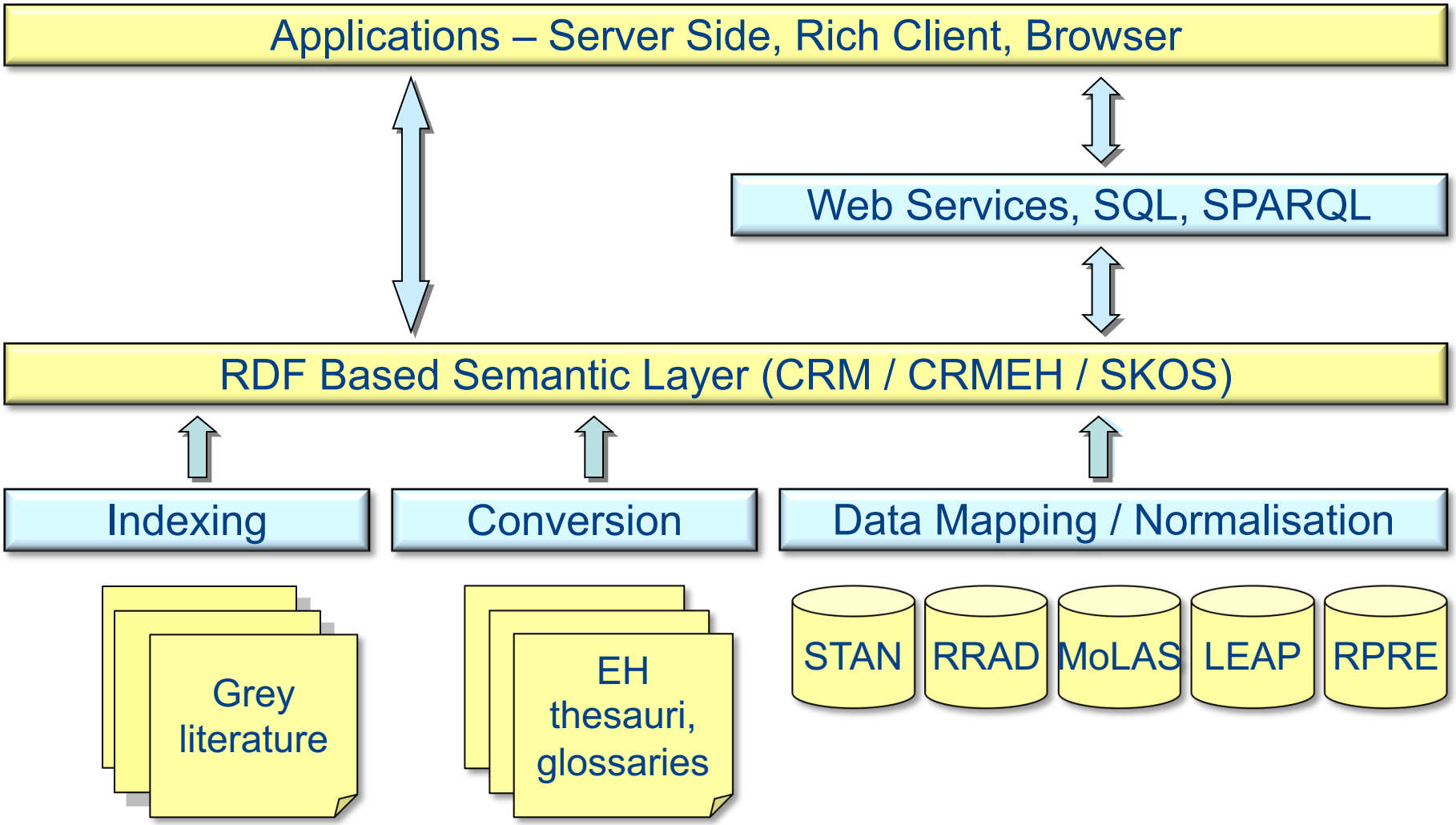
Reflections on KOS based data integration

Douglas Tudhope & Ceri Binding
Hypermedia Research Group,
University of South Wales

douglas.tudhope@southwales.ac.uk
ceri.binding@southwales.ac.uk

ARIADNE is funded by the European Commission's Seventh Framework Programme

STAR Project - General Architecture



Natural Language Processing (NLP)

of archaeological grey literature

Extract key concepts in same semantic representation as for data.

