Multilingual vocabulary mapping in ARIADNEplus

Ceri Binding and Douglas Tudhope
Hypermedia Research Group
University of South Wales (USW)

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

ARIADNEplus is funded by the European Commission's Horizon 2020 Programme

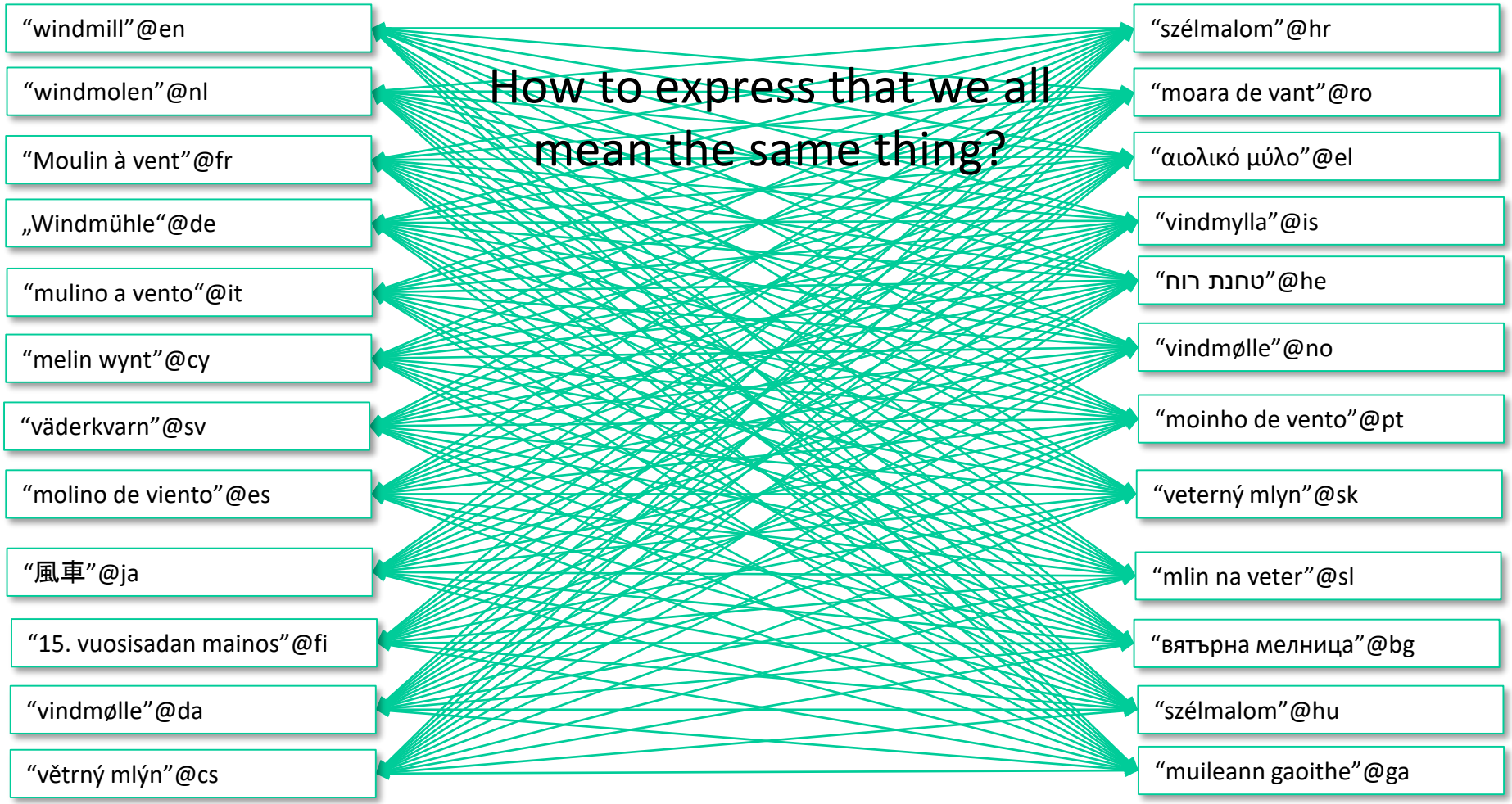
Vocabulary mapping - why?

