Author accepted version
Accepted for publication in Journal of Information Science
SAGE Publishing. ISSN: 01655515
http://journals.sagepub.com/home/jis

Binding C, Tudhope D, Vlachidis A. A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science. In Press. Reprinted by permission of SAGE Publications.

A study of semantic integration across archaeological data and reports in different languages

Ceri Binding, Douglas Tudhope, (Hypermedia Research Group, University of South Wales), Andreas Vlachidis (University of the West of England)

Abstract

This study investigates the semantic integration of data extracted from archaeological datasets with information extracted via NLP across different languages. The investigation follows a broad theme relating to wooden objects and their dating via dendrochronological techniques, including types of wooden material, samples taken, wooden objects including shipwrecks. The outcomes are an integrated RDF dataset coupled with an associated interactive research demonstrator query builder application. The semantic framework combines the CIDOC CRM with the Getty Art and Architecture Thesaurus (AAT).

The NLP, data cleansing and integration methods are described in detail together with illustrative scenarios from the web application Demonstrator. Reflections and recommendations from the study are discussed. The Demonstrator is a novel SPARQL web application, with CRM/AAT based data integration. Functionality includes the combination of free text and semantic search with browsing on semantic links, hierarchical and associative relationship thesaurus query expansion. Queries concern wooden objects (e.g. samples of beech wood keels), optionally from a given date range, with automatic expansion over AAT hierarchies of wood types and specialised associative relationships. Following a 'mapping pattern' approach (via the STELETO tool) ensured validity and consistency of all RDF output. The user is shielded from the complexity of the underlying semantic framework by a query builder user interface. The study demonstrates the feasibility of connecting information extracted from datasets and grey literature reports in different languages and semantic cross-searching of the integrated information. The semantic linking of textual reports and datasets opens new possibilities for integrative research across diverse resources.

1 Introduction

While there is a growing awareness of the benefits to be gained by making research data freely available, the challenges posed for investigators by the isolation and fragmentation of research datasets are well known. Database structure varies and simple differences in table and field format can mislead a search. This is compounded by terminology issues; different words may mean the same thing while the same word can carry different meanings [1] . This is particularly so in archaeology, where a variety of scientific methods are employed and many different excavation recording systems are used. In addition, there are a large number of unpublished grey literature reports resulting from commercial archaeological interventions [2] . Initiatives in different countries have begun to curate these reports in digital libraries. However they are not readily integrated for search purposes with archaeological datasets even though these may be found within the same repository. Meaningful search across data from different institutions is hard to achieve.

"*Given that there is no common schema in use in the archaeological sector and there is extensive variability in the terminology, normal usage of these datasets requires analysis to take place on a site by site basis. Cross-search is extremely limited. Site metadata may allow search at broad location or major time period level. However it is almost impossible to search across datasets directly for, say, examples of a particular type of artefact from a particular period occurring in a particular type of context (e.g. Roman pottery found in early medieval middens). Datasets are increasingly available online but effectively isolated from each other and also with no connection to grey literature (unpublished excavation reports), for example from the ADS digital library. These isolated resources do not support research inquiries that depend on semantic interoperability between differing database structures and terminology, even on such fundamental questions as finding all hearths.*" [3]

This paper reports on a case study, which explores the detailed integration of archaeological reports and datasets in different languages. It investigates the feasibility of semantic interoperability between data extracted from archaeological datasets and data derived from applying natural language processing (NLP) information extraction techniques to grey literature reports. The case study is based on a broad theme of archaeological interest in wooden objects and their dating via dendrochronological techniques, including types of wooden material, samples taken, wooden objects including shipwrecks, dating from dendrochronological analysis. The resources comprise extracts from English and Dutch language datasets together with grey literature archaeological reports in English, Dutch and Swedish languages. The data extracted was transformed to a common interoperable framework and resources were mapped to a common spine subject vocabulary.

The case study builds upon past work by authors on the semantic integration of English language archaeological datasets and grey literature reports (STAR project)[1],

[1] STAR Project - http://hypermedia.research.southwales.ac.uk/kos/star/ (accessed 11 May 2018)

which took some steps towards addressing the issues raised above by Richards and Hardman [3] . A demonstrator Web application showed the capability of supporting search across datasets and information extracted from grey literature reports [4] [5] . A semantic framework for the English language work was provided by the combination of archaeological vocabularies with the CIDOC CRM core ontology (ISO 21127:2014) [6] . The complementary use of controlled vocabularies and ontological structures is suggested where appropriate by the ISO thesaurus standard (section 21) [7] and (as formal metadata and value vocabularies) by the W3C Library Linked Data Incubator Group [8] .

The aim of this case study is to investigate the feasibility of extending these techniques to reports and datasets in different languages, with the ultimate aim of developing tools that can support the investigation of archaeological research questions. The ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) project[2] offered an opportunity to carry this line of research forward. The project provided an e-infrastructure that integrated archaeological datasets and reports from multiple European partners in different languages. An overview of the ARIADNE outcomes is provided by Aloia et al. [9] , which describes the architecture, the underlying data model and semantic framework and the Portal, which provides cross search of the resource discovery metadata.

Within archaeology semantic approaches where both data structures and vocabularies are mapped to common standards based upon a Linked Data framework [10] are seen to offer potential. However, significant challenges and also opportunities remain, including the use of NLP on archaeological reports [11] [3] . The potential for e-research purposes of the under-utilised archaeological grey literature has been recognised in recent years. As part of an initiative to define and prioritise grand challenges for archaeological research, Kintigh [12] highlights the potential of grey literature and the need for natural language processing technologies to extract meaningful information from repositories of archaeological reports. Literal string search is insufficient; addressing research questions requires an ability to extract knowledge. Many of the important questions for archaeology require the ability to deal with reports in more than one language. In the vision set out by Kintigh, machine understanding encompasses the broad sense of a document with the ability to infer implicit knowledge from the document structure to answer complex, faceted queries. This goes beyond current capabilities. This case study takes an initial step by exploring the integration of archaeological reports and data in more than one language.

1.1 Related literature

Sense making practice within archaeological investigation relies upon the practical expertise and experience of the excavation team [13] . Data recording sheets for 'finds' and 'contexts' enable the capture of excavation outcomes in archaeological databases but interpretation (classification of an artefact or feature, assignment of a temporal period) often proceeds in stages and can be subject to revision. Reflexive methodologies have become influential [14] . This has led to the adoption of event-based data modelling approaches within archaeology, where the assignment of an interpretation can be

[2] ARIADNE project http://www.ariadne-infrastructure.eu/ (accessed 11 May 2018)

recorded as an event, allowing potential for further events with other interpretations. For example, Ashley et al. [15] discuss how they employed the event-based CIDOC CRM ontology as a framework in a 'digital mirror' of a more conventional print report on work by Berkeley Archaeologists at the long running Çatalhöyük excavation, influenced by ontological modelling done by English Heritage's Centre for Archaeology [16] . The Berkeley team emphasise the complexity of the mapping process and the need for time-consuming data cleansing with typical archaeological datasets. They note the absence of a "publishing platform that can display a complex and massive content through a friendly interface". They elected to adopt a simpler approach to the CRM class structure by introducing five superclasses for entities in their CIDOC CRM implementation. In our previous English language semantic integration of diverse archaeological datasets and grey literature reports, we also built on the English Heritage model extending the CRM and attempted to hide some of the complexity of the ontology. The Demonstrator Web application provided an archaeological user-friendly interface for a query builder over the RDF data (linked via the CRM and archaeological vocabularies) - various search scenarios are illustrated in [5] . Subsequent work (the STELLAR project[3] and toolkit - see Section 3.6) developed tools and guidelines for third party use, validating them on a different set of UK excavation datasets [4] . Some recent developments aim to impose interoperable semantic structure from the outset at the point of data entry. For example, the Endangered Archaeology in the Middle East & North Africa (EAMENA) project [17] employed the open source ARCHES heritage inventory and management system to build their online resource. ARCHES [18] includes inventory and vocabulary management modules, with a data architecture based around common interoperability standards including the CIDOC CRM.

Kansa et al. [19] advocate a 'data sharing as publication' model to encourage the dissemination and the linking of archaeological datasets via common concepts as Linked Open Data. Their Open Context[4] initiative publishes data and resources from archaeology and related subjects, with review by an editorial board and optional peer review. To date, a relatively simple ontological model has been used to integrate the data. Drawing on experience with Open Context, Faniel et al. [20] investigated archaeologists' experience with data reuse. They argue that in addition to sound linked data procedures, repositories of archaeological data should also provide broader contextual information, relating to data provenance, excavation and analysis methodology, in order to encourage reuse of that data. Other initiatives have focused on spatial or temporal dimensions. The Pelagios initiative makes use of the Pleiades gazetteer[5] (and its URIs) to connect online resources that refer to places in the ancient world via Linked Open Data [21] . 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, with the aim of supporting interoperability while imposing minimal overheads on data providers. In the temporal domain, the PeriodO gazetteer aims to act as a central hub for expressing standard period definitions, in order to link and visualize time period data. PeriodO

[3] Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources (STELLAR project) http://hypermedia.research.southwales.ac.uk/kos/stellar/ (accessed 11 May 2018)
[4] Open Context http://opencontext.org (accessed 11 May 2018)
[5] Pleiades, https://pleiades.stoa.org (accessed 11 May 2018)

defines a data model that includes a name for the period, the temporal bounds, an association with a geographical region on the basis of some literary warrant [22] . ARIADNE project partners expressed temporal metadata for archaeological periods using local vocabularies with start and end dates for each term. The unified list of period vocabularies was represented in PeriodO[6], where URIs identify each period and distinguish the meaning of a period name in different places.

In other application domains, the (UN) FAO's VocBench platform makes available a major linked data effort in the agricultural domain, where the multilingual AGROVOC thesaurus has been mapped to 13 other thesauri [23] . A digital history case study explored the semantic integration of datasets concerning Dutch ships and sailors with resulting linked data [24] . In order to facilitate detailed investigation back to the original data, the datasets were converted to RDF using their own data model and then enriched with links, in order to connect to a common interoperability layer. This built on a previous museum case study that resulted in linked data expression from the Amsterdam Museum [25] . This employed the Europeana Data Model (EDM) as a semantic integration framework, complemented by the Amsterdam Museum thesaurus, which was mapped to the Dutch version of the Getty Art and Architecture Thesaurus (AAT-Ned[7]) in the subject domain (in addition to geographical and person metadata). Other examples of the complementary use of ontologies or formal metadata and value vocabularies include the Europeana cultural heritage portal [26] and the Health Finland prototype [27] .

There have been relatively few studies of information extraction in the archaeology domain. Byrne and Klein [28] investigated the extraction of events in archaeological texts via the identification of verb phrases (and associated event types). Recently Henninger [29] shows the potential for NLP techniques and interoperability standards to enhance the subject metadata of the record of an excavation with information extracted from dig diaries in a case study of the Ness of Brodgar excavation. The Archaeotools project [30] [31] investigated the automatic extraction of various conceptual entities from archaeological grey literature reports, including subject, location and period in order to support what/where/when queries that underlie many archaeological research questions. Rule based approaches were used for regular patterns such as spatial grid references and bibliographies. Machine learning approaches were used for less regular patterns. One issue encountered was the difficulty in distinguishing entities concerning the main focus of a report from cross references to completely different archaeological investigations. The approach adopted was to prioritise entities extracted from the summary of a report if that could be identified or else the first 10% of the text. Negation detection in archaeology has also been explored [32] . The NLP methods employed in this case study build on the English language information extraction techniques developed for STAR, where evaluation delivered competitive results [33] . The grammatical patterns for Relation Extraction were able to extract 'rich phrases' combining CIDOC CRM semantic entities, (via events) such as "*medieval silver*

[6] ARIADNE collection of period definitions in PeriodO, http://n2t.net/ark:/99152/p0qhb66 (accessed 11 May 2018)

[7] http://website.aat-ned.nl/home (accessed 15 May 2018)

coin", "*finds of Roman period*", "*coins dating to AD 350–53*", "*coins belonged to the second half of the 3rd century AD*".

