauri and Taxonomies). Various enrichment services have been developed within the MORe aggregation framework for the LoCloud project (Meghini *et al*., forthcoming). These include:

- Geocoding based on GeoNames;
- Semantic enrichment, where a vocabulary matching service suggests SKOS concepts based on title, descriptions and subject-related information in a metadata record;

- Wikipedia and DBPedia automatic enrichment: a variation of the semantic enrichment that suggests Wikipedia and DBPedia entries, based on the metadata record;
- Language identification: Identifies languages based on a title or description using Apache Tika;
- Thesauri mappings: Allows loading and managing SKOS concepts mappings from SKOSified subject terms to a target SKOS thesaurus.

The Instance Matching application (SAIM[10]) is a browser based interface intended to support instance matching on RDF data, generally though this is somewhat outside the focus of the WP15 work. Consideration was also given to the FP7 supported Linked Data Integration Framework (SILK[11]). SILK is a sophisticated, general semantic web framework for creating links and not designed specifically for mapping thesauri. A SILK 'link specification' allows the comparison of preferred labels from two thesauri (using the Levenshtein distance algorithm). SILK is a complex environment and there is an inevitable learning curve in order to create linkage rules, transformations, aggregations and in the interpretation of match scores. While the SILK application is interactive and allows for some comparison, it can be argued that the main focus of such tools tends to be automatic link generation functionality and bulk automation. The user is not given convenient contextual information (specific to the thesauri) to facilitate informed judgments on the correctness of potential mappings.

Two archaeological domain presentations at the 14th European Networked Knowledge Organization Systems (NKOS) Workshop, at TPDL 2015 in Poznań discussed work on SKOS vocabulary mapping using AMALGAME (Amsterdam Alignment Generation Metatool[12]), developed at Free University of Amsterdam within the ClioPatria framework. Stahn (2015) made a pilot study with three DAI vocabularies, expressing them as SKOS and then mapping via AMALGAME. She makes the point that the matching is on string level only. Kempf (2015) conducted experiments at ZBW (Leibniz Information Centre for Economics) involving the German STW Thesaurus for Economics and an exact language dependent string match on terms. A second phase explored different types of enrichment of the thesaurus data before the mapping process, such as descriptors, synonyms, results from other mappings. The conclusions were that string match could not take account of structural differences between the source and target vocabulary contexts but that the experiment enriching the vocabulary terminology before attempting mapping offered useful potential candidates for further intellectual mapping. AMALGAME gives a choice of similarity metrics and properties to match on. However the process is then based on automatic string match (although the user does decide whether to accept a mapping) and does not take account of the hierarchical context. The work with AMALGAME is continuing under the new project Cultuurlink[13], the Dutch Cultural Heritage Hub, where an alignment service is offered for SKOS thesauri and various thesauri are available. A complex, interactive alignment strategy graphical editor is available. This employs the AMALGAME mapping system and a filter is available that takes some account of hierarchical context.

[10] SAIM, http://saim.aksw.org

[11] SILK, http://wifo5-03.informatik.uni-mannheim.de/bizer/silk

[12] AMALGAME, http://semanticweb.cs.vu.nl/amalgame/

[13] Cultuurlink, http://cultuurlink.beeldengeluid.nl/app/#/start

2.4.2 Vocabulary mapping tools in ARIADNE

Vocabulary mapping is not a trivial exercise. High quality mapping requires domain experts, who may not have expertise in computing semantic technologies. The vocabularies themselves can vary from a small number of keywords from a picklist for a particular dataset to standard national vocabularies with a large number of concepts. In general at the present time, fully automatic vocabulary mapping does not currently deliver mappings of sufficient operational quality; the ARIADNE approach is to focus on quality mappings, in order to support cross search at a medium level of generality. Archaeology domain vocabulary providers tend to have limited resources to spend on the technical aspects of vocabulary mapping tools. It is important to support expert intellectual review and provide enough contextual data to allow vocabulary providers to make an informed decision on the proposed mappings. For example, it might be useful for the user to compare thesaurus structures side by side.

In general for ARIADNE purposes, tools should be light weight, user centered and focused on thesaurus to thesaurus mapping, in order to facilitate user (vocabulary provider) involvement in the mapping process. In some cases, ARIADNE partners drew on their own resources to create the vocabulary mappings to AAT, for example by directly browsing and searching the local vocabulary and the AAT or using local utilities (see ARIADNE D15.1). In addition, two light weight tools were developed by USW to support the domain experts doing the mapping between vocabulary concepts, oriented to different contexts for the vocabularies.

An interactive vocabulary mapping tool was developed for ARIADNE that enables archaeology subject experts create SKOS mapping relationships between local vocabularies available online as SKOS and the AAT (Binding & Tudhope 2016). The lightweight Web application presents concepts from chosen source and target vocabularies side by side, thus presenting context to allow an informed choice when deciding on potential mappings. The tool is designed for vocabularies expressed in RDF/SKOS; it queries external SPARQL endpoints rather than storing local copies of the vocabularies. SKOS mapping relationships (Miles & Bechofer 2009) are employed for expressing the mappings, (e.g. skos:broadMatch or skos:closeMatch). The resulting set of mappings can be saved locally, reloaded and exported to a number of different output formats (JSON for use in ARIADNE).