- Original datasets not necessarily produced with aggregation, consolidation, reuse and cross-search in mind
- I say "*potato*", you say "*pomme de terre*", she says "*maris piper*", he says *"seedling X8/5"*
- Multiple barriers to cross-searching subject metadata

  language, punctuation, spelling, homonyms, synonyms, level of specificity
- Text-based search is limited by all of these
- Need to establish common meaning

- X8/5 'Commendation' in Immunity and Merit Trials, 1963. https://marispiperfifty.wordpress.com/maris-piper/recomendation-of-maris-piper/

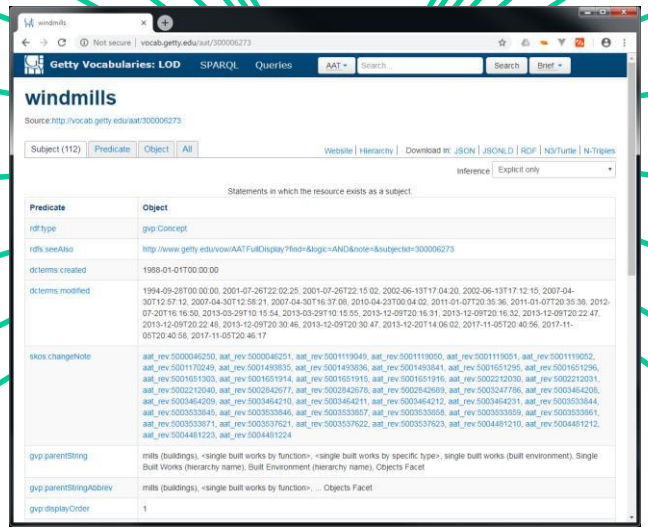
Multilingual subject metadata



Mapping local terms to a central concept

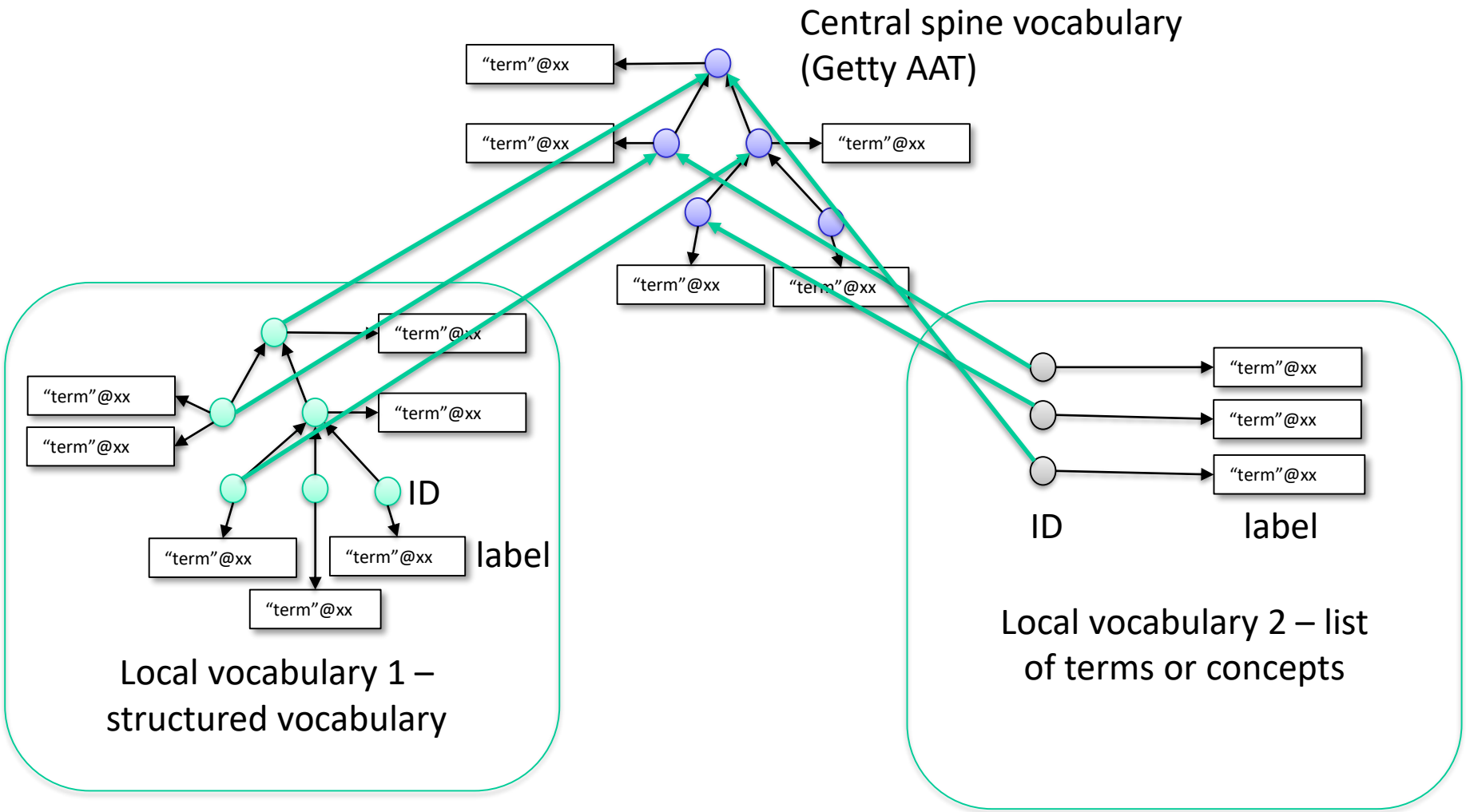
The *words* may be different, but the *concept* is (more or less) the same…