2 Methodology

2.1 Data sources

The multilingual (English, Dutch and Swedish) data sources for the case study originated from the Archaeology Data Service (ADS) [34] , Data Archiving and Networked Services (DANS) [35] and the Swedish National Data Service (SND) [36] . The data included extracts of 5 archaeological datasets, and NLP output from 25 grey literature reports [see Section 2.5 for further details]. In consultation with the ADS, 4 datasets with potential dendrochronology interest were selected, while DANS facilitated an extract from a European dendrochronological database. The data are extracts from these databases for purposes of the case study and should not be regarded as complete. The datasets are:

- Mystery Wreck Project (Flower of Ugie) - Hampshire and Wight Trust for Maritime Archaeology, 2012. http://dx.doi.org/10.5284/1011899 Marine archaeology investigation of material characteristics allowed identification of the wreck as a sailing barque built in 1838.
- Newport Medieval Ship, Newport Museums and Heritage Service, 2014 http://dx.doi.org/10.5284/1020898 The most substantial medieval vessel excavated in UK, finds indicate strong Iberian trading connections
- Dendrochronology Database - Vernacular Architecture Group, 2000 (updated 2015) http://dx.doi.org/10.5284/1039454 Tree-ring dates for over 3700 buildings in UK ranging from cathedrals to cottages.
- Cruck database - Vernacular Architecture Group, 2003 (updated 2015) http://dx.doi.org/10.5284/1031497 Database used to generate the catalogue of cruck (curved timber framed) buildings in the UK, originating as a card index.
- Digital Collaboratory for Cultural Dendrochronology (DCCD) - dendrochronological database http://dendro.dans.knaw.nl/ Digital repository of European tree-ring data of a wide variety of objects, based on the Tree-Ring Data Standard (TRiDaS).

2.2 Workflow and architecture

The general architecture (Figure 1) involved converting all data to populate an integrated RDF triple store, which would then be queried by user interface applications. This necessitated extraction and transformation of data from native formats (grey literature NLP and tabular datasets). Data cleansing was also required to ensure the data was sufficiently normalised for successful integration.

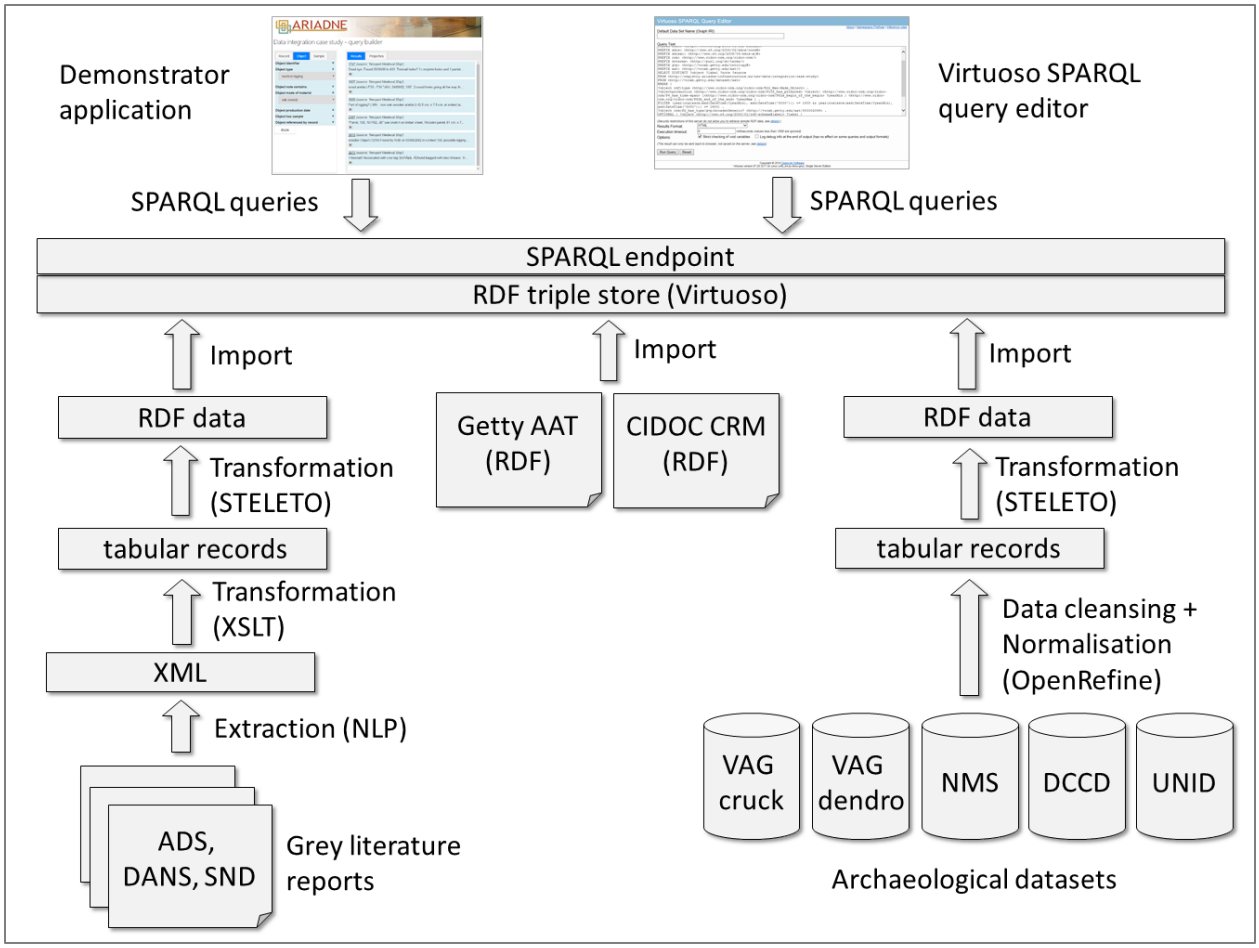

Figure 1 – general workflow and architecture

2.3 Data cleansing

Although datasets originated from multiple sources, cleansing and normalisation processes such as removal of punctuation, consistent capitalisation, whitespace normalisation, splitting of multi-valued cells etc. were commonly applicable to all. This was a detailed and time consuming exercise but without it the semantic alignment of data elements between the datasets would have been less successful; the issue was encountered in ARIADNE generally and is also emphasised by Ashley et al. [15] . The OpenRefine application [37] was used to correct these issues. It also helped in the identification and correction of obvious data anomalies via faceting, clustering, filtering and sorting of column values. As it is important to preserve original data during this process a new column can be created based on existing values that can then be modified without affecting the original, and both the raw and cleaned versions can become separate properties in any subsequent transformation or export of the data.

OpenRefine 'facets' are an aggregated listing of unique data values to expose (and fix) obvious anomalies. The facet example on the left of Figure 2 shows some textual values containing question mark suffixes, sometimes encountered in datasets as an indicator of uncertainty (best practice would have required a separate field for this). Additionally, one record is an example of multiple concatenated values that would

require splitting into 4 separate terms. Sorting this facet listing by count can also help to identify possible anomalies, as when only a few instances of a particular value are present where more might be expected in a table containing many thousands of records. Numeric values and dates can be aggregated and assessed in a similar way.

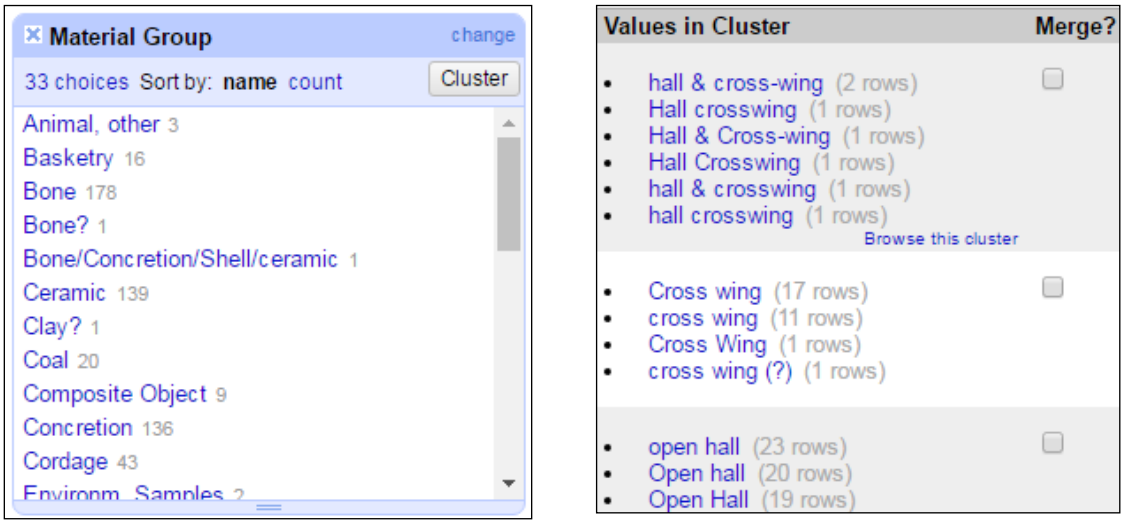

Figure 2 – Use of OpenRefine faceting, sorting and clustering of values to expose and resolve possible anomalies

To the right of Figure 2 is an example of clustering of column values by similarity, to identify different values that may be synonymous representations of the same thing. There is then the option to merge these variant values to a single new value. This form of data cleansing is a prerequisite to efficiently mapping terms to controlled vocabulary concepts.

Date spans were present in a wide variety of textual formats, all of which needed to be normalized to a common format to search them effectively. A small application was created to parse textual values from the data by matching against a series of predefined regular expression patterns covering the most common empirically observed textual expressions of date spans, to determine an appropriate start/end year for all records having some form of associated date information. This was also applied to the NLP output from the reports. By this means, it was possible to create a common numerical year index for the integrated data.

2.4 Mapping Subject terms to a common vocabulary

Using the data cleansing techniques described previously, data values were corrected as appropriate to conform to a limited coherent set of terms that were then mapped to suitable equivalent concept identifiers from the Linked Open Data implementation of the Getty Art & Architecture Thesaurus (AAT) [38] . As demonstrated in the ARIADNE portal [39] , mapping to a common 'spine' concept vocabulary facilitates multilingual cross search over subject metadata in different languages. For example, a search on the (auto-suggest) AAT concept *bowls* returns (amongst other results) Italian records with original subject metadata *bacile*. These records would not be returned if the AAT mapping had not taken place.

Mappings from Dutch controlled vocabularies to AAT concepts had previously been established during the course of ARIADNE. The DCCD team had also developed a vocabulary for the Digital Collaboratory for Cultural Dendrochronology (DCCD), which contains mappings to the AAT. The DCCD vocabulary contains a wide range of object types used in dendrochronological research [40] . Table 1 shows example mappings from DCCD concepts to AAT concepts using SKOS mapping relationships [41] . AAT mappings from a set of relevant Swedish terms were produced for purposes of the case study by SND, as were Swedish translations of a subset of AAT wood types.

Source URI	Source Label	Match type	Target URI	Target Label
dccd:a7a23364-6b80-11e5-ab22-eff9c2a3f34b	"duiker"@nl	skos:exactMatch	aat:300006116	"culvert"@en
dccd:a7a218b6-6b80-11e5-aafd-231cef94b760	"gebouw"@nl	skos:exactMatch	aat:300004790	"building"@en
dccd:a7a24188-6b80-11e5-ab32-e3504b08f149	"graanschuur"@nl	skos:exactMatch	aat:300004929	"granary"@en
dccd:a7a1f96c-6b80-11e5-aad1-af8e72a87100	"gracht"@nl	skos:exactMatch	aat:300006075	"canal"@en
dccd:a7a24804-6b80-11e5-ab3c-e789163eed6c	"heiligdom"@nl	skos:exactMatch	aat:300004575	"sanctuary"@en

Table 1 – example mappings from DANS DCCD vocabulary to Getty AAT concepts

While the various mappings proved useful in aligning many of the cleansed dataset values to AAT concepts, in some cases subjective interpretation of the intention behind the original data values was needed to determine the most appropriate thesaurus concepts.