Figure 1 shows an example where a source concept, *castle*, has just been mapped to the AAT concept, *castles* (*fortification*). The right hand pane shows various AAT concepts that potentially match on the preferred term. The user has selected the AAT concept, *castles* (*fortification*) and both the broader concept (*fortifications*) and the narrower concepts (*châtelets*, *moated castles*, *motte-and-bailey castles*, *qasrs*) are displayed, showing the semantic context in a concise manner. Each of these additional concepts can themselves be browsed in the tool to explore the semantic context. The mapping has been considered a *close match* by the user and the resulting mapping is displayed at the bottom of the screen, together with two other recent mappings. The tool is available as open source with the code freely available on GitHub[14].

[14] Vocabulary Matching Tool source code for local download and installation, https://github.com/cbinding/VocabularyMatchingTool

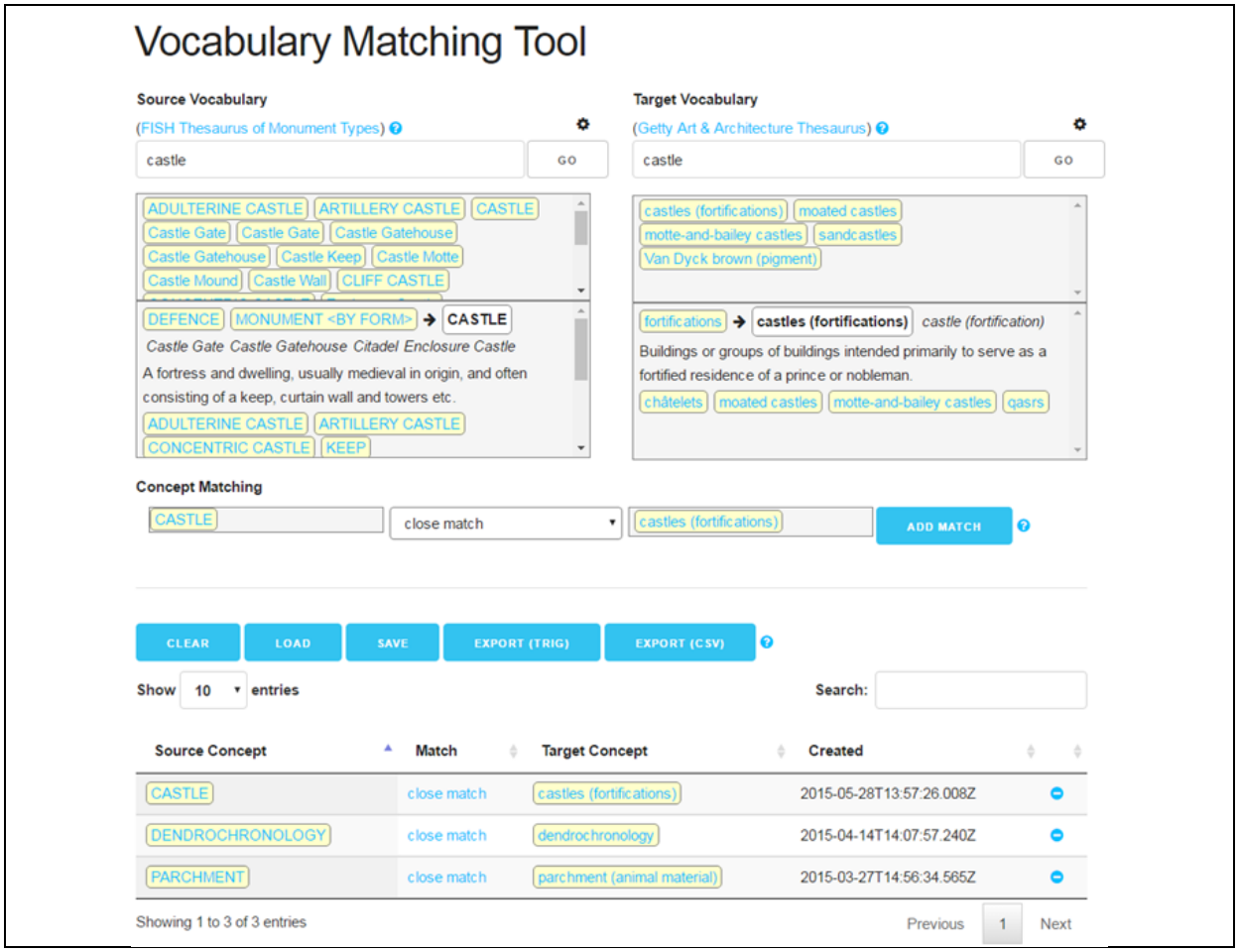

Figure 1: Vocabulary matching tool

A second mapping approach was developed for source vocabularies that are smaller term lists and not yet expressed in RDF. Such term lists are often available or easily represented in a spreadsheet. A standard template with example mappings was designed to support domain experts in the intellectual mapping of terms to the target vocabulary, together with the XSLT conversion of the spreadsheet data to JSON and NTriples formats for upload to ARIADNE. The final template is in part a generalisation of the mapping spreadsheets created by DANS and SND for their mapping work. A CSV transformation produces the representation of the mappings in RDF/JSON format and is available as open source[15]. The user completes the mapping template with the set of source and target (AAT) concepts. The XSLT transformation expresses the mappings in JSON and NTriples formats. The mapping template contained a tab to record metadata for the mapping. In future work, making the mappings available as outcomes in their own right, with appropriate metadata for the mappings would be desirable.

The standard spreadsheet was accompanied by a set of guidelines informed by a pilot exercise (with additional support from the vocabulary team on problematic mappings or existing precedents). In some cases, data cleansing was required before the mapping exercise could proceed. In fact, data cleansing was necessary at various stages of WP15 work. The main tool used for data cleansing by USW was the open source tool OpenRefine (formely GoogleRefine).