- “windmill”@en
- “windmolen”@nl
- “Moulin à vent”@fr
- „Windmühle“@de
- “mulino a vento“@it
- “melin wynt”@cy
- “väderkvarn”@sv
- “molino de viento”@es
- “風車”@ja
- “15. vuosisadan mainos”@fi
- “vindmølle”@da
- “větrný mlýn”@cs



- “szélmalom”@hr
- “moara de vant”@ro
- “αιολικό μύλο”@el
- “vindmylla”@is
- “טחנת רוח”@he
- “vindmølle”@no
- “moinho de vento”@pt
- “veterný mlyn”@sk
- “mlin na veter”@sl
- “вятърна мелница”@bg
- “szélmalom”@hu
- “muileann gaoithe”@ga

Mapping local concepts to a central spine



Multilingual enrichment via AAT

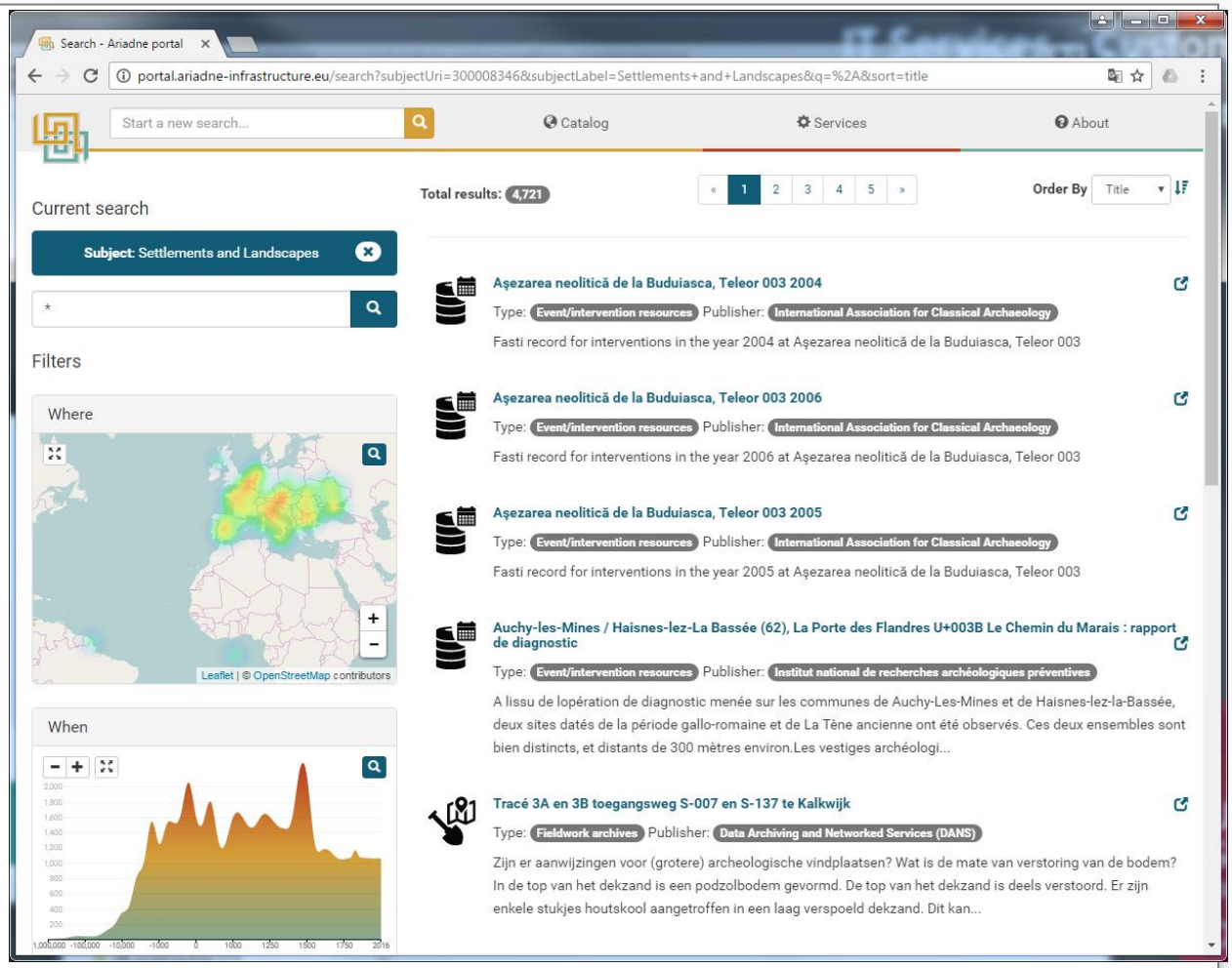
- ARIADNE Registry subject enrichment service derived AAT concepts that augmented subject metadata for partner resources
- When applied to ARIADNE portal this allowed the concept-based search functionality to retrieve records with metadata expressed in different languages via the AAT concepts - the AAT acting as a mapping spine
- When applied to the data integration case studies, we explored the possibility of integrating research data and archaeological grey literature in different languages via the core ontology and value vocabularies

Concept based search in ARIADNE Portal via AAT

ARIADNE Portal Query on AAT subject: *Settlements and Landscapes*

shows

results from IACA (Fasti), INRAP and DANS in multiple languages



ARIADNE Multilingual Data Integration Feasibility Study

- 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
- RDF data - 1.09 million RDF triples
- 23,594 records referencing 37,935 objects
- Demonstration query builder for easier cross-search and browse of integrated datasets
- Concept based query expansion via AAT

Data transformation - STELETO

- Open Source tool for fast bulk transformation of delimited data used in ARIADNE multilingual data integration study
- Uses DotLiquid template engine http://dotliquidmarkup.org/
- Recently used by Historic England for transformation of vocabularies to SKOS RDF for publishing as Linked Open Data https://heritagedata.org/live/schemes.php
- https://github.com/cbinding/STELETO

Mappings

In ARIADNE I, concepts from 27 vocabularies from 12 data partners were mapped to Getty AAT

Mappings by individual partners ranging from a few to over 1600 concepts following guidelines

6416 mappings total

Most at similar level of generality
Some partner vocabs more specialised than AAT but in a few cases AAT was more specialised