The issue of how to represent ‘non-information’ values within the datasets proved surprisingly complex. These may be completely unstated values – e.g. NULL values or empty strings originating from an empty database field, alternatively they may take the form of *known unknowns* - string values confirming the lack of information e.g. “NOT KNOWN”, “BLANK”, “NULL”, “NOTHING”, “VOID”, “NOT SPECIFIED”, “UNSPECIFIED”, “UNCERTAIN”, “MISSING” or “EMPTY”. These are not necessarily synonymous terms, there are fine-grained semantics involved as to whether a term is describing an unstated/unknown value that is known to exist, or whether the existence of the value itself is what is uncertain - and what (if anything) can be implied where a property is not stated at all, or is stated as being an empty value. This issue is compounded in extracting data from information systems where closed world semantics are assumed (e.g. typical relational databases) into an environment supporting an open world assumption (the Semantic Web) where any stated values may be ambiguously

contradicted any number of times, and unstated values may be stated elsewhere at any time (*Anyone can say Anything about Anything*). Since these wider issues were out of scope for the case study, the solution adopted was to omit RDF triples for any unstated values, and to map any stated *known unknowns* to a limited set of AAT concepts (Table 2) judged to most closely represent the semantics of each of the values.

URI	Term(s)	Scope note
aat:300400511	N/A (information indicator), N.A., n.a., n/a, not applicable	Indication usually represented as an abbreviation, in texts, databases, tables, and lists when the topic or element is not relevant to the instance at hand.
aat:300400513	other (information indicator)	Indication in texts, databases, tables, and lists when the topic or element for the instance at hand is some value beyond the specific values provided.
aat:300400512	unavailable (information indicator)	Indication in texts, databases, tables, and lists when information for the instance at hand is not readily available.
aat:300379012	undetermined (information indicator)	Indication in texts, databases, tables, and lists when information for the instance at hand is not determined. For information that is unavailable to the cataloguer or other information provider, rather than being in general undetermined, prefer "unavailable."
aat:300386154	unidentified	General term referring to a person, people, place, or thing for which the identity has not been established.

Table 2 – AAT concepts representing 'non-information'

2.5 Natural Language Processing

Three separate Named Entity Recognition (NER) pipelines were built for processing English, Dutch, and Swedish text using the GATE platform [42] . NER is a subtask of Information Extraction aimed at the recognition and classification of units of information to predefined categories [43] (some of the archaeological entities in the case study are more specialised than the typical NER focus). The design of the pipelines followed a rule-based information extraction approach supported by a controlled vocabulary implemented as a GATE resource, originating from the Getty Art and Architecture Thesaurus. This builds on a previous study of extracting entities and relationships of interest from English language archaeological grey literature [33] . In addition to the new multilingual dimension, the case study followed a wood related focus relevant to dendrochronology analysis, including the broad classes object, sample, (wood) material, date ranges. The date extraction techniques primarily addressed numeric temporal values such as '1040 AD' with the exception of the English pipeline, which also targeted temporal appellations, such as '*sixteenth century*'. Wood material related both to tree types (e.g. *oak, beech, mahogany*) and wood products (e.g. *lumber, plywood*). The process delivered an intermediate output of XML format containing inline mark-up of the various entities and properties identified within the text, which was then transformed to the same RDF format as the data originating from databases (see section 2.6).

Overall, 25 documents relating to dendrochronology were selected for the investigation: 11 English, 9 Dutch and 5 Swedish reports, contributing a total of 501,871

Tokens (words and punctuations). The ADS Grey literature archives[8] were searched for reports relating to "*dendrochronology*", while Dutch partners provided a sample of 9 Dutch reports from the DANS EASY archive and Swedish partners provided 5 reports based on a focus on wood material and dendrochronological analysis. Different strategies were explored for identifying potentially relevant material. An extract of relevant sections from the Swedish reports was produced manually for the case study. The Dutch pipeline explored the potential for automatic detection of dendrochronology related sections. A gazetteer of approximately 40 Dutch words and phrases relevant to dendrochronology discussion was compiled. A pre-processing component identified and extracted relevant sections by matching the gazetteer input and expanding on 3 sentences before and after each match. Overlapping sections were normalised and the identified passages were extracted and compiled into a new document collection. The issue is further explored in Section 4.1.

The rules for the Dutch and English pipelines were driven by a hierarchical subset of AAT concepts, while the Swedish pipeline exploited vocabulary that had been mapped to AAT concepts. The AAT subsets were taken from the hierarchies, Architectural Elements[9] and Wood and Wood Products[10]. The hierarchies were retrieved from the Getty AAT SPARQL end-point and transformed via XSLT scripts to GATE enabled OWL-Lite structures. The corresponding preferred labels (skos:prefLabel) were employed for the English and Dutch pipelines respectively. With respect to temporal appellations, the English NER pipeline employed the Historic England Periods thesaurus[11].

The NER pipelines perform in a cascading order of 5 subsequent phases. The first phase employs a set of domain independent NLP modules such as, Tokenizer, Part of Speech Tagger, and Lemmatiser which produce an output of Tokens necessary for the operation of the subsequent domain dependent phases. The second phase is responsible for producing the Lookup matching that is driven by the controlled vocabulary whereas the third phase employs contextual (hand-crafted) rules for classifying the Lookup output to the respective entities of interest. During the fourth phase the entity classification output is validated and matches that classify as verbs or stop-words are discarded. The output of the NLP pipelines was mapped to CIDOC-CRM entities as described in the following section.

The following examples illustrate the English, Dutch and Swedish NLP output (before transformation to RDF), with colour coding indicating the semantic entities identified (Legend: objects, materials, dates, samples):

---

[8] ADS Library of Unpublished Fieldwork Reports (Grey Literature Library) http://archaeologydataservice.ac.uk/archives/view/greylit/ (accessed 11 May 2018)

[9] AAT Architectural Elements. http://vocab.getty.edu/aat/300000885 (accessed 11 May 2018)

[10] AAT Wood and Wood Products. http://vocab.getty.edu/aat/300011913 (accessed 11 May 2018)

[11] Historic England Periods. http://purl.org/heritagedata/schemes/eh_period (accessed 11 May 2018)

English
The calculation of the common felling period for each dated timber from this floor suggests a construction date between AD 1682 and c AD 1699.
Two timbers dated from the west wing roof produce felling dates in the winter of AD 1735/6 and the spring of AD 1736.
The results identified that one board was datable by tree-ring dating techniques, with this board felled in either the late-sixteenth century or early seventeenth century.
Dutch
Dendrochronologisch onderzoek door Stichting RING in Amersfoort wijst uit dat de eik waaruit de paal is vervaardigd, is geveld tussen 55 en 69 na Chr.
De dateringen op basis van dendrochronolo- gisch onderzoek van het hout uit de sporen 6 en 9 wijzen uit dat een eventuele de reparatie voor 62 na Chr.
Swedish
Två prover togs från åtelpålen och kunde genom en dendrokronologisk analys dateras till 1730-tal.
Prov 1 som var bearbetat virke av ek daterades till fällningsår vinterhalvåren 1536/37.

2.6 Data conversion

Two significant issues for data integration using the CIDOC CRM as the semantic framework have been the complexity of the process and the potential for creating multiple valid mapping expressions (chains of CRM entities and relations) for the same underlying semantics in different databases [44] . Different valid CRM expressions can result in integrated data that do not 'join up' for practical retrieval purposes unless an additional index is created or specific queries introduced for each mapping variant. For this reason, in a previous UK data integration exercise in collaboration with ADS, we followed a *mapping pattern* template-based approach. This offers an easier entry for users to map their data to the CIDOC CRM (or other) ontology when it is possible to make templates available for key use cases (such as cross search). Data manipulation skills are required but not necessarily detailed knowledge of semantics or the ontology. In the STELLAR project, ADS archaeologists were able to use the toolkit and guidelines to extract and publish archaeological linked data (see discussion in [4]). Another current example of a pattern based approach can be found in the Linked Art Project, which aims to provide a shared model for describing art with Linked Open Data. The Linked Art Data Model [45] comprises a subset of the CIDOC CRM complemented by Getty Vocabulary LOD (including AAT) concepts. The model is expressed as a series of interlinking components, where community driven best practice patterns describe how each component should be practically implemented using a primary target serialization format of JSON-LD.

An application (STELETO) was developed for the case study, derived from a core subset of the original STELLAR functionality, reduced to the minimum required for frequently encountered tabular data conversion tasks. Non-core features were omitted (e.g. XSL transformation option and GUI interface) and the command line options were simplified in order to make typical batch processing operations more straightforward. STELETO [46] is a cross-platform command line application (open source) that performs

bulk transformation of delimited text tabular data into other textual formats via a custom template.

Contents of example CSV delimited text input file (mydata.csv):

```
id,bt,en,fr
001,,animals,animaux
002,001,vertebrates,vertébrés
003,001,invertebrates,invertébrés
004,002,mammals,mammifères
005,003,insects,insects
```

Contents of example STG template file to perform the conversion operation (mytemplate.stg):

```
delimiters "{" , "}"

HEADER(options) ::= <<
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix : <{options.baseuri}> .
: a skos:ConceptScheme .
>>

RECORD(data, options) ::= <<
:{data.id} a skos:Concept ; skos:inScheme : ;
      skos:prefLabel "{data.en}"@en, "{data.fr}"@fr .
{if(data.bt)}
:{data.id} skos:broader :{data.bt} .
:{data.bt} skos:narrower :{data.id} .
{else}
:{data.id} skos:topConceptOf : .
: skos:hasTopConcept :{data.id} .
{endif}
>>
```

STELETO command line:

```
C:\path\to\STELETO.exe -f -d:","
  -i:"c:\path\mydata.csv" -t:"c:\path\mytemplate.stg"
  -o:"c:\path\myoutput.ttl" -p:baseuri:"http://temp/"
```

Contents of resultant output file – CSV input converted to valid SKOS TURTLE RDF (myoutput.ttl):

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix : <http://myscheme> .
: a skos:ConceptScheme .

:001 a skos:Concept ; skos:inScheme : ;
      skos:prefLabel "animals"@en, "animaux"@fr .
:001 skos:topConceptOf : .
: skos:hasTopConcept :001 .
```

```
:002 a skos:Concept ; skos:inScheme : ;
      skos:prefLabel "vertebrates"@en, "vertébrés"@fr .
:002 skos:broader :001 .
:001 skos:narrower :002 .

:003 a skos:Concept ; skos:inScheme : ;
      skos:prefLabel "invertebrates"@en, "invertébrés"@fr .
:003 skos:broader :001 .
:001 skos:narrower :003 .

:004 a skos:Concept ; skos:inScheme : ;
      skos:prefLabel "mammals"@en, "mammifères"@fr .
:004 skos:broader :002 .
:002 skos:narrower :004 .