[15] ARIADNE subject mappings: Spreadsheet template and conversion, https://github.com/cbinding/ARIADNE-subject-mappings

2.4.3 Vocabulary mappings conducted for ARIADNE

Figure 2 summarises the mappings produced. In total 6416 mappings were conducted, with mappings by individual partners ranging from a few to over 1600 terms. By December 2016, concepts from 27 vocabularies employed by 12 project partners have been mapped to the AAT; 17 of the vocabulary mappings were conducted with the spreadsheet template (or a similar partner spreadsheet), 2 using the online interactive mapping tool (i.e. when the source vocabulary was available in RDF/SKOS) and 8 using the partner's own (intellectual/manual) resources.

		Match type						
Source	Vocabularies mapped to AAT	No match	skos:exactMatch	skos:closeMatch	skos:broadMatch	skos:narrowMatch	skos:relatedMatch	Totals
ADS	FISH Archaeological Objects Thesaurus (subset)	0	197	96	118	0	0	411
ADS	FISH Building Materials Thesaurus (subset)	0	8	4	0	0	0	12
ADS	FISH Thesaurus of Monument Types (subset)	0	141	107	141	0	1	390
ADS	Historic England Components Thesaurus (subset)	0	7	1	1	0	0	9
ADS	Historic England Maritime Craft Thesaurus (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
Discovery	National Inventory of Architectural Heritage	0	211	59	78	7	0	355
IACA	FASTI Monument Types	2	23	80	24	0	0	129
ICCU	ICCD RA and PICO thesaurus (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
NIAM-BAS	AIS AKB database - subject terms	0	50	85	92	11	0	238
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	**2**	**0**	**379**	**128**	**327**	**7**	**0**	**841**
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**
NIAM-BAS	**1**	**0**	**50**	**85**	**92**	**11**	**0**	**238**
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:	**27**	**10**	**3176**	**1172**	**1726**	**316**	**16**	**6416**
	Proportion:	**0.16%**	**49.50%**	**18.27%**	**26.90%**	**4.93%**	**0.25%**	**100%**

Figure 2: Mappings conducted as part of WP15

The mappings were converted by USW to JSON format for use in the ARIADNE Registry to help augment partner subject metadata with appropriate AAT concepts. The ARIADNE data catalogue

employs the MoRe (Metadata & Object Repository) aggregator[16] to harvest the metadata provided by the project partners utilising the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). A bespoke AAT subject enrichment service has been developed by DCU that imports the partner vocabulary mappings and applies them to the partner subject metadata, deriving an AAT concept (both preferred label and URI) to augment the existing subject metadata in the data catalogue. The catalogue metadata is supplied to the ARIADNE portal, where the search functionality can retrieve multilingual records related via the AAT concepts.

2.4.4 Vocabulary mappings within the ARIADNE portal

The enriched AAT concepts afford a multilingual capability in the ARIADNE Portal, allowing a search on an AAT concept to return results from subject metadata in different languages.