Source	Vocabularies mapped to AAT	Match type						Totals
		No match	skos:exactMatch	skos:closeMatch	skos:broadMatch	skos:narrowMatch	skos:relatedMatch	
ADS	FISH Archaeological Objects Thesaurus (subset)	0	197	96	118	0	0	411
ADS	FISH Building Materials Thesaurus (subset)	0	8	4	0	0	0	12
ADS	FISH Thesaurus of Monument Types (subset)	0	141	107	141	0	1	390
ADS	Historic England Components Thesaurus (subset)	0	7	1	1	0	0	9
ADS	Historic England Maritime Craft Thesaurus (subset)	0	13	8	3	0	0	24
DAI	ARACHNE - books	0	13	4	0	0	0	17
DAI	ARACHNE - collections	0	8	2	1	0	0	11
DAI	ARACHNE - inscriptions	0	18	1	0	0	0	19
DAI	ARACHNE - buildings and structures	0	81	37	44	14	0	176
DAI	ARACHNE - multi-part monument	0	51	35	22	0	0	108
DAI	ARACHNE - topographic objects	0	46	7	2	0	0	55
DANS	DCCD vocabulary	0	245	9	82	0	0	336
DANS	EASY - Complextypen	3	3	59	34	0	15	114
Discovery	Irish Monument Types	0	168	69	249	0	0	486
Discovery	National Inventory of Architectural Heritage	0	211	59	78	7	0	355
IACA	FASTI Monument Types	2	23	80	24	0	0	129
ICCU	ICCD RA and PICO thesaurus (subset)	0	642	94	310	258	0	1304
INRAP	PACTOLS thesaurus (subset)	0	1161	121	346	6	0	1634
MNM-NOK	site types	0	0	34	7	0	0	41
NIAM-BAS	AIS AKB database - subject terms	0	50	85	92	11	0	238
OEAW	DFMROE DB	0	4	0	0	3	0	7
OEAW	Franzhausen Kokoern DB	0	5	2	2	1	0	10
OEAW	UK Material Pool DB	0	7	0	4	5	0	16
OEAW	UK Thunau DB	0	3	1	0	0	0	4
SND	combined terms list	5	71	156	144	11	0	387
ZRC-SAZU	ZBIVA vocabulary	0	0	25	5	0	0	30
ZRC-SAZU	ARKAS vocabulary	0	0	76	17	0	0	93
ADS	**5**	**0**	**366**	**216**	**263**	**0**	**1**	**846**
DAI	**6**	**0**	**217**	**86**	**69**	**14**	**0**	**386**
DANS	**2**	**3**	**248**	**68**	**116**	**0**	**15**	**450**
Discovery	**2**	**0**	**379**	**128**	**327**	**7**	**0**	**841**
IACA	**1**	**2**	**23**	**80**	**24**	**0**	**0**	**129**
ICCU	**1**	**0**	**642**	**94**	**310**	**258**	**0**	**1304**
INRAP	**1**	**0**	**1161**	**121**	**346**	**6**	**0**	**1634**
MNM-NOK	**1**	**0**	**0**	**34**	**7**	**0**	**0**	**41**
NIAM-BAS	**1**	**0**	**50**	**85**	**92**	**11**	**0**	**238**
OEAW	**4**	**0**	**19**	**3**	**6**	**9**	**0**	**37**
SND	**1**	**5**	**71**	**156**	**144**	**11**	**0**	**387**
ZRC-SAZU	**2**	**0**	**0**	**101**	**22**	**0**	**0**	**123**
Totals:	**27**	**10**	**3176**	**1172**	**1726**	**316**	**16**	**6416**
	Proportion:	**0.16%**	**49.50%**	**18.27%**	**26.90%**	**4.93%**	**0.25%**	**100%**