:005 a skos:Concept ; skos:inScheme : ;
      skos:prefLabel "insects"@en, "insectes"@fr .
:005 skos:broader :003 .
:003 skos:narrower :005 .
```

Figure 3 – STELETO data conversion example

Figure 3 illustrates a simple illustrative example CSV to SKOS data conversion using STELETO, showing the input data file, the template, the command line options used and the resultant output. Internally the STELETO application utilizes the *StringTemplate* engine [47] to transform the delimited text input data according to the specified template. STELETO looks for the presence of 3 optional named templates: HEADER (called once at the start of processing), RECORD (called once per data record) and FOOTER (called once at the end of processing). These user-defined templates represent patterns of text to be written to the output, with embedded named placeholders that are replaced with the corresponding named data field values at runtime. To enforce strict model-view separation, templates support only necessary functionality (simple conditional statements based on the presence/absence of data values). The output may be any textual format as prescribed by the template used. In Figure 3 an RDF semantic graph structure in Turtle format is produced, consisting of 31 triples that describe 5 multilingual SKOS Concepts belonging to a single Concept Scheme and connected via bidirectional hierarchical relationships. This output can be imported directly into RDF aware applications, combined with other RDF data, queried using SPARQL, and visualized (Figure 4).

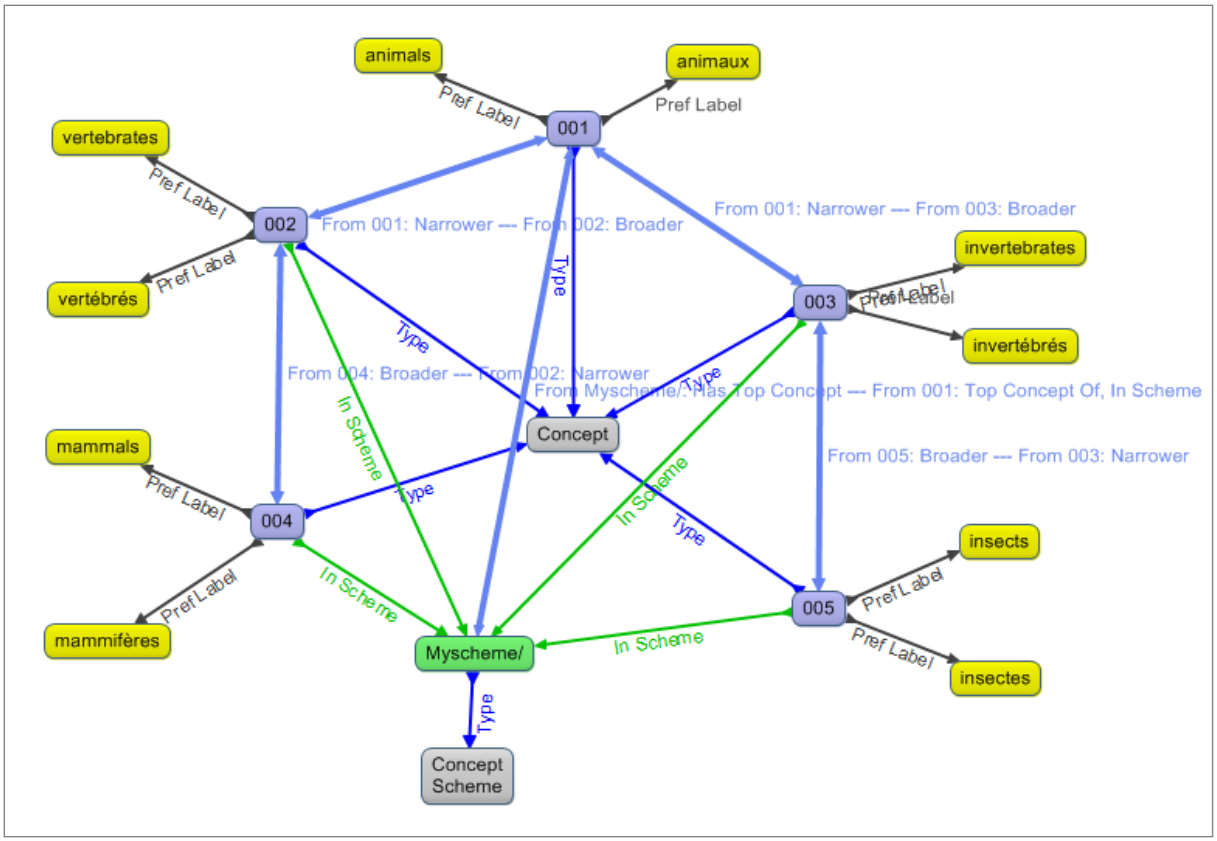

Figure 4 – graphical representation of the resultant RDF semantic graph structure

A custom template was produced specifically for this case study generating NTriples format serialization of RDF data representing CIDOC CRM entities and properties. Use of the same template for all data conversion (both for datasets and NLP output) in the case study ensured validity and consistency of all RDF output.

The GATE NLP output consisted of a series of XML files (one file per original source document). Each XML file contained text extracted from the corresponding original source document, with inline embedded XML elements representing a number of custom entities identified by the GATE processing (*Sample, SamplePhrase, Date, woodMaterials, archElements Has_Time-Span* etc.). Element containment represented a link between elements (e.g. an *archElements* entity containing a *woodMaterials* element indicated an object *made of* a material. This information was extracted from all the GATE output XML files using a batch XSL transformation process, creating a set of consistent delimited text data files for subsequent input to the STELETO application.

2.7 Data integration

The semantic framework used for the case study (Figure 5) was a subset of the CIDOC Conceptual Reference Model (CRM) [6] [48] . The Tree Ring Data Standard (TriDaS) [49] for dendrochronological data could have been an alternative choice for an overall data model. The CRM was used since the case study was situated within the broader ARIADNE framework, with a view to informing discussion on wider semantic

integration for archaeology research goals. All crm:E55_Type (conceptual entity) information was mapped to concepts originating from the Getty AAT.

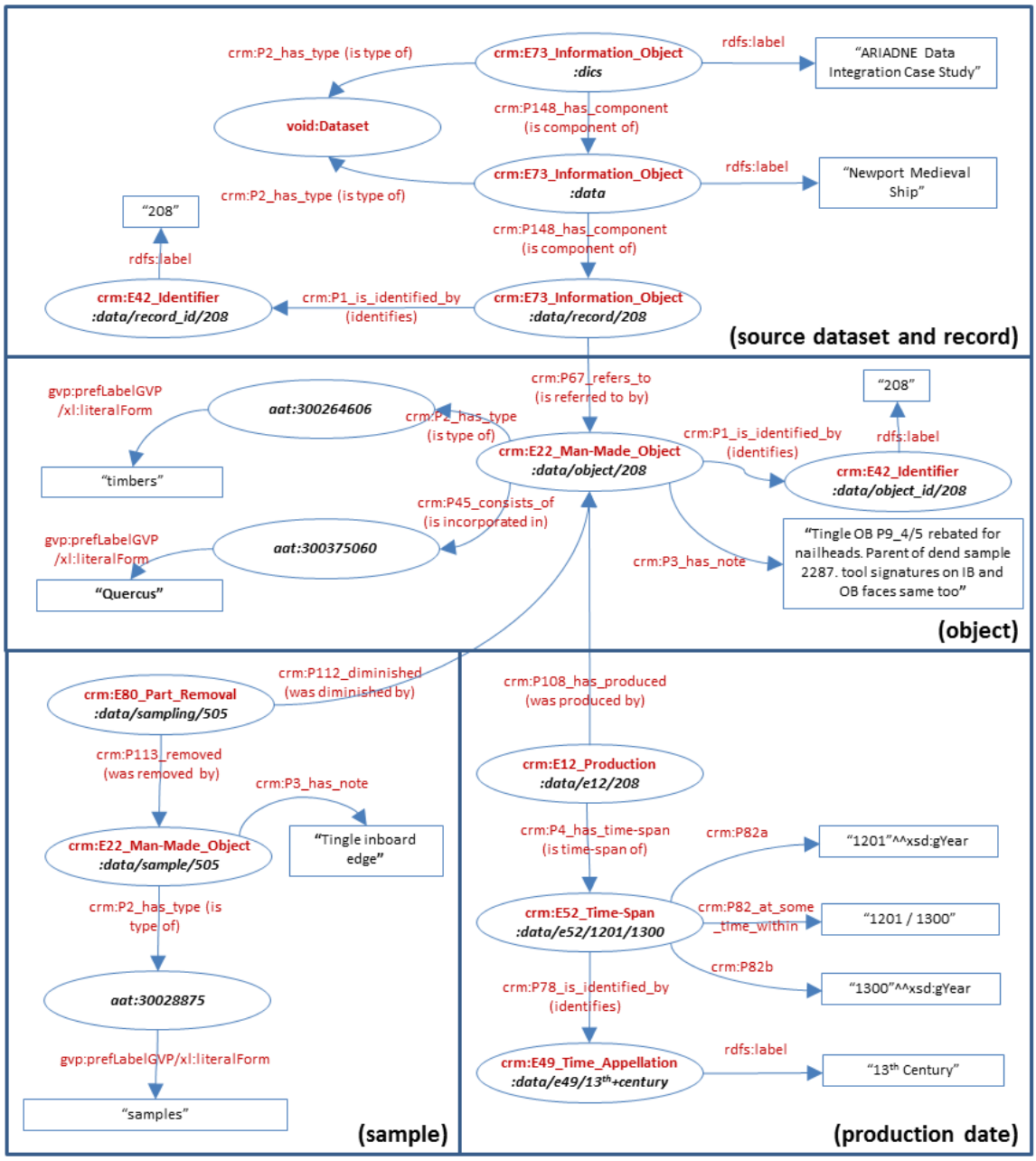

Figure 5 – semantic framework model used for the data integration case study

The template created for this work produced bidirectional relationships between entities by default; being more explicit in this way reduces the requirement for end users of the data to undertake semantic reasoning and assists more flexible query formulation.

2.7.1 Integration results

The resultant RDF data produced was consolidated as a single named graph into a Virtuoso triple store [50] to support cross search. A total of 1.09 million RDF triples were

produced, representing 23,594 multilingual records and referencing 37,935 objects. Virtuoso full-text indexing was configured for the consolidated data, allowing a more flexible combination of syntactic and semantic querying.

3 Demonstrator web application

Queries can be formulated directly at the SPARQL endpoint. However, this can prove difficult without a detailed knowledge of the underlying data schema and particular query syntax supported. Therefore, a query builder application [51] was developed for the case study, as a demonstration of techniques to achieve easier searching and browsing of the integrated RDF dataset. The application performs hierarchical thesaurus concept expansion and allows a combination of both free text search and structured semantic search. The demonstrator is a bespoke application and user interface for the case study, building on and taking forward the general approach followed in STAR: single page integrated application, query builder performing interactive background generation and execution of SPARQL queries, AJAX remote server interaction, JSON responses and a JavaScript "widget" component based approach (using the JQuery UI Widget Factory [52]).

The query builder supports point and click interactive formulation of structured queries, dynamically building a correctly formatted SPARQL 1.1 query in the background to be executed against the consolidated RDF data accessed via the SPARQL endpoint. Queries conform to the model described in Figure 5, targeting *records* referring to *objects* or to *samples*, which then have certain properties that can be specified. Some query builder controls allow selecting a single property value from a limited list of possible values generated from the data (e.g. record sources, object types / materials), some controls allow free text searching within textual notes, and a specialised date selection control allows limiting the query scope to a particular date range (start year → end year) using dual sliders. Expanding and specifying any property value automatically adds it to the query; collapsing any property removes it from the query. These features facilitate quick experimentation and incremental interactive query building. Figure 6 illustrates an example usage of the query builder (on the left hand side) to construct and execute a query, rendering the results on the right hand side. It shows a query with object type and date range based on the ability to query over the CRM structure via an object production event.

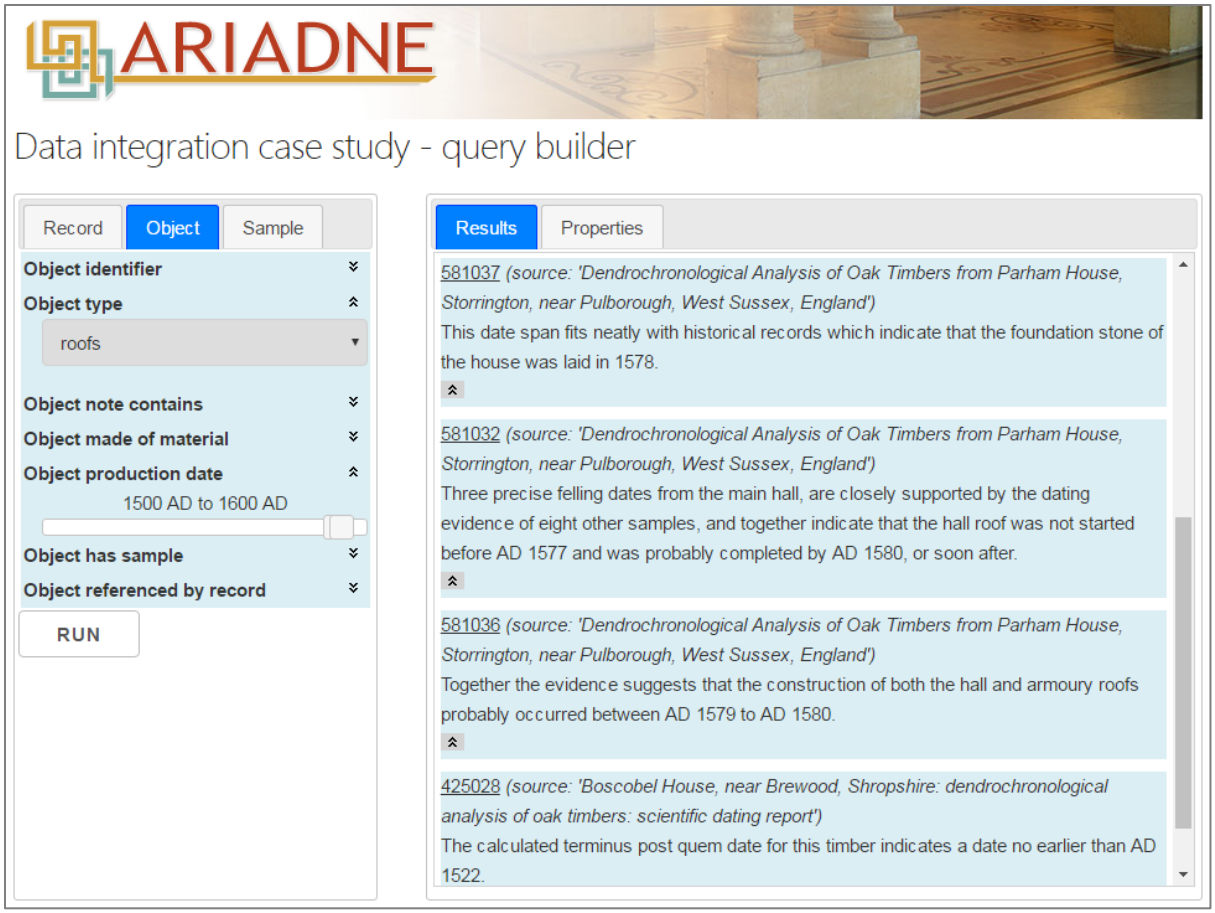

Figure 6 - Demonstrator query builder

The application facilitates the formulation of structured queries without necessarily requiring knowledge of the details of the underlying data structure or of SPARQL 1.1 syntax. The query is on *roofs* having a production date in the range *1500-1600 AD*. The results displayed in this particular example originate from the outcome of NLP processing of textual reports where the process associated an identified instance of an object type with a date range.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX aat: <http://vocab.getty.edu/aat/>
SELECT DISTINCT ?object ?label ?note ?source
FROM <http://registry.ariadne-infrastructure.eu/usw-data-integration-
case-study>
FROM <http://vocab.getty.edu/dataset/aat>
WHERE {
?object rdf:type <http://www.cidoc-crm.org/cidoc-crm/E22_Man-
Made_Object> .
```

```
?objectproduction <http://www.cidoc-crm.org/cidoc-
crm/P108_has_produced> ?object; <http://www.cidoc-crm.org/cidoc-
crm/P4_has_time-span> [<http://www.cidoc-crm.org/cidoc-
crm/P82a_begin_of_the_begin> ?yearMin ; <http://www.cidoc-
crm.org/cidoc-crm/P82b_end_of_the_end> ?yearMax ] .
FILTER (year(coalesce(xsd:DateTime(?yearMin), xsd:DateTime('5000'))) >=
1500 && year(coalesce(xsd:DateTime(?yearMax), xsd:DateTime('5000'))) <=
1600) .
?object crm:P2_has_type/gvp:broaderGeneric?
<http://vocab.getty.edu/aat/300002098> .
OPTIONAL { ?object <http://www.w3.org/2000/01/rdf-schema#label> ?label
}
OPTIONAL { ?object <http://www.cidoc-crm.org/cidoc-crm/P3_has_note>
?note }
OPTIONAL { ?object <http://www.cidoc-crm.org/cidoc-
crm/P67i_is_referred_to_by> [ a <http://rdfs.org/ns/void#Dataset>;
<http://www.w3.org/2000/01/rdf-schema#label> ?source ] }
}
```