Allows unified searching of different datasets and grey literature

in terms of same underlying CRM-based conceptual structure

Output as RDF triples in Demonstrator and as XML with greylit

“ditch containing prehistoric pottery dating to the Late Bronze Age”

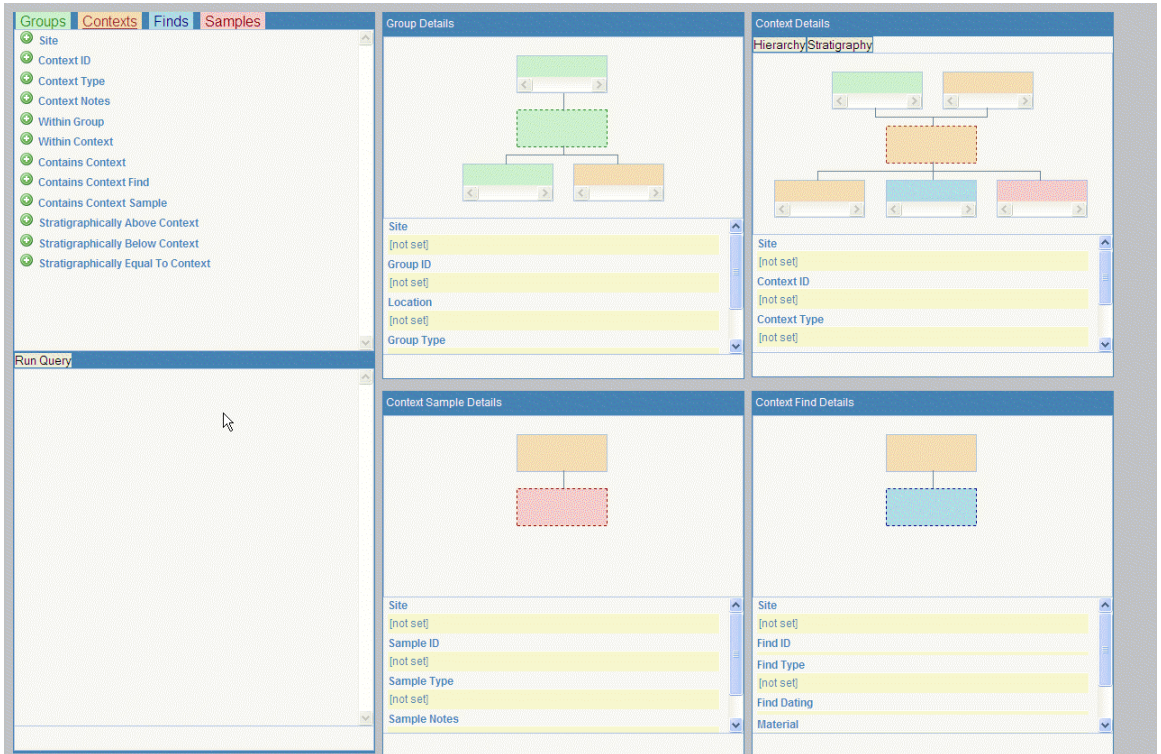
EHE1002.ContextFindProductionEvent	
prehistoric pottery dating to the Late Bronze Age	
EHE0009.ContextFind	EHE0039.TimeSpanAppellation
pottery [#ehg027.2]	Late Bronze Age [#134734]

EHE1004.ContextFindDepositionEvent	
ditch containing prehistoric pottery	
EHE0007.Context	EHE0009.ContextFind
ditch [#ehg003.20]	pottery [#ehg027.2]