Expressing vocabulary matches

	A	B	C	D	E	F
1	**sourceLabel**	**matchURI**	**targetLabel**	**targetURI**	**Source-Hierarchy**	**Source-ScopeNote**
14	begravningsplats	skos:closeMatch	burial sites	http://vocab.getty.edu/aat/300387004	FMIS	historisk tid
15	begravningsplats, enst	skos:closeMatch	burials	http://vocab.getty.edu/aat/300263485	FMIS	historisk tid avsedd för en
16	bengömma	skos:closeMatch	remains	http://vocab.getty.edu/aat/300265420	FMIS	märgkluvna ben i skyddat
17	bergshistorisk lämning	skos:closeMatch	mine structures	http://vocab.getty.edu/aat/300006423	FMIS	bergshantering som inte
18	bildristning	skos:broadMatch	rock carvings	http://vocab.getty.edu/aat/300080131	FMIS	eller slipade bilder av
19	björngrav	skos:broadMatch	graves	http://vocab.getty.edu/aat/300005907	FMIS	björnben
20	blästbrukslämning	skos:closeMatch	bloomeries	http://vocab.getty.edu/aat/300379639	FMIS	(lågteknisk
21	blästplats (sammansat	skos:closeMatch	bloomeries	http://vocab.getty.edu/aat/300379639	FMIS	lämningar efter
22	boplats	skos:closeMatch	buried settlements	http://vocab.getty.edu/aat/300387241	FMIS	förhistorisk tid vistats och
23	boplatsgrop	skos:broadMatch	pits (earthworks)	http://vocab.getty.edu/aat/300008027	FMIS	uppgrävda materialet
24	boplatslämning övrig	skos:broadMatch	buried settlements	http://vocab.getty.edu/aat/300387243	FMIS	boplatslämningar som inte
25	boplatsvall	skos:closeMatch	sunken huts	http://vocab.getty.edu/aat/300137527	FMIS	omger eller avgränsar en
26	borg	skos:broadMatch	fortifications	http://vocab.getty.edu/aat/300006888	FMIS	kombinationer av murar,
27	bro	skos:exactMatch	bridges (built works)	http://vocab.getty.edu/aat/300007836	FMIS	väg, järnväg, kanal eller
28	brott/täkt	skos:broadMatch	extracting complexes/	http://vocab.getty.edu/aat/300000388	FMIS	utnyttjats för utvinning elle
29	brunn	skos:broadMatch	wells (structures)	http://vocab.getty.edu/aat/300006207	FMIS	åtkomst till färskvatten
30	brytningsyta	skos:closeMatch	quarries (extracting complex	http://vocab.getty.edu/aat/300000402	FMIS	bergart eller mineral för
31	byggnad annan	skos:closeMatch	buildings (structures)	http://vocab.getty.edu/aat/300004792	FMIS	kulturhistoriskt värde. (OBS
32	byggnadsminne	skos:closeMatch	listed buildings	http://vocab.getty.edu/aat/300343491	FMIS	byggnad i privat ägo
33	bytomt	skos:closeMatch	greens (open spaces)	http://vocab.getty.edu/aat/300008164	FMIS	eller mantalssatt
34			landings (marine structures)	http://vocab.getty.edu/aat/300007928	FMIS	Stenröjd uppdragningsplats
35	dammvall	skos:broadMatch	dikes	http://vocab.getty.edu/aat/300170882	FMIS	kunna ansamla eller
36	depåfynd	skos:closeMatch	hoards (groupings)	http://vocab.getty.edu/aat/300195474	FMIS	föremål som kan antas ha

- Simple approach in ARIADNE was a spreadsheet template for term lists and vocabularies
- Partner domain experts specified mappings from source terms to Getty AAT concepts, following examples and guidelines, with assistance where required
- Resulting mappings were transformed to appropriate format for ingest to ARIADNE semantic framework
- Mappings facilitated concept based multilingual searching and browsing

What will we need in ARIADNEplus?

- Identify subject metadata relating to local datasets
  - Thesauri / glossaries / gazetteers, authority files, term lists
    or maybe just a list of distinct terms from a particular data field
- Consider data cleaning (where necessary)
- Our starting point is to reuse / extend existing ARIADNE mappings
- We can assist in producing new mappings
- Vocabulary mapping tool (first version) on a Virtual Research Environment on D4Science platform
- https://vmt.ariadne.d4science.org/vmt/