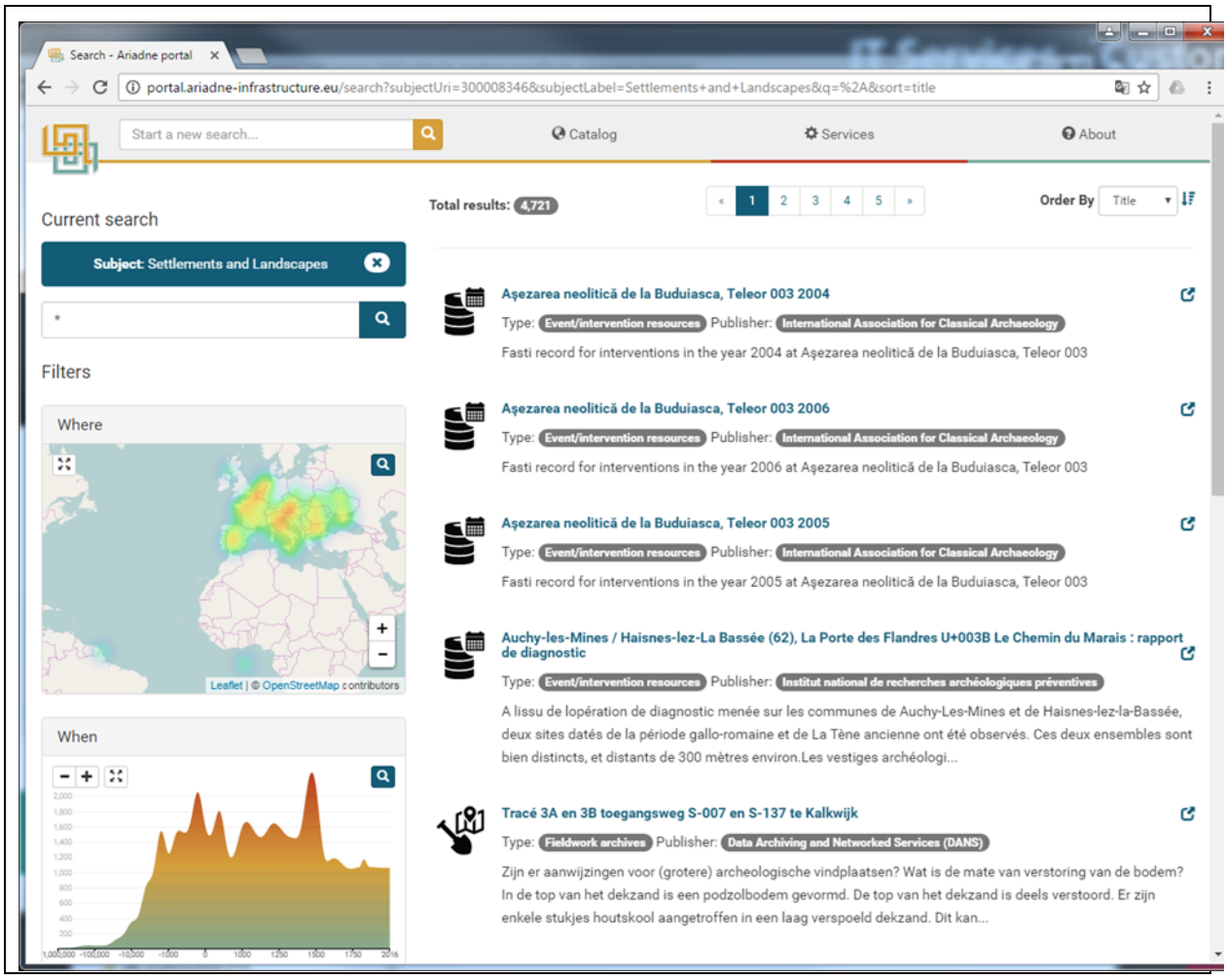

Figure 3: Portal Query on AAT subject: *Settlements and Landscapes* showing results from AIAC (Fasti Online), INRAP and DANS, with multiple languages (December 2016).

Figure 3 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, named Subject, is populated by the MoRe enrichment process described above; effectively the Subjects are AAT concepts, which have been mapped from the native vocabulary subject metadata of the data

[16] MoRe (Metadata & Object Repository) aggregator, http://more.dcu.gr

resources in the Portal. Figure 3 shows that a simple query on the single AAT concept, Settlements and Landscapes, is able to retrieve results in multiple languages for records originating from AIAC (Fasti), INRAP and DANS.

In addition to semantic linking within the metadata of the ARIADNE Portal, work in WP15 also explored semantic linking at the more detailed level of individual items within ARIADNE datasets. Extracts from the datasets were additionally linked to archaeological reports as a result of NLP semantic annotation work performed as part of WP16. The semantic framework for the linking was the combined CIDOC CRM and Getty AAT. The following section discusses a case study that investigated this approach and the resulting Demonstrator.

2.5 Wood/Dendrochronology case study on item level data integration

A joint WP15 and WP16 effort conducted an exploratory investigation of the semantic integration of extracts from archaeological datasets with information extracted via NLP (Natural Language Processing) across different languages. The case study[17] was based on a loose theme of archaeological interest in wooden objects and their dating via dendrochronological techniques. The work was undertaken by USW on the technical side, in collaboration with DANS and SND as regards Dutch and Swedish archaeological datasets, reports and vocabularies.

The case study investigates whether it is possible to achieve a degree of semantic interoperability between archaeological datasets and data derived from applying NLP information extraction techniques to the textual content of grey literature reports in different languages. The semantic framework combined the CIDOC CRM with the Getty AAT. The case study has a broad theme relating to wooden material including shipwrecks, with a focus on indications of types of wooden material, samples taken, wooden objects with dating from dendrochronological analysis, etc. The resources comprise five English and Dutch language datasets and grey literature reports, together with Swedish archaeological reports. ADS datasets included two shipwreck datasets, the Newport Medieval Ship[18] and the Mystery Wreck Project (Flower of Ugie)[19], together with the Vernacular Architecture Group dendrochronology[20] and Cruck[21] databases. DANS facilitated an extract from the database of the international Digital Collaboratory for Cultural Dendrochronology (DCCD[22]).

The CIDOC CRM was used to connect the data elements and the NLP entities, which include Object, Sample, Material, Place (in some cases), date ranges. A spine vocabulary was identified from the AAT hierarchies for Material and Objects. Mappings from Dutch terms to AAT concepts mostly existed already from WP15 mapping work, while mappings from Swedish terms to AAT concepts were produced by SND, as part of their WP16 effort.

The NLP focus was on concepts relevant to the theme, such as samples, materials, objects and temporal information. The NLP techniques were able to generate XML output from English, Dutch and Swedish texts, which was then transformed to the same RDF format as the instance data

[17] For a discussion of other ARIADNE case studies of item level integration, see ARIADNE D14.2 Pilot Deployment Experiments

[18] http://dx.doi.org/10.5284/1020898

[19] http://dx.doi.org/10.5284/1011899

[20] http://dx.doi.org/10.5284/1039454

[21] http://dx.doi.org/10.5284/1031497

[22] http://dendro.dans.knaw.nl

extracted and mapped to the CRM/AAT. NLP derived RDF statements do not necessarily carry the same degree of reliability as those derived from the datasets; the Dutch and Swedish NLP pipelines are at a prototype stage. The NLP outcomes include some false positives. See ARIADNE D16.4 (Second Report on Natural Language Processing) for more details on the NLP techniques, a discussion of limitations and further work. In future work, some indication of the provenance of the RDF data could be included in the CRM model, which would allow informed judgments of the reliability of the information.

Data cleaning was necessary and achieved via OpenRefine[23]. The output from GATE was expressed as XML and was converted to RDF conforming to the CIDOC CRM with connections also made to the AAT. This was achieved by means of a mapping/extraction tool, STELETO, developed by USW for ARIADNE and freely available as open source[24]. STELETO is a 'lite' cross platform version of the STELLAR.CONSOLE application developed for the STELLAR Project[25] (Binding *et al*. 2015). It is a general delimited text data conversion utility, with simpler command line functionality and efficient performance. STELETO converts tabular input data to any textual output format via a custom (user-defined) template. It was used by USW to convert the case study data (from both dataset and text streams) so that it conformed to the semantic framework of the CIDOC CRM and AAT. In total, 1.09 million RDF triples resulted from 23,594 records with 37,935 objects.