Figure 7 – SPARQL 1.1 query as constructed by the query builder application

The resultant underlying SPARQL 1.1 query as built and executed by the query builder is shown in Figure 7.

3.1 Thesaurus query expansion

Mapping the data to Getty AAT concepts provided common points of reference between discrete datasets, facilitating cross querying of multilingual data. Another use of thesauri in search systems is to employ the semantic structure for query expansion (QE). Shiri et al. [53] review the use of thesauri in search system user interfaces. An 'explode' command is sometimes used in commercial search systems to give a form of narrower expansion by simply adding narrower terms to the original (string match) query. However, this can result in mismatches when terms are homographs. [54] reviews thesaurus-based query expansion (QE) and reports on a (pre linked data) study of concept-based QE over the AAT's semantic relationships and facet structure on the Science Museum's collections database where the thesaurus was integral to the user interface. Here the QE algorithm automatically expanded over all thesaurus relationships subject to a threshold of semantic distance.

It should be remembered that thesaurus QE is not necessarily equivalent to logical inference but rather an expansion of the scope of a query based on the thesaurus semantic structure with probable relevance of any additional results for the user to choose from. Depending on the thesaurus, the broader relationship can subsume more specialised sub-types of hierarchical relationship. The vast majority of the AAT's hierarchical relationships are 'broaderGeneric' (species/genus relationship) but the AAT also contains a few 'broaderPartitive' (part/whole relationships) and the composition of the two subtypes in QE can sometimes bring in unexpected results depending on the query. In fact, the Demonstrator only uses the specialised broaderGeneric relationship, which will yield reliable results in QE (see [55] which discusses the composition of thesaurus hierarchical relationships in some detail). Another approach, as followed in [54] , is to associate a cost (varying by relationship) with each traversal of the thesaurus structure and thus limit the extent of any query expansion or prioritise particular relationships.

Gavel and Andersson [56] discuss results from QE over the Medical Subject Headings (MeSH) thesaurus on a Swedish bibliographic database, taking advantage of the multilingual entry vocabulary when mapping query terms to thesaurus concepts. In this case, the QE algorithm directly made use of the tree-based representation of the thesaurus concept in indexing so that it was possible to achieve narrower expansion by truncating the MeSH tree number. Taking advantage of the (tree-based) format of the index term identifiers gives an efficient implementation of narrower expansion. However it limits the QE algorithm to a particular thesaurus (or identifiers that follow a particular tree structure) and it does not allow expansion over the other thesaurus semantic relationships.

The case study made use of *property paths* (a new feature introduced in SPARQL 1.1) to perform semantic query expansion over the hierarchical and associative links between vocabulary concepts. In Figure 8, a query on the underlying concept for the term "willow" is expanded automatically to include all *narrower* concepts in the hierarchical structure via the specialised broaderGeneric relationship. Thus a query at a general level can also retrieve resources indexed more specifically. In addition, it is possible to expand over other thesaurus relationships, as discussed below.

aat:300264091	Materials Facet
aat:300010357	. Materials (hierarchy name)
aat:300010358	. . materials (matter)
aat:300206573	. . . <materials by origin>
aat:300265629	. . . . biological material
aat:300124117	. . . . . plant material
aat:300011913	. . . . . <wood and wood products>
aat:300011914	. . . . . wood (plant material)
aat:300011915	. . . . . <wood by composition or origin>
aat:300011916	. . . . . hardwood
aat:300012498	. . . . . **willow (wood)**
aat:300012500	. . . . . black willow (wood)
aat:300012502	. . . . . Japanese willow (wood)
aat:300012504	. . . . . western black willow (wood)
aat:300012508	. . . . . white willow (wood)

Figure 8 - hierarchical structure for AAT concept 300012498 "willow (wood)"

Differences in indexing and descriptions were observed, in that references to wooden materials within the datasets and grey literature might use reference names of materials or family/genus/species terms interchangeably. Although not intended as a formal scientific taxonomy, the Getty AAT does include a hierarchical structure of family/genus/species concepts - an example is shown in Figure 9.

aat:300264089	Agents Facet
aat:300265673	. Living Organisms (hierarchy)
aat:300390503	. . living Organisms (entities)
aat:300265677	. . . Eukaryota (domain)
aat:300132360	. . . . Plantae (kingdom)
aat:300265706	. . . . . Angiospermae (division)
aat:300375593	. . . . . Magnoliopsida (class)
aat:300374936	. . . . . Malpighiales (order)
aat:300374937	. . . . . Salicaceae (family)
aat:300375384	. . . . . **Salix (genus)**

aat:300375393	. . . . . Salix alba (species)
aat:300375392	. . . . . Salix bakko (species)
aat:300375391	. . . . . Salix cardiophylla (species)
aat:300375390	. . . . . Salix gilgiana (species)
aat:300375387	. . . . . Salix lucida (species)
aat:300375389	. . . . . . Salix lucida ssp caudate
aat:300375385	. . . . . Salix nigra (species)

Figure 9 - Taxonomic structure for AAT concept 300375384 "Salix (genus)"

The AAT includes specific specialized associative relationships (Figure 10), connecting the concepts found within the *Materials* hierarchy and the *Agents (Living Organisms)* hierarchy.

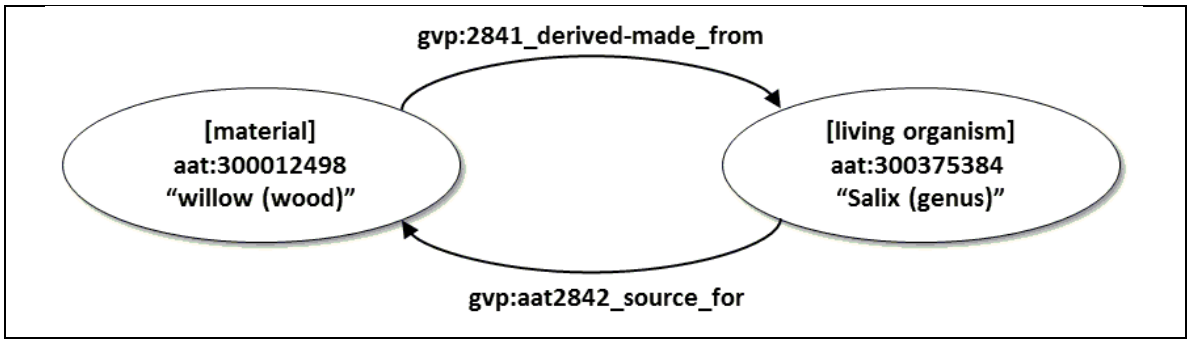

Figure 10 - AAT specific RT specialization

For more effective search we employed these specialized associative relationships in query expansion. For example, based on the data and relationships shown in Figure 8, Figure 9 and Figure 10, a query for resources linked to *willow* also retrieves resources linked to *Salix (*and any/all of their respective hierarchical descendant concepts), as seen in Table 3.

```
PREFIX aat: <http://vocab.getty.edu/aat/>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX xl: <http://www.w3.org/2008/05/skos-xl#>

select ?uri str(?lbl AS ?label) WHERE {
    ?uri
gvp:broaderGeneric?/(gvp:aat2842_source_for|gvp:aat2841_derived-
made_from)? aat:300375384 .
    OPTIONAL { ?uri gvp:prefLabelGVP [xl:literalForm ?lbl] }
}
```

uri	label
aat:300375384	Salix (genus)
aat:300375385	Salix nigra (species)
aat:300375387	Salix lucida (species)
aat:300375390	Salix gilgiana (species)
aat:300375391	Salix cardiophylla (species)
aat:300375392	Salix bakko (species)
aat:300375393	Salix alba (species)
aat:300012498	willow (wood)
aat:300012500	black willow (wood)
aat:300012502	Japanese willow (wood)
aat:300012504	western black willow (wood)

aat:300012508	white willow (wood)

Table 3 – SPARQL query on AAT concepts exploiting hierarchical and associative relationships, with results

This principle can be observed clearly in the demonstrator application (Figure 11), where querying for records referring to e.g. *"pine (wood)"* retrieves Swedish records referring to aat:300012620 *"pine (wood)"*, English records referring to aat:300343658 "*Pinus (genus)"* and Dutch records referring to aat:300343781 "*Pinus sylvestris (species)"* - a hierarchical descendant of aat:300343658 "*Pinus (genus)"*. Without the expansion only the Swedish records would be retrieved.