STAR Demonstrator – search for a conceptual pattern

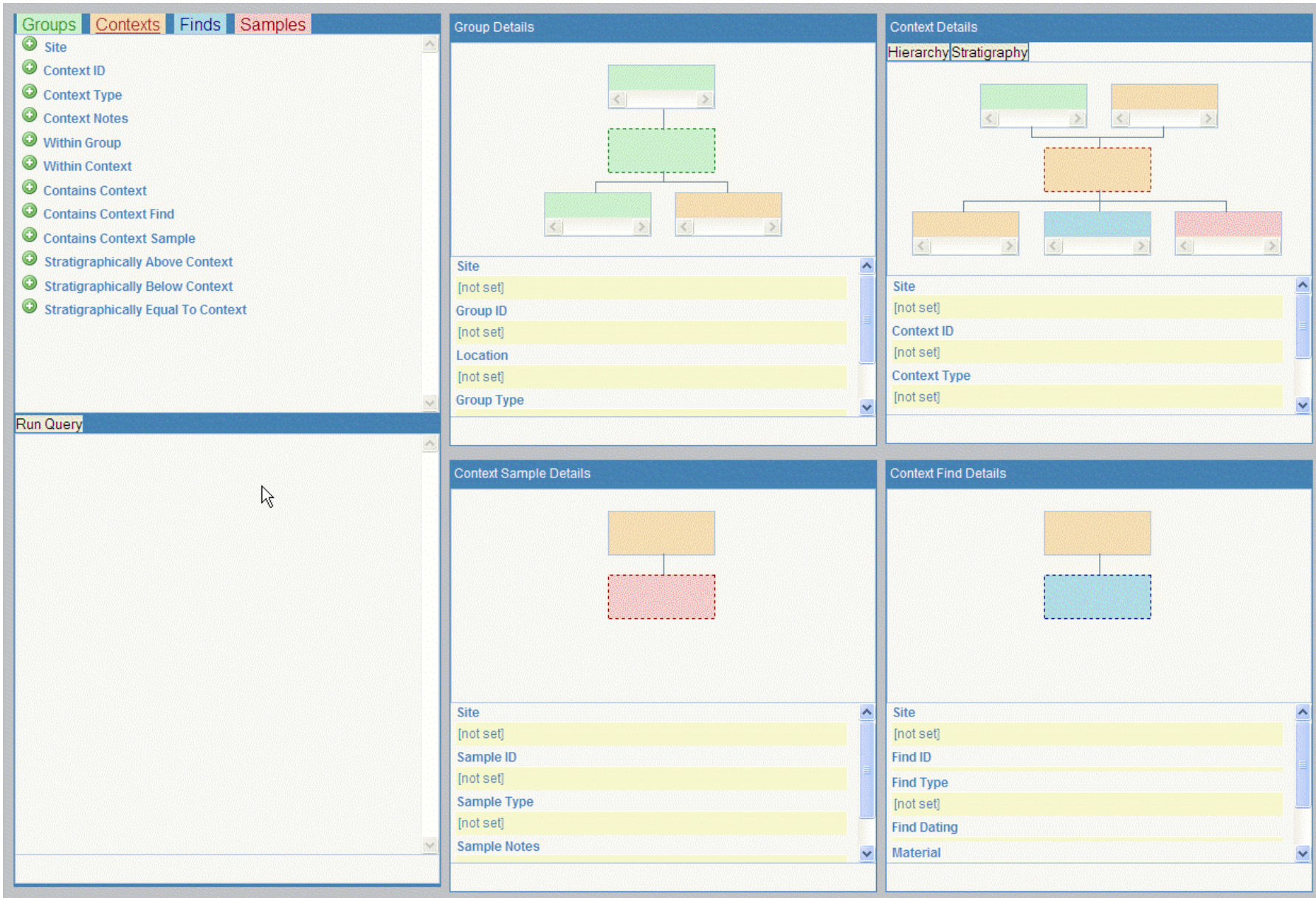
An Internet Archaeology publication on one of the (Silchester Roman) datasets we used in STAR discusses the finding of a *coin* within a *hearth*.
-- does the same thing occur in any of the grey literature reports?

Requires comparison of extracted data with NLP indexing in terms of the ontology.