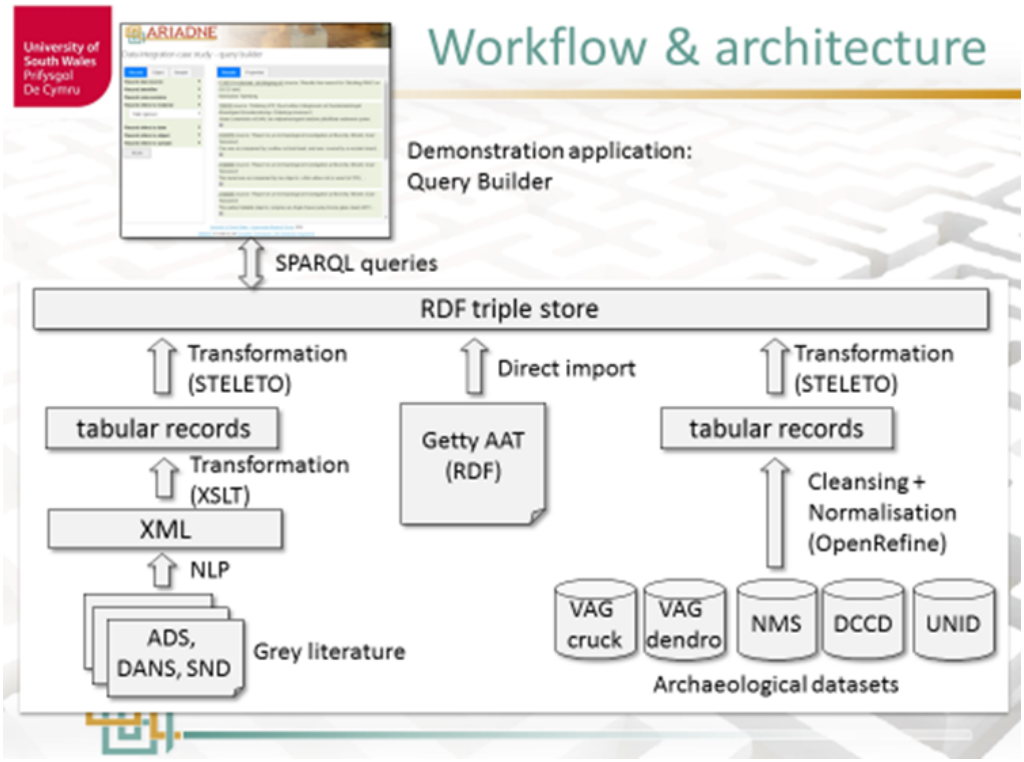

Figure 4: Demonstrator architecture and workflow

The overall architecture and workflow is shown in Figure 4. The research Demonstrator cross searches the data extracted from datasets and Dutch, English and Swedish text reports via SPARQL queries. The Demonstrator is a SPARQL query builder, developed by USW, that seeks to hide the

[23] OpenRefine, http://openrefine.org

[24] STELETO, https://github.com/cbinding/STELETO

[25] STELLAR Project, http://hypermedia.research.southwales.ac.uk/kos/stellar/

complexity of the underlying ontology. This Web application demonstrates the potential for alternative user interfaces to a plain SPARQL endpoint building on techniques developed in the SENESCHAL project[26]. As the user selects from the interface, an underlying SPARQL query is automatically constructed in terms of the corresponding ontological entities. It is possible to search across all datasets (the default) or select a dataset to search individually. A set of interactive controls offer search and browsing of the extracted archaeological data. The controls are designed to be browser agnostic and the Demonstrator will run in most modern internet browsers.