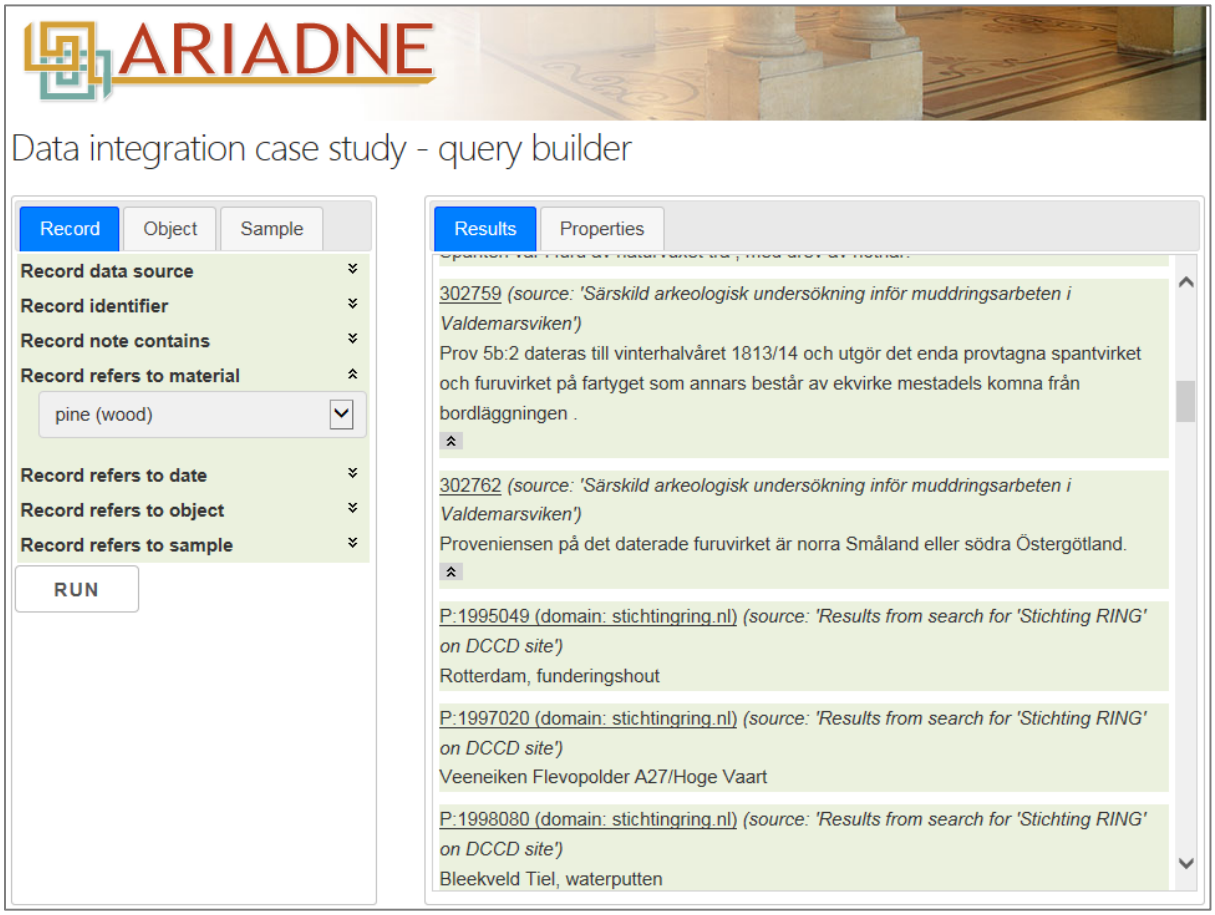

Figure 11: Query on records referring to "pine" – results include records retrieved via query expansion

3.2 Demonstrator scenarios

The following scenarios illustrate other aspects of the Demonstrator. The Newport medieval ship proved to contain rich data on the different kinds of wood used to construct the ship. Figure 12 shows a query of nautical rigging elements with query expansion on object types and materials. Different kinds of rigging device (*treenail, dead eye, sheave, parrel,* etc.) are retrieved, made of a variety of wood (*oak, ash and some elm, alder, boxwood*).

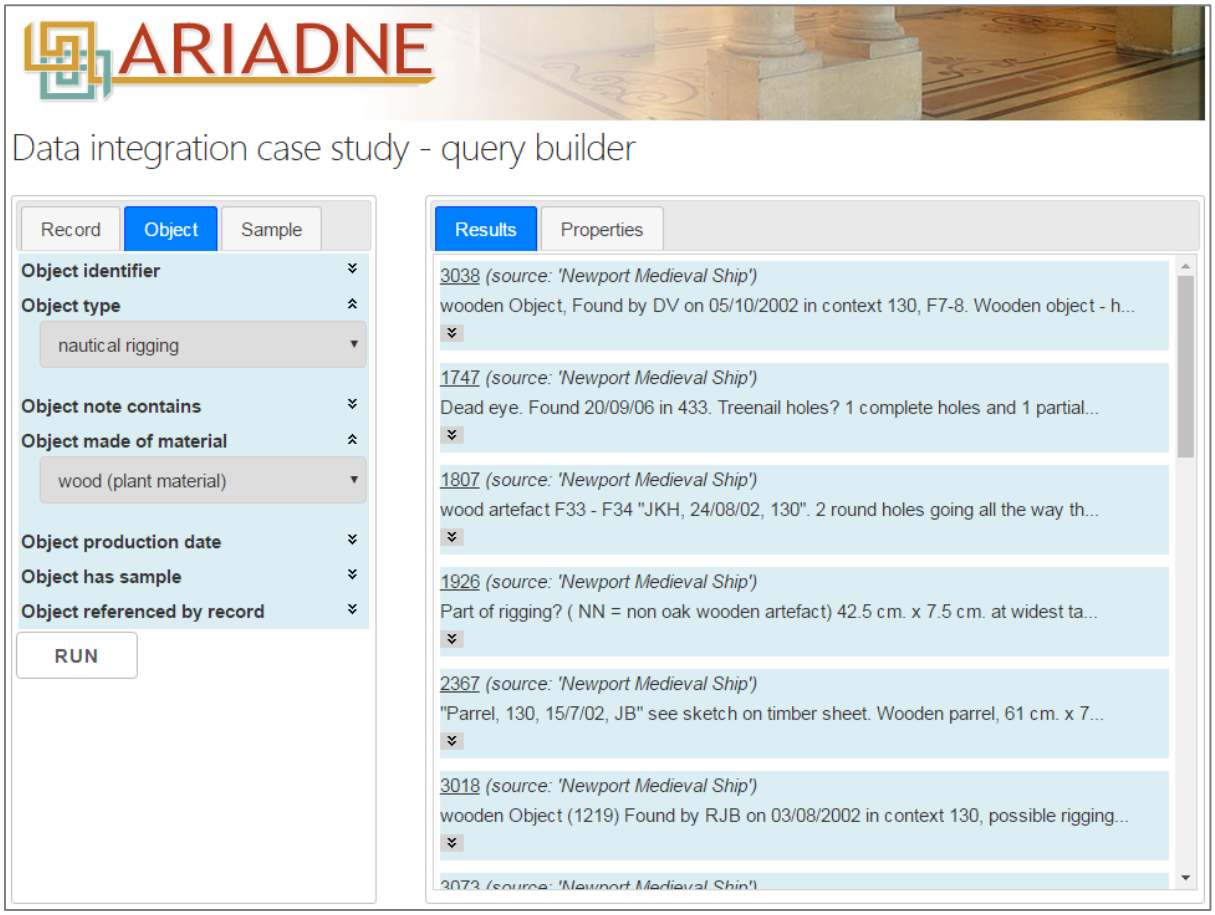

Figure 12: Demonstrator query on nautical rigging with different types of wood from the Newport medieval ship

Different datasets and reports hold information on keels and their construction. Figure 13 shows results from the DCCD data base and the Newport medieval ship for a query on *oak keels*.

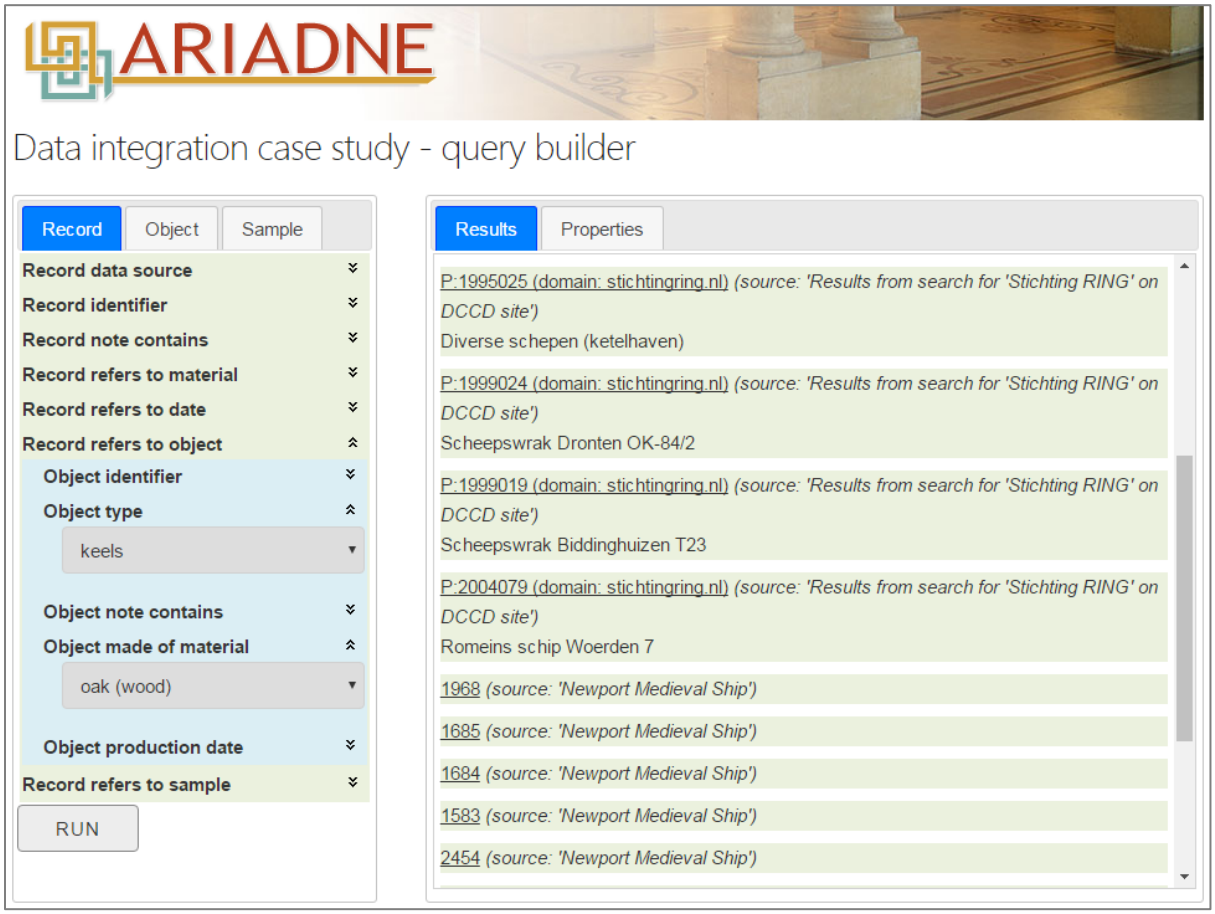

Figure 13: Demonstrator query on records referring to (objects) keels made of oak

Sometimes it is important to identify results where sampling has taken place, perhaps as an indication of reliability of the dating. Figure 14 is an example of a very specific query showing a Swedish report with a record of *pine* from a specific date, which has been sampled (and shows an expansion via the associative relationship to *Pinus (genus)*).