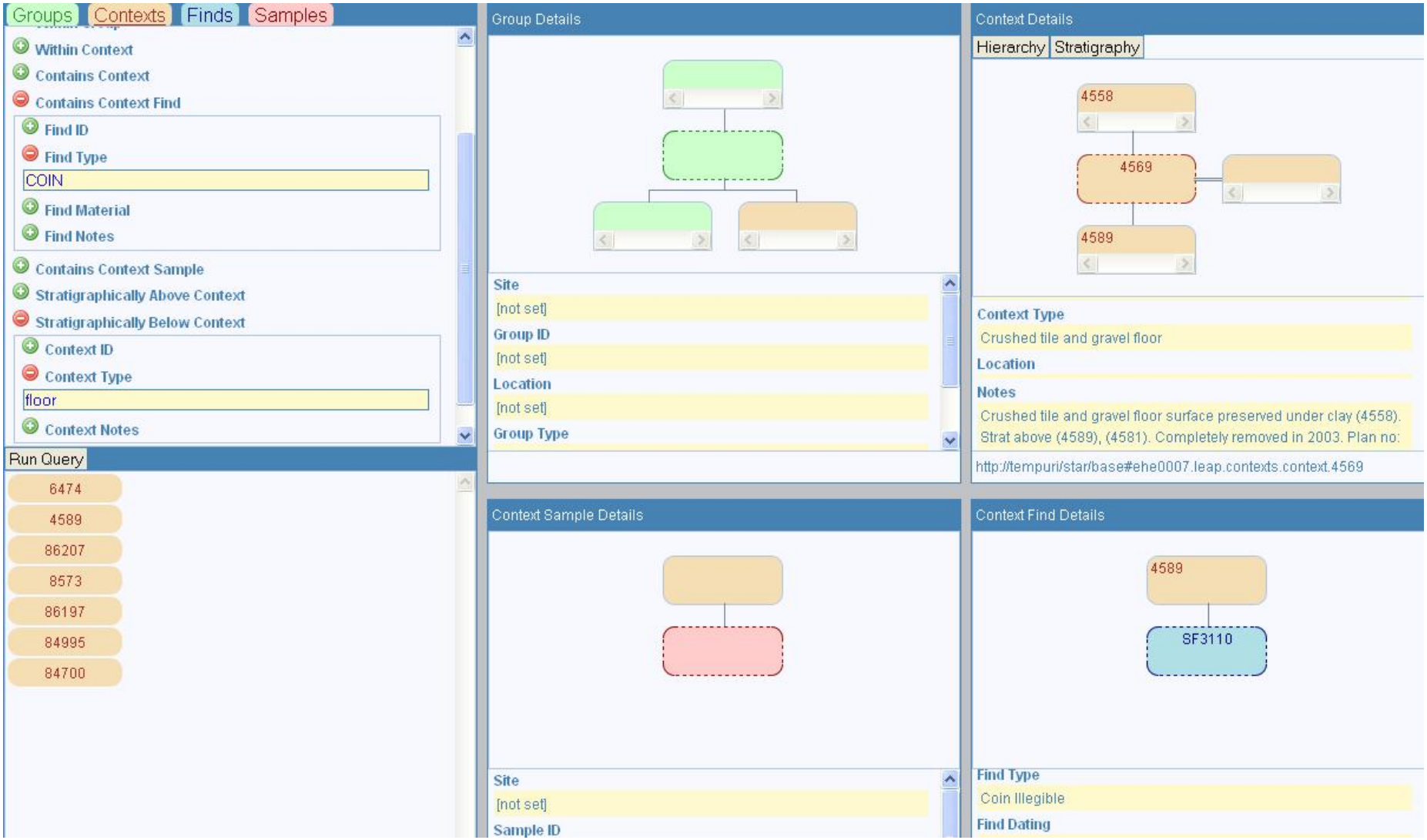


STAR Demonstrator – search for a conceptual pattern

Research paper reports finding a *coin in hearth* – exist elsewhere?



