University of South Wales
Prifysgol De Cymru

Vocabulary Linked Data Publication and Mapping

Ceri Binding, Douglas Tudhope
University of South Wales
ceri.binding@southwales.ac.uk

University of South Wales
Prifysgol De Cymru

ARIADNE Project

- **A**dvanced **R**esearch **I**nfrastructure for **A**rchaeological **D**ataset **N**etworking in **E**urope
  - http://www.ariadne-infrastructure.eu/
  - 4 year FP7 project: February 2013 – January 2017
  - 24 European partner organisations
  - Multiple languages, multiple controlled vocabularies
  - Millions of consolidated metadata records from partners
- Consolidating metadata does not make it any more interoperable
- Adoption of common metadata schema plus use of suitable controlled vocabularies are the real keys to interoperability

University of South Wales Prifysgol De Cymru

SENESCHAL Project

- AHRC funded project: March 2013 – February 2014
- Project aims:
  - Linked Open Data publication of Cultural Heritage vocabularies
  - Widening access to key vocabulary resources
  - Facilitating improved consistency for existing and future metadata
- Project outcomes:
  - 14 vocabularies converted to SKOS format and made available online as Linked Open Data (see www.heritagedata.org)
  - Associated vocabulary web services and functional 'widget' user controls
  - Experimental alignment of legacy data sets to thesauri
  - Experimental inter-thesaurus concept matching exercise
- ARIADNE project is making use of some of these LOD thesauri

University of South Wales
Prifysgol De Cymru

5 Star deployment scheme for Linked Open Data

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

[http://www.w3.org/DesignIssues/LinkedData.html]

- This "5 Star" scheme therefore refers to data *format*, not data *quality*
- Also much LOD emphasis to date has been on the *quantity* of data; less focus on the *quality* (e.g. The big LOD diagram)
- Difficult to locate information on exactly how links between data items have been created, how rigorous was the methodology?
- The quality of links may vary – e.g. Automatically derived links vs. manual links, the quality of the underlying data itself may also vary
- We reuse this data, divorced from its original context, as the foundation for new applications. Shouldn't we be more concerned?

Inter-thesaurus matching exercise - data

- Seneschal outcomes included:
  - HE (formerly EH) Monument Types Thesaurus
  - RCAHMS Monument Types Thesaurus
  - RCAHMW Monument Types Thesaurus
- The RCAHMS & RCAHMW thesauri both derived originally from HE Monument Types
  - Ideally shared conceptual knowledge about the domain would not be split along modern political boundaries (and partially duplicated) ...but it is
  - Should be good potential for inter-thesaurus links? Matching exercise undertaken as part of SENESCHAL project

Inter-thesaurus matching exercise - approach

- *Levenshtein* edit distance algorithm
  - Measures optimal number of character edits required to change one string into another
  - Accommodates small spelling differences
- Bulk alignment process
  - Removed bracketed qualifiers from vocabulary terms to give the algorithm a better chance
    - Doesn't penalise a match between e.g. BANK □ BANK (EARTHWORK), but conversely reintroduces homonyms, so a suggested 100% match may be completely wrong…
  - Compared each preferred term from one thesaurus to each term from another thesaurus – obtained best scoring textual matches
  - Similarity threshold introduced to suppress low scoring matches. Edit distance algorithms *always* produce a match, even if it is a bad one!

University of South Wales
Prifysgol De Cymru

Inter-thesaurus matching exercise - results

(small extract, matching on preferred terms only)

RCAHMS Monument Type	Best match HE Monument Type	Score
GALVANIZING WORKS	GALVANIZING WORKSHOP	85%
PENSTOCKS	PENSTOCK	88%
FLAX KILN	FLARE KILN	80%
CUP AND RING MARKED ROCK	CUP AND RING MARKED STONE	84%
GUNCOTTON STORE	GUNCOTTON STOVE	93%
GOOD STATION	GOODS STATION	92%
STAITH	STAITHE	85%
TEXTILE PRINT WORKS	TEXTILE PRINTING WORKS	86%
GRAVE	GRAVE	100%
CIST	CIST	100%
ENCLOSED CREMATION CEMETERY	ENCLOSED CREMATION CEMETERY	100%
HOFFMAN KILN	HOFFMANN KILN	92%
ROAD BLOCK	ROADBLOCK	90%
ANTI AIRCRAFT DEFENCES	ANTI AIRCRAFT DEFENCE SITE	84%
TAKEAWAY	TAKE-AWAY	88%
SETTLING POND	RETTING POND	84%
SUSPENSION FOOTBRIDGE	SUSPENSION BRIDGE	80%
SESSION HOUSE	SESSIONS HOUSE	92%
ALUMINA WORKS	ALUMINIUM WORKS	80%
SHIP BREAKING YARD	SHIP BREAKERS YARD	83%

RCAHMS monuments to HE monuments

RCAHMS Object Type	Best match FISH Object Type	Score
CANDLEHOLDER	CANDLE HOLDER	92%
MANUFACTURING AND PROCESSING	MANUFACTURE AND PROCESSING	89%
CRUSIE	CRUSE	83%
INORGANIC MATERIAL	ORGANIC MATERIAL	88%
PERSONAL ADORNMENT	PERSONAL ORNAMENT	83%
BALANCE	BALANCE	100%