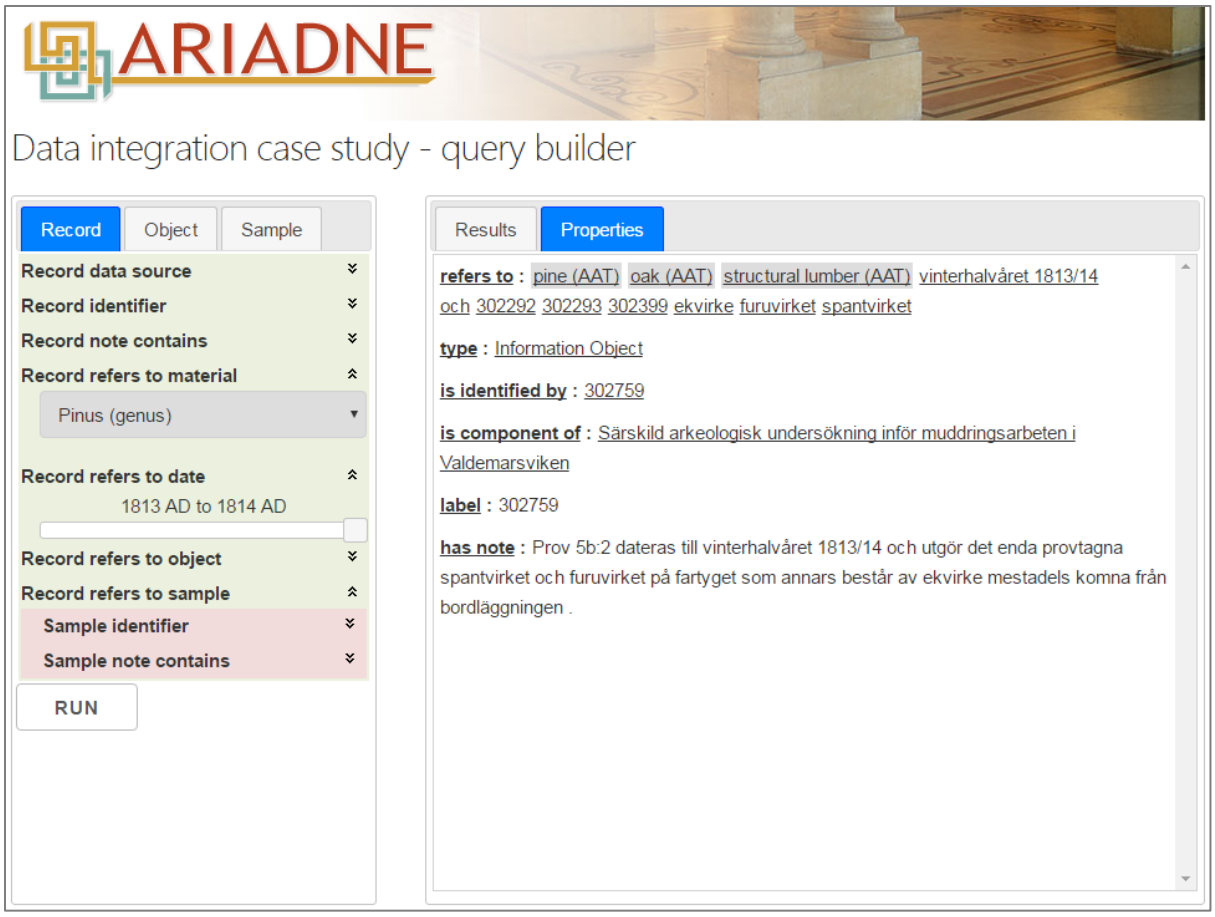

Figure 14: Demonstrator query on records referring to (material) *Pinus (genus)* from a specific date which has been sampled

4 Discussion and limitations

The CIDOC CRM relies on existing syntactic interoperability as a prerequisite [57] . In our experience syntactic interoperability is fairly rare in practice; substantial data cleansing and validation is often required. Many legacy datasets were not intended for cross search or integration purposes and rules on data entry may not have been strictly enforced or even specified. As described in Section 2.3, some of the problems stem from good intentions in data entry in attempting to provide richer data or more context than the data model affords. If these issues are not addressed then semantic integration can fail due to low level issues. Within the study, data cleansing required a significant amount of time and this should be budgeted for in semantic integration projects. While OpenRefine proved a useful tool, there is scope for further work that would identify where there is a need for data cleansing and provide a toolkit and outline common steps for simplifying the work.

The case study shows the potential of larger scale work to address broad archaeological research questions made possible by the integration of data and information resources. The constraints of the study's timescale and resources imposed some pragmatic limitations. The datasets and reports used in the study were selected from available open research data as loosely relating to dendrochronology rather than being considered as supporting a particular archaeological research question. While this was appropriate for the overall aim of the study regarding the technical feasibility of the technologies to achieve meaningful semantic interoperability, it places limits on how far the data derived for the demonstrator can address any archaeological research question. In an operational research toolkit, an initial selection phase would locate and request the key thematically related datasets and reports for a particular research question. This could involve addressing any issues of access and permission. Some thought should be given as to the intended use of the integrated data; in our view successful semantic integration requires significant resource and should be justified by an associated investigation on a domain research question. An investigation might also gather data for overviews or visualisations over time, such as the changing uses of wood and other material, or the evolution of trading patterns.

STELETO was used both for data conversion of relevant extracts from the 5 archaeological datasets and also the data resulting from the NLP information extraction from the archaeological reports (currently object production events are derived from the datasets and English language NLP data). The pattern-based templates ensure consistency both of the ontology mappings as discussed in Section 2.5 and also the lower level implementation details - differing RDF linked data implementation expressions can also thwart interoperability.

In contrast to previous work in the STAR project, where a more detailed archaeological extension of the CIDOC CRM ontology was employed (for a discussion of granularity, see [5]), for the purposes of the case study the semantic framework comprised high level entities of the CRM, further described by types from the Getty AAT Thesaurus. This is somewhat similar to the use of a few broad concepts described by Ashley et al. [15]) for their work with Çatalhöyük excavation data (their approach involved creating superclasses). The AAT narrower concept expansion functionality in the SPARQL 1.1 demonstrator application (section 3.1) enabled the capability to query at a high level of generality and still retrieve specific results, or to directly query at a lower level of detail. This was further elaborated by the query expansion between AAT facets via the specialised associative relationships, allowing a connection to be made between lay and scientific terminology for wood types. A review and discussion on the potential for specialising the thesaurus associative relationship can be found in [58] .

While the Query Builder web application is a prototype, it illustrates that more domain application oriented user interfaces are possible for searching RDF datasets than the common SPARQL endpoint or high level browsing interfaces. In an operational user interface, more elaborate auto-suggest elements would be employed in the query pane and more context and navigation options provided on the results panes. The user interface is not automatically derived from the underlying ontology; any major change to the data model might necessitate alterations to the interface (albeit changes to the high level

entities are unlikely). It demonstrates that the user interface can hide much of the complexity of the underlying data model and the associated query syntax, facilitating more straightforward searching and browsing of the dataset without requiring specialist knowledge. Web technology progresses quickly and the growth of frameworks such as Angular/React/VUE (etc.) indicate promising future directions for reusable interactive components and platform/device neutral applications for accessing, caching, integrating and visualizing semantic knowledge originating from SPARQL endpoints and data APIs.

4.1 Natural Language Processing

For purposes of the case study, a lenient information extraction strategy was followed in order not to miss potential examples and false positives can be found in the NLP output in all languages. Further work is needed for operational versions. Nonetheless, the principle of semantic data integration from text documents and databases has been demonstrated. The case study was able to generate CRM/AAT based output via NLP techniques from English, Dutch and Swedish texts in the same format as the instance data extracted and mapped to the CRM/AAT.

Even in operational systems, RDF statements resulting from inherently ambiguous natural language do not carry the same degree of reliability as those derived from the datasets. An indication of the provenance of the RDF data and the workflow involved should be included in the semantic framework, which would allow judgments of the reliability of the information. More work is also required on the appropriate semantic model for expressing data extracted via NLP since natural language is less precise or more general than in databases. For example, in some cases a report may refer to a specific object (a particular artefact find from an excavation which has been preserved), whereas in other cases a report may refer to artefacts encountered (but not individual instances), or a report may make a general statement about particular types of artefacts. Depending on the intended use case(s) of the information extraction exercise, it may or may not be important to model these distinctions.

More work is needed on Relation Extraction algorithms that could assert CRM properties between entities. The English language NLP output is based on grammatical patterns for Relation Extraction, building on previous work [33] . These extract contextual relationships between objects and dates or material. For the Dutch and Swedish reports, simpler techniques are used that do not attempt connections between entities extracted (other than co-occurrence within the same sentence). Future work would apply a more contextualised information extraction approach to Dutch and Swedish reports similar to the English language work. The development of techniques for the annotation of compound noun forms is also important for Dutch and Swedish pipelines, along with refinements to their stemming and part of speech components.

An operational system would require enlarged vocabularies drawing on relevant resources for the research questions, if necessary adapting the terminology for NLP purposes. In the study, some English and Dutch terms were classed as 'stop words' and excluded from matching due to the high potential for producing false positives within the context of the case study. Polysemous Swedish terms, such as *lager*, would be good

candidates for stop words. An extended glossary of contextual date indicators is also important given archaeology's focus on dating.

Ambiguity between *material* and *object* senses proved challenging in some cases (for both machine and human annotators). For example, in the Swedish reports, it was difficult to distinguish between say a pine tree *(tall)* and material made from *pine*. In fact, archaeological reports do not always make clear distinctions and the issue of whether the semantic distinction is important for the use case (research question) should be considered.

As discussed in section 2.5, the case study explored different methods for identifying passages of particular relevance for information extraction. This is an important issue, given the length of many archaeological reports. Sections which follow their own structure, such as tables or references, should either be omitted or merit a specialised NLP component. References can contain instances of names as homographs, which can result in false positives. Others sections, such as a historical review, may make side references to other excavations or previous work and in such cases the entities extracted may not represent the core subject matter or results of the report. The ability to detect different types of document section automatically would be valuable, although this is made difficult by the variety of report formats and writing styles encountered. Practical approaches can attempt to focus information extraction on report abstracts or conclusions, prioritise the start of a document or attempt to make use of the frequency of particular annotations in a document.

5 Conclusions

There are a number of contributory factors to achieving successful data integration and full interoperability. Data cleansing was a vital step before conversion. Use of a common data schema/ontology allows the data structure to be cross searched, in this case orienting to high level entities from the CIDOC CRM ontology in combination with concepts from the Getty AAT. Following a 'mapping pattern' approach with the same template for all data conversion (using the STELETO tool) ensured validity and consistency of all RDF output. By referencing/mapping terms to a common controlled vocabulary (AAT concepts), commonality could be distinguished within the data - even where the records originated from different sources using different data schema and were even expressed in different languages. In addition, the AAT thesaurus structure was utilized to automatically expand queries both hierarchically and via specialized associative relationships.

The Demonstrator implementation is a novel SPARQL web application, with CRM/AAT based data integration. Functionality incudes hierarchical and associative thesaurus expansion and combination of free text and semantic search with browsing on semantic links. The Demonstrator hides the complexity of the underlying semantic framework from the user. This simplification of course does not permit the construction of arbitrary queries and reduces the potential to explore or quantify the underlying graph of entities and relationships. However the Demonstrator does not preclude such investigation as it is directly querying a SPARQL endpoint which is also accessible by

the end user – though effective direct queries would require knowledge of the underlying ontological model and SPARQL syntax. The option of a graphical user interface (or possibly an API) can shield end users from the need to fully understand these details. Query builder user interfaces can generate optimized queries, and can assist query formulation by providing controlled lists of possible values to choose from. They can simultaneously execute multiple asynchronous query requests against multiple remote data endpoints and APIs, consolidating, filtering, sorting and presenting the results.