Type of match between concepts

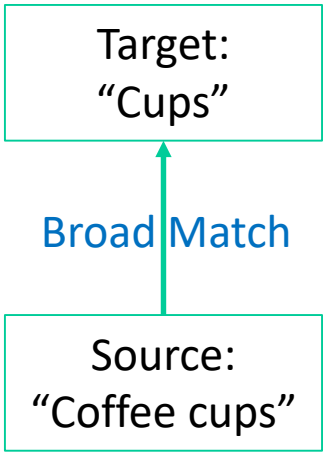
Exact Match

Don’t rely on label match; consider full context – meaning and scope of concepts

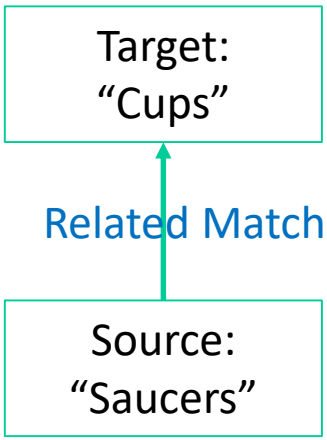
Close Match

Where scope or context of concepts suggests conceptual slight differences

Some/all rule for generic hierarchical relationships



Some cups are coffee cups; All coffee cups are cups

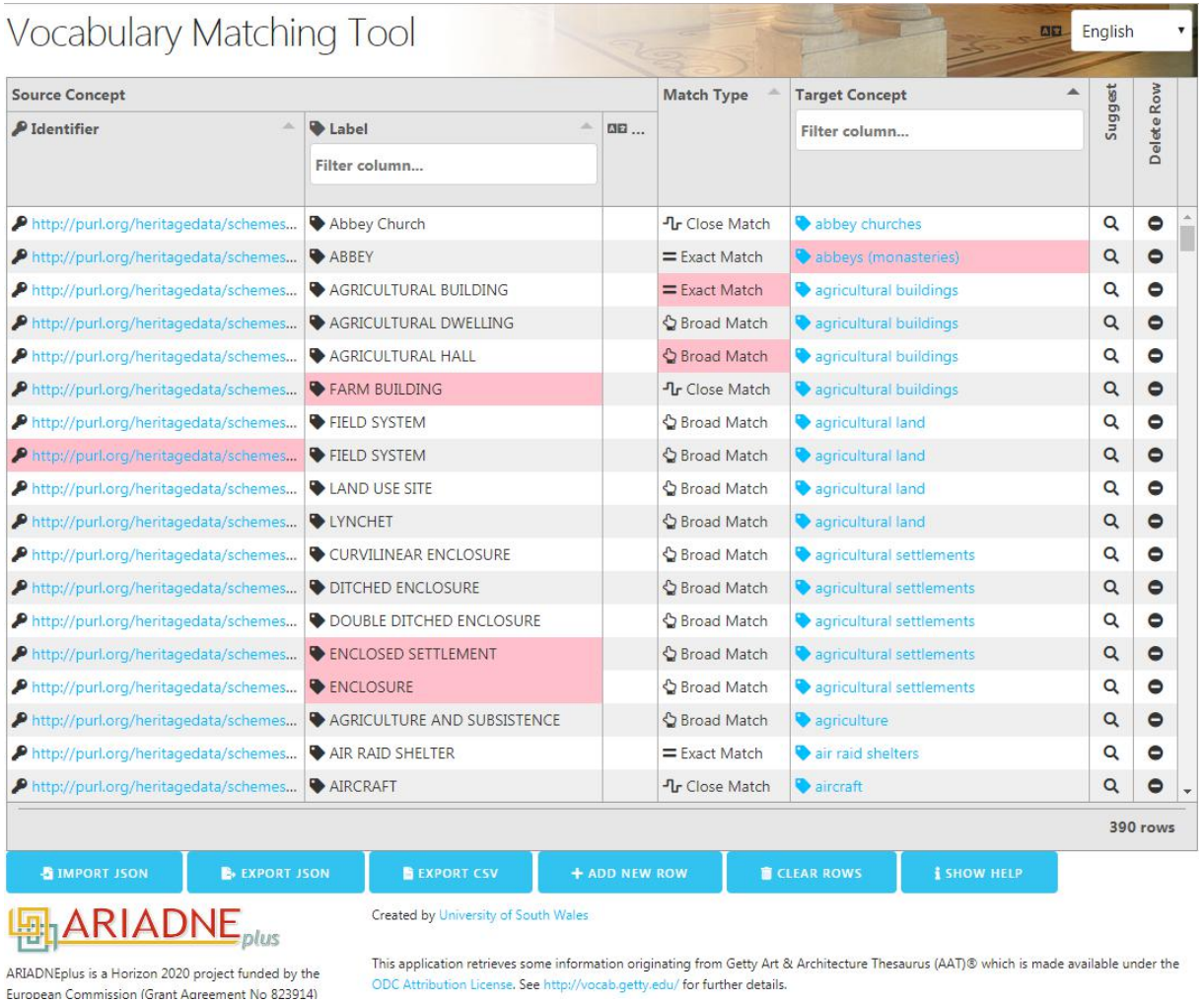


Some other association between concepts. Wherever possible prefer one of the other match types

Vocabulary Mapping Tool

https://vmt.ariadne.d4science.org/vmt/

- For matching subject terms / concepts to AAT concepts
- Search & browse AAT
- Decide match by examining scope and context of source / target
- Can input existing mappings
- Variety of export formats



RDF serialisations of mappings

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!--Example mappings expressed in RDF-XML serialization format ()-->
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">