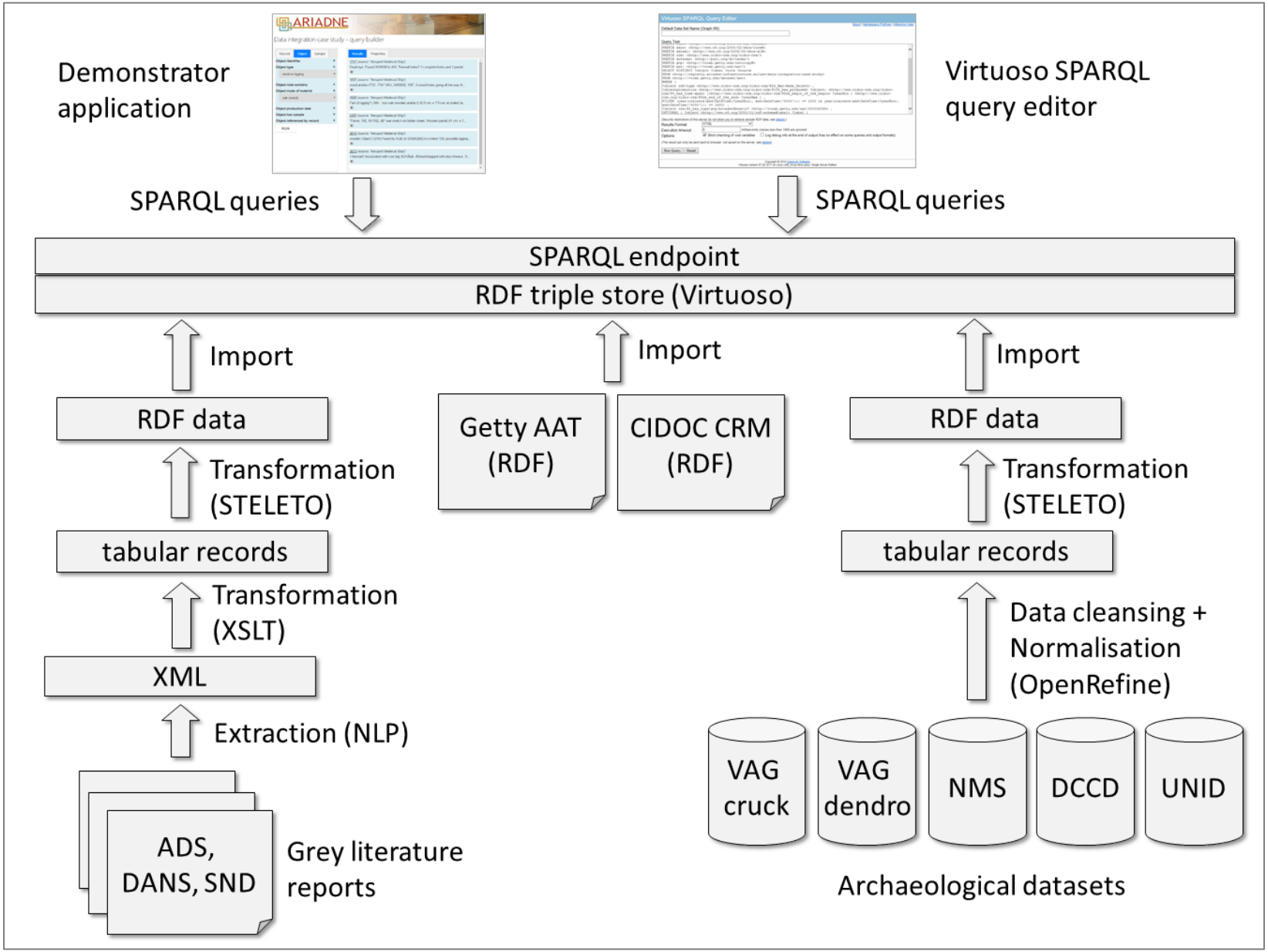
Stratigraphic query



Feasibility Study of Research Data Integration
- part of European ARIADNE project

- Extracts of 5 archaeological datasets, output from NLP on extracts from 25 grey literature reports
- broad theme of wooden material, objects and samples dated via dendrochronological analysis
- Multilingual - English, Dutch and Swedish data/reports
- Data integration via CIDOC CRM and Getty AAT
- 1.09 million RDF triples
- 23,594 records
- 37,935 objects
- Demonstration query builder for easier cross-search and browse of integrated datasets
- Concept based query expansion via AAT

General workflow and architecture



STELETO data conversion application

- A simpler, cross-platform version of the (previous project) STELLAR.Console application
- Performs bulk transformation of tabular delimited data via user-defined templates
- Cross platform (tested on Linux and Windows)
- Open source (https://github.com/cbinding/steleto)
- Flexible (can produce any textual output format)
- Simple, fast

ARIADNE vocabulary mapping to Getty AAT

- Subject metadata in different languages , so potentially:
  - useful resources missed
  - false results from homographs (eg 'coin’ French for corner, 'boot’ German for boat and 'monster’ Dutch for sample)
- Scalable solution – employ hub architecture

- Getty AAT adopted (available as LOD)
- Interactive (intellectual) mapping tools developed
  - generates SKOS mapping relationships in JSON and other formats
- Mapping guidelines produced
- 6416 concepts (27 vocabularies, 12 partners) mapped

NLP methods

- Rule based Named Entity Recognition (NER) pipelines for English, Dutch, and Swedish text using GATE platform
- Builds on previous English language NLP work on archaeological grey literature
- Supported by a controlled vocabulary based on Getty AAT with mappings to Dutch and Swedish vocabulary
- Intermediate XML output with inline mark-up transformed to same RDF format as for datasets
- Different strategies explored for identifying potentially relevant material (manual, automatic)