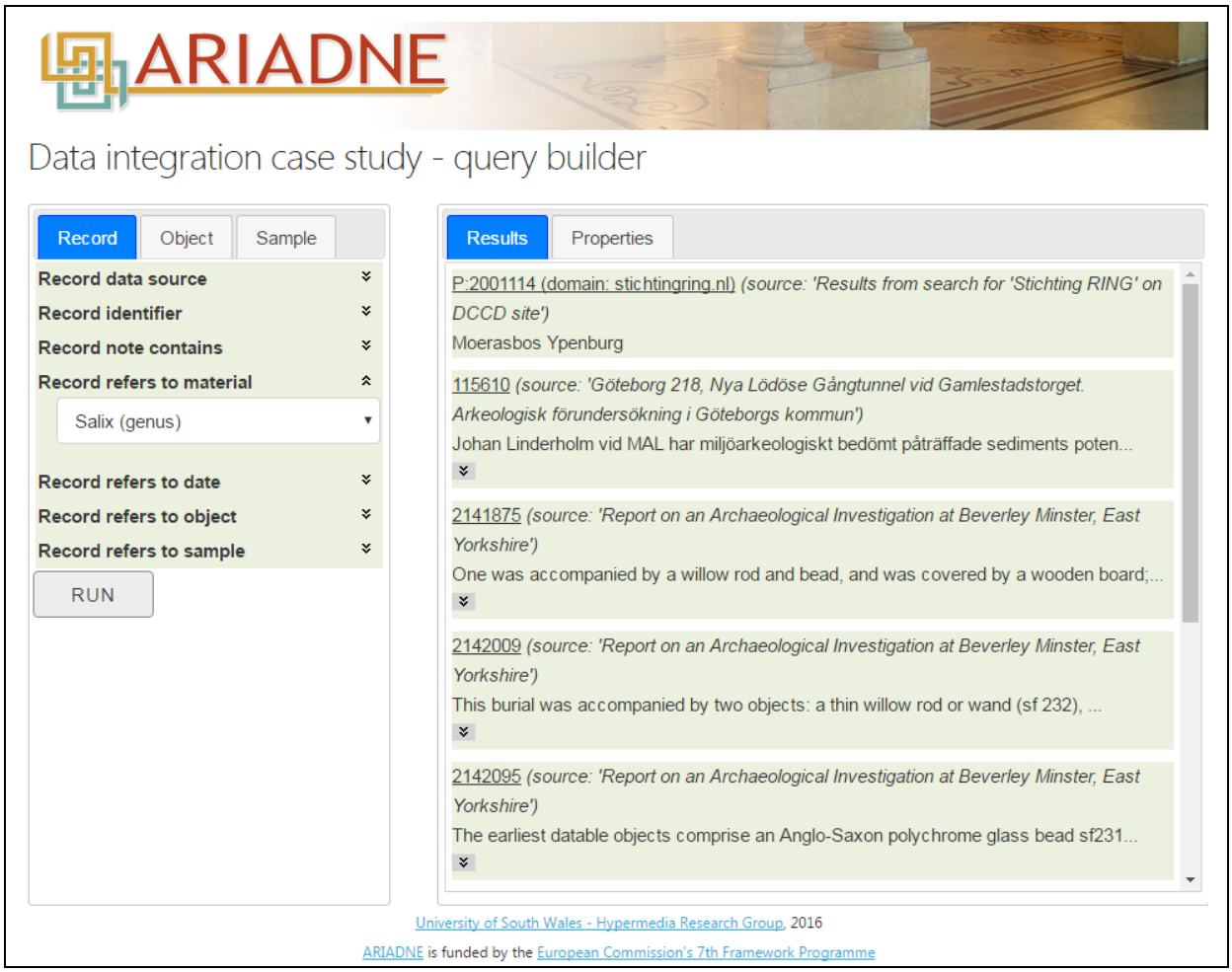

Figure 5: Data integration case study demonstrator application

The demonstrator can perform semantically structured queries, free-text queries, or a combination of both. Drop-down lists of all datasets, AAT materials and AAT object types used in the data are populated at startup, and a dual slider control is initialized to represent the minimum and maximum years for any object production dates present in the data. This provides useful selectable options to assist query formulation. Hierarchical expansion has been implemented over the semantic structure of the Getty AAT and results from narrower concepts are included when available.

Figure 5 illustrates the case study query builder application, performing a structured query on records linked to the AAT concept "Salix (genus)". The multilingual results originate from Dutch, Swedish and English records, generated from databases and textual reports. Some of the results shown are referring to the AAT concept "willow" – this is because the application is leveraging the

[26] SENESCHAL Project, http://hypermedia.research.southwales.ac.uk/kos/SENESCHAL/

underlying AAT structure to automatically expand the original query across relevant semantic relationships, so improving recall without sacrificing precision.

Figure 6 shows a query on objects declared as being of type "roofs", with a production date stated as being somewhere within the date range 1500 to 1600 AD. The results originate from data derived via NLP on a textual report, indicating some instances where the process has associated an object with a date (or a date range). Examination of the textual notes associated with each object suggest that the NLP results are correct.