    <skos:Concept rdf:about="http://snd.gu.se/sv/catalogue/keyword/bengomma">
        <skos:prefLabel xml:lang="sv">bengömma</skos:prefLabel>
        <skos:closeMatch rdf:resource="http://vocab.getty.edu/aat/300265420"/><!--remains-->
    </skos:Concept>

    <skos:Concept rdf:about="http://snd.gu.se/sv/catalogue/keyword/bergshistorisk-lamning-ovrig">
        <skos:prefLabel xml:lang="sv">bergshistorisk lämning övrig</skos:prefLabel>
        <skos:closeMatch rdf:resource="http://vocab.getty.edu/aat/300006423"/><!--mine structures-->
    </skos:Concept>

    <skos:Concept rdf:about="http://snd.gu.se/sv/catalogue
        <skos:prefLabel xml:lang="sv">bildristning</skos:p
        <skos:broadMatch rdf:resource="http://vocab.getty.
    </skos:Concept>

    <skos:Concept rdf:about="http://snd.gu.se/sv/catalogue
        <skos:prefLabel xml:lang="sv">björngrav</skos:pref
        <skos:broadMatch rdf:resource="http://vocab.getty.
    </skos:Concept>
</rdf:RDF>
```

```turtle
# Mappings expressed in Turtle RDF serialization format
@prefix data: <http://snd.gu.se/sv/catalogue/keyword/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix aat: <http://vocab.getty.edu/aat/> .

data:bengomma a skos:Concept;
    skos:prefLabel "bengömma"@sv ;
    skos:closeMatch aat:300265420 . # remains

data:bergshistorisk-lamning-ovrig a skos:Concept ;
    skos:prefLabel "bergshistorisk lämning övrig"@sv ;
    skos:closeMatch aat:300006423 . # mine structures

data:bildristning a skos:Concept;
    skos:prefLabel "bildristning"@sv ;
    skos:broadMatch aat:300080131 . # rock carvings

data:bjorngrav a skos:Concept;
    skos:prefLabel "björngrav"@sv ;
    skos:broadMatch aat:300005907 . # graves
```

Expanded entry vocabulary?

- Considering a multilingual dictionary service for archaeological terminology as a search tool, building on Wikidata multilingual resources and other sources

  eg https://www.wikidata.org/wiki/Q11761

Cf Joachim Neubert, NKOS 2017 (and also see DCMI 2018)

- *Wikidata as a linking hub for knowledge organization systems? Integrating an authority mapping into Wikidata and learning lessons for KOS mappings*
- http://ceur-ws.org/Vol-1937/paper2.pdf
  http://zbw.eu/stw/version/latest/mapping/wikidata/about.en.html

Mapping Guidelines

- Aim to support search and browsing (rather than logical inferencing), hence a rough subject mapping is ok
- Usually just make one match (the best one) for source concept
  - no need to express multiple relationships to AAT concepts
    as this is provided gratis via the AAT's semantic structure
- The exception is where the source concept relates to two genuinely different AAT concepts
- Use one of the SKOS mapping properties (in case the search functionality is able to make distinctions)
- Mappings should be made to AAT *concepts* rather than guide-terms (inside <>). If an AAT guide term appears as a match in the tool, consider a narrower or broader concept in the AAT.

Ontology vs Thesaurus?

- What is the appropriate balance between ontology and vocabulary? How much to handle via the ontology and how much to handle via the thesaurus (or other vocabulary)?

➔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. …

References

- ARIADNE. http://www.ariadne-infrastructure.eu
- ARIADNE Portal. http://portal.ariadne-infrastructure.eu/
- 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
- Binding C, Tudhope D, Vlachidis A. (2019) A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science, 45(3), 364-386. 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

THANK YOU!

ARIADNEplus is a project funded by the European Commission under the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914.

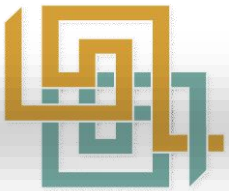
The views and opinions expressed in this presentation are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

Contact:

ceri.binding@southwales.ac.uk

douglas.tudhope@southwales.ac.uk

http://www.ariadne-infrastructure.eu/



Vocabulary mapping tool

- https://heritagedata.org/vmt2/ (temporary address)

- The vocabulary matching tool is now working on the new Virtual Machine on D4Science platform
- https://vmt.ariadne.d4science.org/vmt/