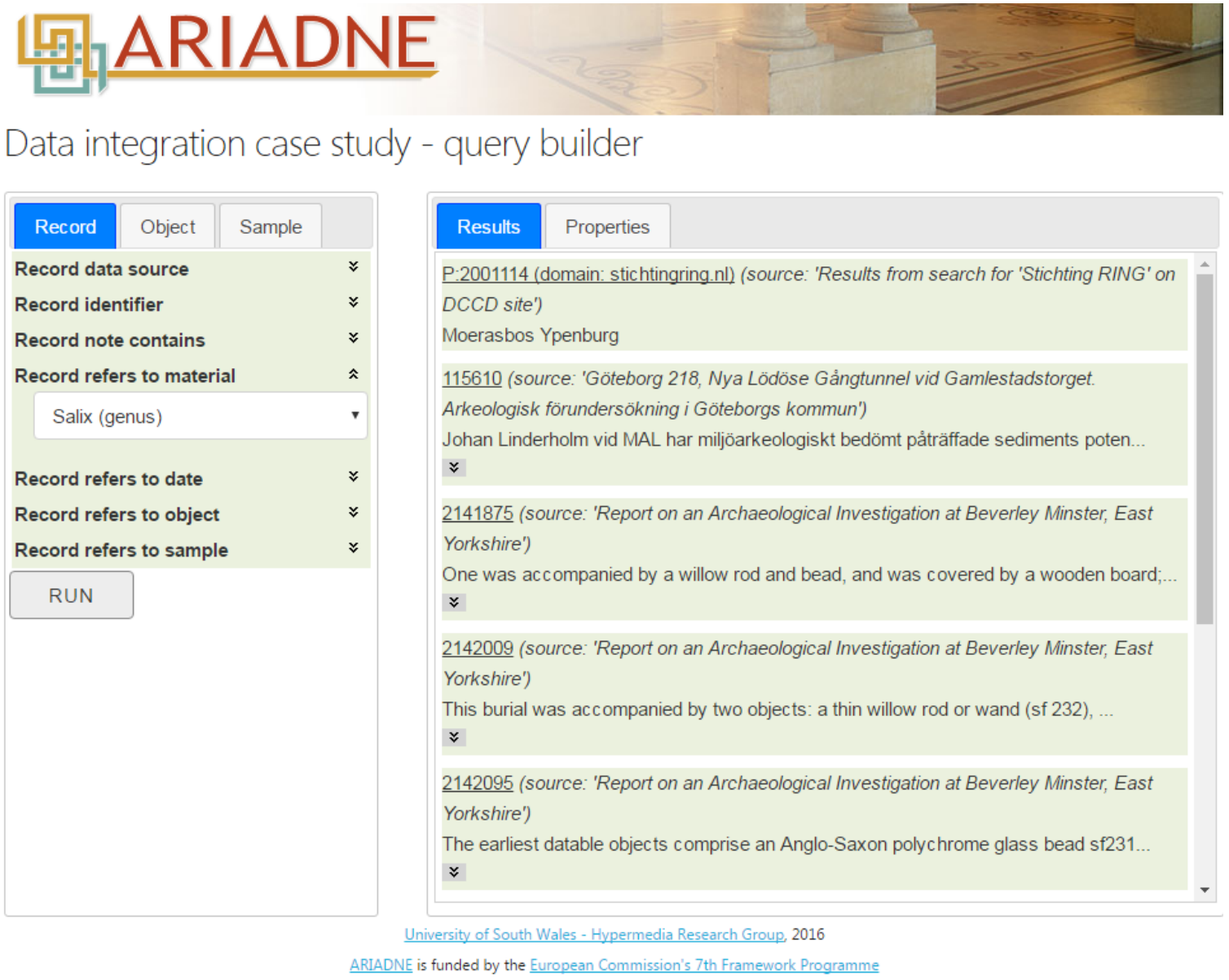
Illustrative examples of NLP output

Examples illustrating English, Dutch and Swedish NLP output (before transformation to RDF), with colour coding objects, materials, dates, samples):

Two timbers dated from the west wing roof produce felling dates in the winter of AD 1735/6 and the spring of AD 1736.

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.

Prov 1 som var bearbetat virke av ek daterades till fällningsår vinterhalvåren 1536/37.



Query Builder (query on left, results on right): Records referring to material “Salix (genus)”
Shows English, Dutch & Swedish results, originating from NLP and database records