Grey literature reports are an underutilised resource which can be combined with datasets for meta research and large scale studies. NLP methods have the potential to extract specific items of information not found in the report metadata, which can be useful for many research questions. The case study demonstrates the feasibility of connecting information (at a detailed level) extracted from datasets and also grey literature reports to the same RDF data format allowing semantic cross-searching of the integrated information. The semantic integration of the contents of textual reports and datasets opens new possibilities for research across diverse resources not previously combined.

In future work, we aim to evaluate these methods addressing real research challenges that require the semantic integration of different datasets and textual information. This will require the active participation of domain experts as collaborators in providing use cases and research questions for the novel combination of resources, the terminology and vocabulary used in relevant subject domains and as users in the refinement and evaluation of the resulting research toolkit application.

6 Acknowledgements

This work was supported by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193 (the ARIADNE project). Thanks are due to ARIADNE project partners generally and those who facilitated access to archaeological datasets and reports, including Julian Richards (Archaeology Data Service), Paul Boon & Hella Hollander (KNAW-DANS), Esther Jansma (Universiteit Utrecht), Milco Wansleeben (Universiteit Leiden) and Jeremy Azzopardi (Swedish National Data Centre) who contributed vocabulary, AAT translations and mappings and commented on Swedish NLP issues.

7 References

[1] ISO 25964-1:2011. Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. https://www.niso.org/schemas/iso25964#part1 (2011, accessed 11 May 2018).

[2] Falkingham G. A Whiter Shade of Grey: a new approach to archaeological grey literature using the XML version of the TEI Guidelines. Internet Archaeology 2005; 17. http://dx.doi.org/10.11141/ia.17.5 (accessed 11 May 2018).

[3] Richards J and Hardman C. Stepping Back from the Trench Edge: an archaeological perspective on the development of standards for recording and publication. In: (Eds

Greengrass and Hughes). The Virtual Representation of the Past. Ashgate, 2008, pp. 101-112.

[4] Binding C, Charno M, Jeffrey S, May K and Tudhope D. Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data. International Journal on Semantic Web and Information Systems 2015; 11(1), 1-29.

[5] Tudhope D, May K, Binding C, Vlachidis A. Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology 2011; 30, https://doi.org/10.11141/ia.30.5 (accessed 11 May 2018).

[6] ISO 21127:2014. Information and documentation -- A reference ontology for the interchange of cultural heritage information. https://www.iso.org/standard/57832.html (accessed 15 May 2018)

[7] ISO 25964-2:2013. Information and documentation - Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies. https://www.niso.org/schemas/iso25964#part2 (accessed 11 May 2018).

[8] Isaac A, Waites W, Young J and Zeng M. (Eds.). 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/ (accessed 11 May 2018).

[9] Aloia N, Binding C, Cuy S, Doerr M, Fanini B, Felicetti A, Fihn J, Gavrilis D, Geser G, Hollander H, Meghini C, Niccolucci F, Nurra F, Papatheodorou C, Richards J, Ronzino P, Scopigno R, Theodoridou M, Tudhope D, Vlachidis A and Wright H. Enabling European Archaeological Research: The ARIADNE E-Infrastructure. Internet Archaeology 2017; 43. https://doi.org/10.11141/ia.43.11 (accessed 11 May 2018).

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

[11] May K, Binding C and Tudhope D. Barriers and opportunities for Linked Open Data use in archaeology and cultural heritage. Archäologische Informationen 2015; 38, 173-184. DGUF. http://dx.doi.org/10.11588/ai.2015.1.26162 (accessed 11 May 2018).

[12] Kintigh K. Extracting Information from Archaeological Texts. Open Archaeology 2015; 1: 96–101.

[13] Olsson M. Making sense of the past: The embodied information practices of field archaeologists. Journal of Information Science 2016; 42(3): 410–419.

[14] Hodder I. The Archaeological Process. Oxford: Blackwell, 1999.

[15] Ashley M, Tringham R and Perlingieri C. Last House on the Hill: Digitally remediating data and media for preservation and access. Journal of Computing and Cultural Heritage 2011; 4 (4), Article 13. http://dx.doi.org/10.1145/2050096.2050098 (accessed 11 May 2018).

[16] Cripps P, Greenhalgh A, Fellows D, May K and Robinson D. Ontological modelling of the work of the Centre for Archaeology. CIDOC CRM Technical Paper. 2004; http://old.cidoc-crm.org/docs/Ontological_Modelling_Project_Report_Sep2004.pdf (accessed 11 May 2018).

[17] EAMENA. Endangered Archaeology in the Middle East and North Africa project. http://eamena.arch.ox.ac.uk/ (accessed 11 May 2018).

[18] Myers D, Dalgity A and Avramides I. The Arches heritage inventory and management system: a platform for the heritage field. Journal of Cultural Heritage Management and Sustainable Development 2016; 6 (2), 213-224. Emerald.

[19] Kansa E and Kansa S. We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication. Journal of Eastern Mediterranean Archaeology and Heritage Studies 2013; 1 (1), 88-97.

[20] Faniel I, Kansa E, Kansa S, Barrera-Gomez J and Yakel E. The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. Proc. 13th ACM/IEEE-CS Joint Conference on Digital Libraries, 2013, pp 295-304. New York: ACM.

[21] Barker E, Simon R, Isaksen L and de Soto Cañamares P. The Pleiades Gazetteer and the Pelagios Project. In: (Berman, Merrick Lex; Mostern, Ruth and Southall, Humphrey eds.) Placing Names: Enriching and Integrating Gazetteers. Bloomington: Indiana University Press, 2016; pp. 97–109.

[22] Shaw R, Rabinowitz A, Golden P and Kansa E. A sharing-oriented design strategy for Networked Knowledge Organization Systems. International Journal on Digital Libraries 2016; 17(1), 49-61.

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

[24] de Boer V, Leinenga J, van Rossum M and Hoekstra R. Dutch Ships and Sailors Linked Data Cloud. Proc. International Semantic Web Conference 2014; Riva del Garda, Italy, pp 229-244.

[25] de Boer V, Wielemaker J, van Gent J Hildebrand M, Isaac A, van Ossenbruggen J and Schreiber G. Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. Proc. 9th Extended Semantic Web Conference 2012; Lecture Notes in Computer Science 7295. Springer, pp 733-747.

[26] Europeana Data Model Primer. (Isaac A. ed) http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf (2013, accessed 11 May 2018).

[27] Suominen O, Hyvönen, E, Viljanen K and Hukka E. HealthFinland-a National Semantic Publishing Network and Portal for Health Information. Journal of Web Semantics 2009; 7(4), 287-297.

[28] Byrne K and Klein E. Automatic Extraction of Archaeological Events from Text. In: (Frischer, B, Crawford J and Koller D eds.). Making History Interactive. Proc. 37th Computer Application in Archaeology Conference, Williamsburg, 2010; pp. 48-56. Oxford: Archaeopress.

[29] Henninger M. From mud to the museum: Metadata challenges in archaeology. Journal of Information Science. 2017 online first. https://doi.org/10.1177/0165551517741790 (accessed 11 May 2018)

[30] Jeffrey S, Richards J, Ciravegna F, Waller S, Chapman S and Z. Zhang Z. The Archaeotools project: faceted classification and natural language processing in an archaeological context. In (Coveney P ed). Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures, Special Themed Issue of the Philosophical Transactions of the Royal Society A 2009; 367, pp 2507-2519.

[31] Richards Jeffrey S, Waller S, Ciravegna F, Chapman S and Zhang Z. The Archaeology Data Service and the Archaeotools project: faceted classification and natural language processing. In: (Kansa S, Kansa E and Watrall E eds.). Archaeology 2.0 and Beyond: New Tools for Collaboration and Communication. 2011; Los Angeles: Cotsen Institute of Archaeology Press, pp.31-56.

[32] Vlachidis A and Tudhope D. Negation detection and word sense disambiguation in digital archaeology reports for the purposes of semantic annotation. Program 2015; 49(2), 118 – 134.

[33] Vlachidis A and Tudhope D. A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology 2016; 67(5), 1138-1152.

[34] Archaeology Data Service (ADS) http://archaeologydataservice.ac.uk/ (accessed 11 May 2018).

[35] Data Archiving and Networked Services (DANS) https://dans.knaw.nl/ (accessed 11 May 2018).

[36] Swedish National Data Service (SND) https://snd.gu.se/en (accessed 11 May 2018).

[37] OpenRefine application http://openrefine.org/ (accessed 11 May 2018).

[38] Getty Art & Architecture Thesaurus as Linked Open Data, Getty Vocabulary Program http://vocab.getty.edu/ (accessed 11 May 2018).

[39] ARIADNE Portal. http://portal.ariadne-infrastructure.eu/ (accessed 11 May 2018).

[40] Jansma E, van Lanen R, Brewer, P and Kramer R. The DCCD: A digital data infrastructure for tree-ring research. Dendrochronologia 2012; 30(4), 249-251.

[41] Miles A and Bechhofer S. (eds.). SKOS Simple Knowledge Organization Reference [W3C recommendation, August 18, 2009] 2009; http://www.w3.org/TR/skos-reference/ (accessed 11 May 2018).

[42] Cunningham H, Maynard D, Bontcheva K. and Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications, Proc. 40th Annual Meeting of Association for Computational Linguistics 2002; New Brunswick, pp 168-175.

[43] Nadeau D., and Sekine S. A survey of named entity recognition and classification, Lingvisticae Investigationes 2007; 30(1), 3 – 26

[44] Nußbaumer P and Haslhofer B. Putting the CIDOC CRM into Practice – Experiences and Challenges. Technical Report TR-200, 2007; University of Vienna. https://eprints.cs.univie.ac.at/404/ (accessed 11 May 2018).

[45] Linked Art Data Model. https://linked.art/model/ (accessed 11 May 2018).

[46] STELETO open source code https://github.com/cbinding/steleto/ (accessed 11 May 2018).

[47] StringTemplate templating engine http://www.stringtemplate.org/ (accessed 11 May 2018).

[48] CIDOC Conceptual Reference Model (CRM) http://www.cidoc-crm.org/ (accessed 11 May 2018).

[49] Tree Ring Data Standard (TRiDaS) http://www.tridas.org/ (accessed 11 May 2018).

[50] OpenLink Virtuoso Universal Server https://virtuoso.openlinksw.com/ (accessed 11 May 2018).
[51] ARIADNE data integration case study demonstrator http://ariadne-lod.isti.cnr.it/ (accessed 11 May 2018).
[52] JQUERY UI Widget Factory https://jqueryui.com/widget/ (accessed 11 May 2018).
[53] Shiri A, Revie C and Chowdhury G. Thesaurus-enhanced search interfaces. Journal of Information Science 2002; 28 (2), 111–122.
[54] Tudhope D, Binding C, Blocks D and Cunliffe D. Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation 2006; 62 (4), 509-533.
[55] Alexiev V, Isaac A and Lindenthal J. On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI). International Journal on Digital Libraries 2016; 17(1), 39–48. Springer. https://link.springer.com/article/10.1007/s00799-015-0162-2 (accessed 11 May 2018)
[56] Gavel Y and Andersson P. Multilingual query expansion in the SveMed+ bibliographic database: A case study. Journal of Information Science 2014; 40(3), 269-280.
[57] Crofts N, Doerr M, Gill T, Stead S and Stiff M. (eds). Definition of the CIDOC Conceptual Reference Model, v5.0.4. 2011; http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf (accessed 11 May 2018).
[58] Tudhope D, Alani H and Jones C. Augmenting thesaurus relationships: possibilities for retrieval. Journal of Digital Information 2001; 1(8), http://journals.tdl.org/jodi/article/view/181/160 (accessed 11 May 2018).