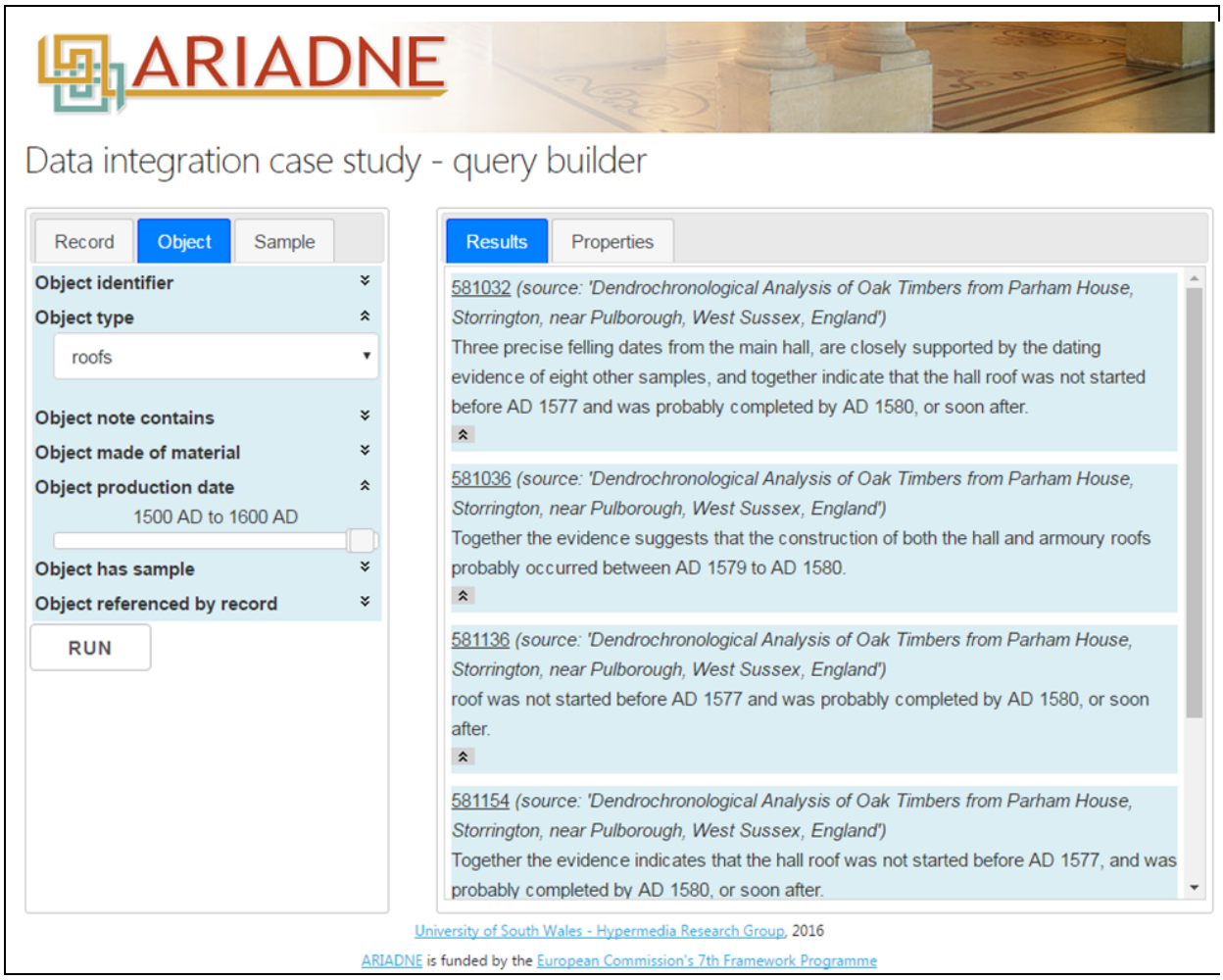

Figure 6: Search on specific object type and date range

The Demonstrator is able to cross search and browse information extracted from datasets and reports in different languages via the common framework based on the CIDOC CRM and AAT; a degree of semantic integration between data extracted from text documents and databases has been achieved. In future work, the intention is to explore possibilities of the semantic techniques for larger scale efforts on multilingual integration of datasets with reports. Another aim is to explore possibilities for more intuitive user interfaces for searching RDF datasets than the usual SPARQL endpoint. The Demonstrator is available for use via the ARIADNE Portal services[27]. It is also available as open source with the code freely available on GitHub[28].

[27] Wood/Dendrochronology Demonstrator, http://portal.ariadne-infrastructure.eu/services

[28] Wood/Dendrochronology Demonstrator code, https://github.com/cbinding/ARIADNE-data-integration

3 Summary and lessons learned

Brief summary

The review and development work described in this chapter focuses on the aspects of semantic linking and annotation particularly relevant to ARIADNE. The nature of semantic linking and annotation within the ARADNE context is discussed and reference is made to other relevant ARIADNE deliverables. Semantic linking within ARIADNE is considered within the spatial, temporal and subject dimensions. The subject dimension is considered in depth, starting with a review of linking tools considered relevant to ARIADNE followed by a discussion of the ARIADNE approach and the vocabulary mapping tools used within ARIADNE. The Getty AAT proved an appropriate vocabulary mapping hub that afforded a multilingual search capability in the ARIADNE Portal via the semantic enrichment of partner subject metadata with derived AAT concepts. The characteristics of the different types of mappings of partner vocabularies to the Getty AAT produced within ARIADNE are described. In total 6416 mappings were conducted, with mappings by individual partners ranging from a few to over 1600 terms. By December 2016, concepts from 27 vocabularies employed by 12 project partners have been mapped to the AAT. The vocabulary mapping tools developed for ARIADNE are available as open source via Github.

A case study conducted an exploratory investigation of the semantic integration of extracts from archaeological datasets with information extracted via NLP across different languages. The investigation followed a broad theme relating to wooden material including shipwrecks, with a focus on types of wooden material, samples taken, wooden objects with dating from dendrochronological analysis, etc. The Demonstrator developed for this wood/dendrochronological case study is described with illustrative screendumps; it is available for general use. The user is shielded from the complexity of the underlying semantic framework (based on the CIDOC CRM and Getty AAT) by the Web application user interface. The Demonstrator highlights the potential for archaeological research that can interrogate grey literature reports in conjunction with datasets. Queries concern wooden objects (e.g. samples of beech wood keels), optionally from a given date range, with automatic expansion over hierar-chies of wood types.

Lessons learned

- The spatial, temporal and subject dimensions are key to archaeology and the main vehicle for semantic linking. Cleansing and normalisation via semantic enrichment is necessary in each dimension.
- In the subject dimension, ARIADNE requires high quality vocabulary mappings, which cannot be achieved by automatic means alone, although appropriate mapping tools can help. It is important to support expert intellectual review and provide contextual data.
- The Getty AAT proved an appropriate vocabulary mapping hub that afforded a multilingual search capability in the ARIADNE Portal.
- The Wood/Dendrochronology case study demonstrated the feasibility of connecting information (at a detailed level) extracted from datasets and grey literature reports in different languages and semantic cross-searching of the integrated information. The case study suggests that the semantic linking of textual reports and datasets opens up possibilities for integrative archaeological research across diverse resources.

- Data cleansing is necessary for semantic data integration together with tools that ensure a consistent representation of the extracted data in terms of the semantic framework.
- NLP derived RDF statements do not carry the same degree of reliability as those derived from the datasets and NLP outcomes may include some false positives. An indication of the provenance of the RDF data should be included in the semantic framework, which would allow judgments of the reliability of the information.
- The Wood/Dendrochronology demonstrator shows that a Web application can hide the complexity of the underlying semantic framework from the user and allow querying and browsing the information without expertise in SPARQL; it illustrates that more intuitive user interfaces are possible for searching RDF datasets than the usual SPARQL endpoint or browsing hyperlinks.