Leveraging thesaurus structure

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


AAT Taxonomic structure for concept 300375384 (not a formal Scientific taxonomy)
--------------------------------------------------------------------------------
Agents Facet                                         aat:300264089
-  Living Organisms (hierarchy)                      aat:300265673
-  - - living Organisms (entities)                   aat:300390503
-  - - - Eukaryota (domain)                          aat:300265677
-  - - - - Plantae (kingdom)                         aat:300132360
-  - - - - - Angiospermae (division)                 aat:300265706
-  - - - - - - Magnoliopsida (class)                 aat:300375593
-  - - - - - - - Malpighiales (order)                aat:300374936
-  - - - - - - - - Salicaceae (family)               aat:300374937
-  - - - - - - - - - salix (genus)                   aat:300375384
-  - - - - - - - - - - salix lucida (species)        aat:300375387
-  - - - - - - - - - - - Salix lucida ssp caudata    aat:300375389
```

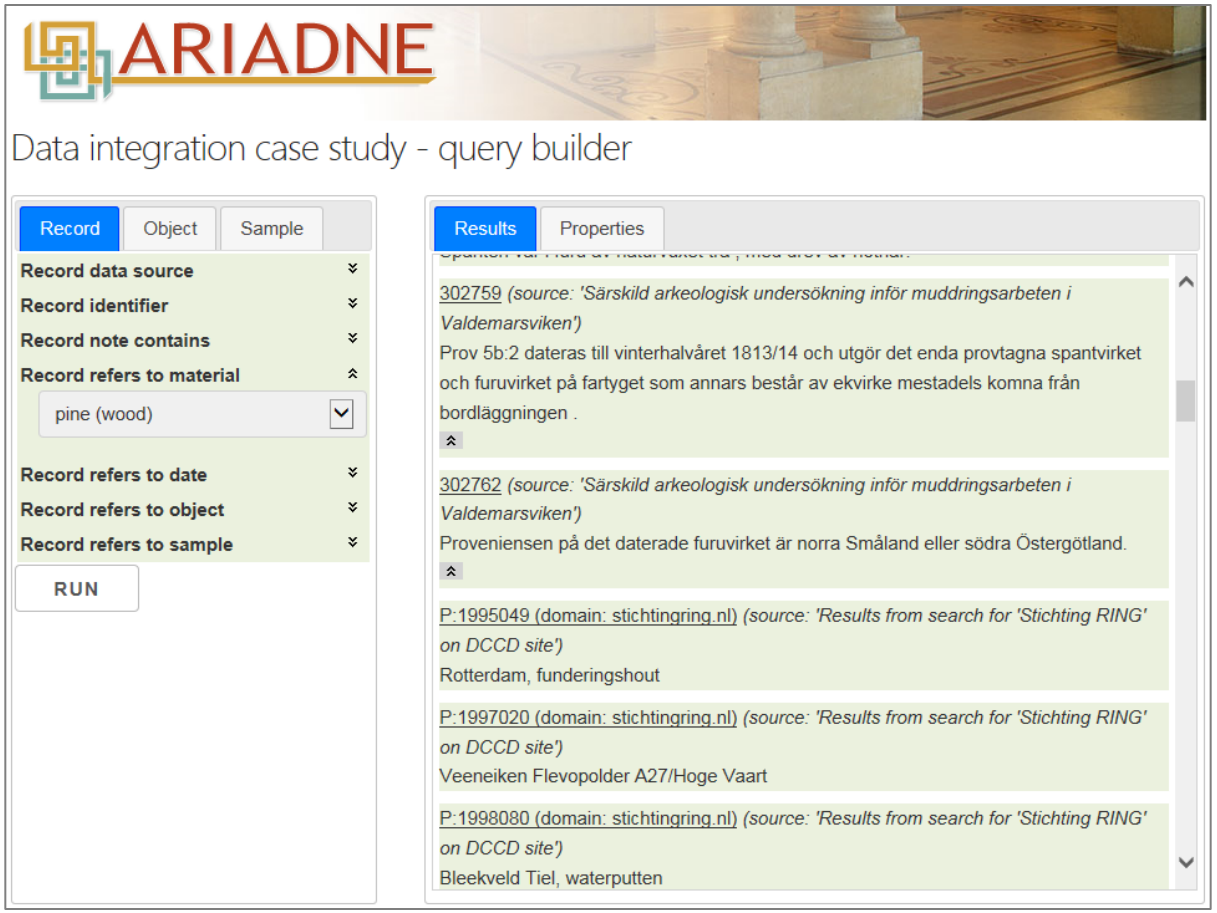
References to wood in datasets (and grey literature) often use material/family/genus/species interchangeably.

For more effective search employ the link between the material (type of wood) and the agent (living organism) in AAT this is a specific GVP RT specialisation and its reciprocal (inverse) relationship. e.g.:

aat:300012498 gvp:2841_derived-made_from aat:300375384 .
"willow (wood)" derived/made-from "Salix (genus)" .
aat:300375384 gvp:aat2842_source_for aat:300012498 .
"Salix (genus)" source for "willow (wood)" .

A search on e.g. "willow (wood)" can retrieve the Material [aat:300012498], the Agent [aat:300375384] and their respective hierarchical descendant concepts.

Leveraging thesaurus structure



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)”

Design decisions

- KOS-based development efforts involve design choices
- Usually impractical to develop parallel implementations to compare major design alternatives and thus not easy to know the consequences of one design choice over another
- Reflecting on some major design decisions encountered during the two projects, with a view to informing future work …

Design decisions 1

- How to select datasets, how much to model

  How much of the source datasets and reports should be extracted, aligned to KOS and expressed as linked data? Should it be a subset *(USW case studies)* or as much as possible *(which is possibly usual CRM schema based approach)*?

- How to match datasets, reports, research questions
  - *An operational project should budget resources to locate key datasets and reports to address a particular research question (addressing issues of access and permission)*
- Should native schema of the source datasets be maintained in the resulting integration *(in Dutch Ships and Sailors linked data cloud – datasets converted to RDF using own data model and enriched with links to connect to interoperability layer)* or replaced by the new semantic framework *(USW case studies)*?

Design decisions 2

- Appropriate balance of application modeling detail, expressed between ontology and vocabulary side. How much to handle via the ontology and how much to handle via the thesaurus (or other vocabulary)? How much detail is it worthwhile to model?
  - *Not go beyond original data semantics … Depends on use cases*

➔ISO 25964 Part 2 (ch21)

One of the fundamental purposes of an ontology is reasoning, including generic tasks such as:

— inferring class membership for individuals;

— inferring relationships between classes and properties; and

— checking the consistency of a knowledge base

… Whereas the role of most of the vocabularies described in this part of ISO 25964 is to guide the selection of search/indexing terms, or the browsing of organized document collections, the purpose of ontologies in the context of retrieval is different. Ontologies are not designed for information retrieval by index terms or class notation, but for making assertions about individuals, e.g. about real persons or abstract things such as a process. …

Design decisions 3

- How to mitigate the possibility of creating alternative (valid) ontology mapping expressions of the same underlying semantics from different sources and thus make cross search and interoperability difficult?

➔

- *Mapping pattern based approach (in our case the template based STELLAR/STELETO tools)*
  http://hypermedia.research.southwales.ac.uk/resources/STELLAR-applications/

- *Similarly see Linked Art project (also using CRM and AAT)*
  https://linked.art

Design decisions 4

- Both projects required substantial data cleansing. How represent the new information, what is the relationship with the source dataset? - *replaced by new semantic framework?*

  Examples encountered

  - obvious spelling errors, reordering of words
  - Additional prefixes or suffixes (e.g. "*red hill (possible)*", "*trackway (cobbled)*", "*croft?*", "*portal dolmen (re-erected)*")
  - attempts at providing additional structure within a single field (e.g. "*pottery;ceramic tile;iron objects;glass*")
  - very specific compound phrases (e.g. "*side wall of pot with lug*")
  - how to represent 'non-information' values?
    - unstated NULL values or empty strings
    - *known unknowns* "not known", "blank", "null", "nothing", "void", "not specified", "unspecified", "uncertain", "missing", "empty".

Design decisions 5

- How to express information extracted via NLP? How much certainty to associate with the derived data, what kinds of elements are represented (archaeological texts often refer to types of object or material rather than named specific individual items)?
- How to express results from search over both data and textual reports, how to express the provenance of the subject metadata extracted and also the method by which it was extracted?

➔ Future work identifying passages of particular relevance for NLP information extraction (or sections to avoid).
STAR project focused mainly on report abstracts

References

- ARIADNE. http://www.ariadne-infrastructure.eu
- ARIADNE Portal. http://portal.ariadne-infrastructure.eu/
- Data Integration study Demonstrator. http://ariadne-lod.isti.cnr.it/description.html
- STELETO open source code. https://github.com/cbinding/steleto/
- Binding C. & Tudhope D. 2016. Improving Interoperability using Vocabulary Linked Data. International Journal on Digital Libraries, 17(1), 5-21
- Tudhope D, May K, Binding C, Vlachidis A.2011. Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, https://doi.org/10.11141/ia.30.5 (open access),

Recent paper on second study

Binding C, Tudhope D, Vlachidis A. (2018) A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science, Sage. https://doi.org/10.1177/0165551518789874 - see below for OA version.

Open Access versions of Hypermedia Research Group's KOS papers are available from https://bit.ly/2ocaHC6

Acknowledgments

- Andreas Vlachidis (NLP)

Thank you

ARIADNE is a project funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193.