4 References

AMALGAME, http://semanticweb.cs.vu.nl/amalgame/

ARIADNE (2015) website: ARIADNE at Linked Pasts, http://www.ariadne-infrastructure.eu/News/ARIADNE-at-Linked-Pasts

Binding C. & Tudhope D. (2016): Improving Interoperability using Vocabulary Linked Data. In: International Journal on Digital Libraries, 17(1), 5-21 [doi:10.1007/s00799-015-0166-y]

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

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

Charles V., Freire N. & Isaac A. (2014): Links, languages and semantics: linked data approaches in the European Library and Europeana. Proceedings IFLA World Library and Information Congress 2014, Lyon, http://ifla2014-satdata.bnf.fr/pdf/iflalld2014_submission_Charles_Freire_Isaac.pdf

CULTUURLINK, http://cultuurlink.beeldengeluid.nl/app/#/start

Heritage Data. Linked Data Vocabularies for Cultural Heritage, http://www.heritagedata.org

Isaac A., Waites W., Young J. & Zeng M. (eds.) (2011): Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets [W3C Incubator Group Report, October 25, 2011]. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/

ISO 25964-2:2013. Information and documentation - Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies. http://www.niso.org/schemas/iso25964/#part2

ISO25964-1:2011. Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. http://www.niso.org/schemas/iso25964/#part1

Kempf A., Neubert J., Faden M. (2015): The Missing Link – A Vocabulary Mapping Effort in Economics. 14th European Networked Knowledge Organization Systems (NKOS) Workshop, TPDL 2015 Poznań.

Liang A. & Sini M. (2006): Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. In: New Review of Hypermedia and Multimedia, 12(1): 51-62.

Meghini C. *et al*. (forthcoming): ARIADNE: A Research Infrastructure for Archaeology. In: Journal on Computing and Cultural Heritage.

Miles A. & Bechofer S. (2009): SKOS Simple Knowledge Organization System Reference. W3C Recommendation, https://www.w3.org/TR/2009/REC-skos-reference-20090818/#L4138

Niccolucci F. & Hermon S. (2015): Representing gazetteers and period thesauri in four-dimensional spacetime. In: International Journal on Digital Libraries, 17(1): 63-69 [doi:10.1007/s00799-015-0159-x]

Open Annotation Ontology, http://www.openannotation.org

OpenRefine, http://openrefine.org

Pelagios, http://commons.pelagios.org

PeriodO client, http://n2t.net/ark:/99152/p0

PeriodO website, http://perio.do

PeriodO (Github), https://github.com/periodo

Peripleo, https://github.com/pelagios/peripleo#peripleo-api

Pleiades, https://pleiades.stoa.org

Recogito, http://recogito.pelagios.org

Richards J., Tudhope D. & Vlachidis A. (2015): Text Mining in Archaeology: Extracting Information from Archaeological Reports, pp. 240-254, in: Barcelo J. & Bogdanovic I. (eds.): Mathematics and Archaeology. CRC Press

SAIM Instance Matching Application, http://saim.aksw.org

SENESCHAL project: Semantic ENrichment Enabling Sustainability of arCHAeological Links. University of South Wales, Hypermedia Research Group, http://hypermedia.research.southwales.ac.uk/kos/seneschal/

Shaw R., Rabinowitz A., Golden P. & Kansa E. (2015): A sharing-oriented design strategy for Networked Knowledge Organization Systems. International Journal on Libraries, 17(1): 49-61 [doi:10.1007/s00799-015-0164-0]

Shaw R., Rabinowitz A., Golden P. & Kansa E. (2015): Report on and demonstration of the PeriodO period gazetteer. 14th European Networked Knowledge Organization Systems (NKOS) Workshop, TPDL 2015 Poznań.

SILK - Silk Link Discovery Framework, http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

Simon R., Isaksen L., Barker E. & de Soto Cañamares P. (2016): Peripleo: a Tool for Exploring Heterogenous Data through the Dimensions of Space and Time. In: Code4Lib Issue 31, 2016-01-28, http://journal.code4lib.org/articles/11144

Simon R., Isaksen L., Barker E. & de Soto Cañamares P. (forthcoming): The Pleiades Gazetteer and the Pelagios Project, in: Berman M., Mostern R. & Southall H. (eds.): Placing Names: Enriching and Integrating Gazetteers. Indiana University Press.

SKOS mapping relationships, http://www.w3.org/TR/skos-reference/#L4138

SKOS. Simple Knowledge Organization System. W3C, http://www.w3.org/2004/02/skos/

SPARQL 1.1 Query Language. W3C. (2013), http://www.w3.org/TR/sparql11-query/

Stahn L. (2015). Vocabulary Alignment for archaeological Knowledge Organization Systems. 14th Workshop on Networked Knowledge Organization Systems, TPDL 2015 Poznań.

STAR Project: Semantic Technologies for Archaeological Resources. University of South Wales: Hypermedia Research Group, http://hypermedia.research.southwales.ac.uk/kos/star/

STELETO, http://github.com/cbinding/STELETO/

STELLAR Project: Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources. University of South Wales: Hypermedia Research Group, http://hypermedia.research.southwales.ac.uk/kos/STELLAR/

Stiller J., Petras V., Gäde M. & Isaac A. (2014): Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences, pp. 238-247, in: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. Springer LNCS 8740.

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