RCAHMS objects to FISH objects

RCAHMS Maritime Craft Type	Best match HE Maritime Type	Score
MOTOR GUN BOAT	MOTOR GUNBOAT	92%
HOUSEBOAT	HOUSE BOAT	90%
CONTAINER SHIP	CONTAINER SHIP	100%
LIBERTY SHIP	LIBERTY SHIP	100%
COLLIER	COLLIER	100%
DUMB HOPPER BARGE	(no match above threshold)	

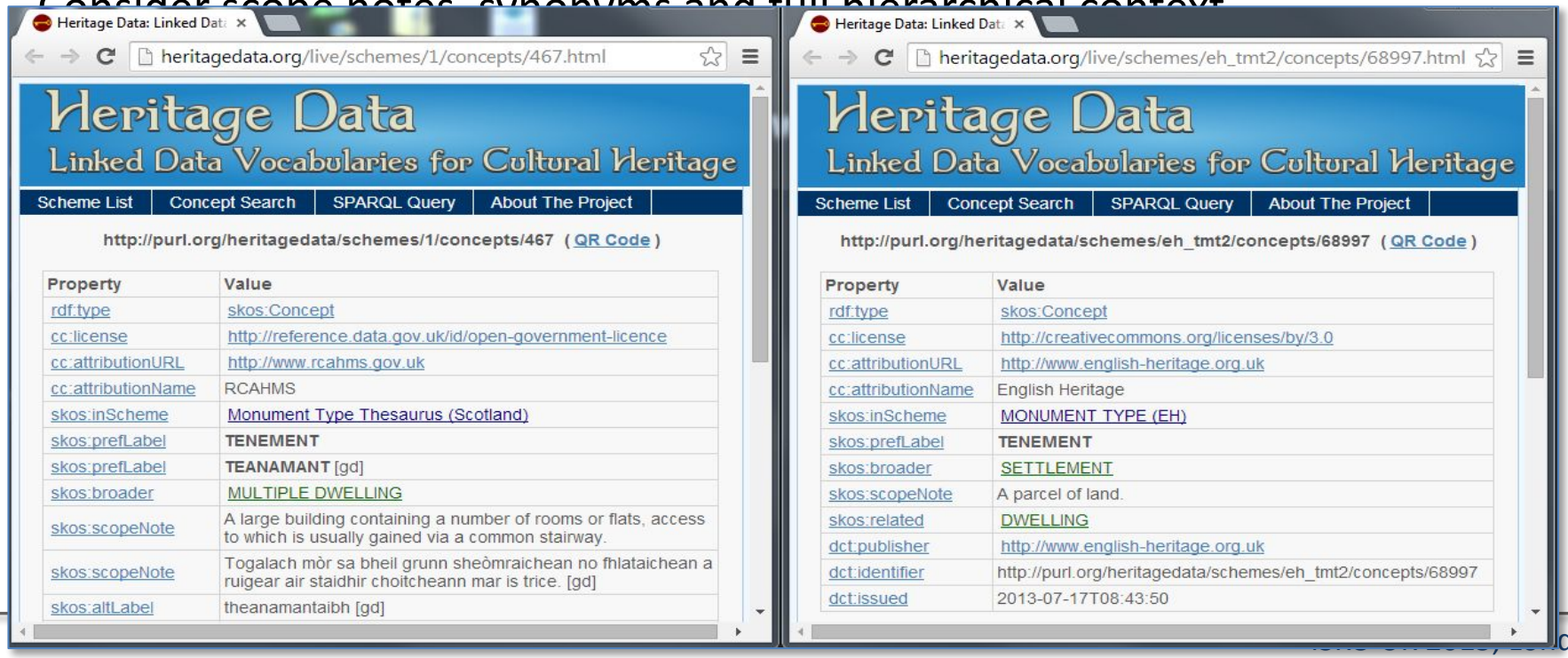
RCAHMS maritime to HE maritime

Technically correct (!) because in using an edit distance measure we **asked for** a *syntactic* match. But what we actually **wanted** was a *semantic* match

University of South Wales
Prifysgol De Cymru

Solution - compare *concepts*, not just *terms*

- ISO 25964-2:2013 notes the need for caution in creating mappings (between thesaurus concepts), stating *"…it is better to have no mapping at all than to establish a misleading one"*
- Automated matching can produce false positives and false negatives. Requires human checking and intervention
- Taking term matches at face value is an inadequate approach - 100% exact match between 2 terms is syntactic NOT semantic; it does NOT mean a concept match
- Consider scope notes, synonyms and full hierarchical context



Comparing concepts

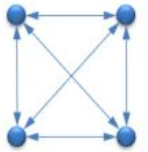
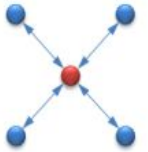
- *Syntactic matching* - may be inexact matching, employing stemming, string matching algorithms (e.g. using the *Levenshtein* edit distance approach as described previously). May need to strip term 'qualifiers', and consider white space, punctuation, capitalisation, case sensitivity etc. Terms may require prior translation in the case of multilingual terminology matching.
- *Scope note evidence* – there may be full or partial (or no) overlap in scope between concepts, realistically this contextual evidence requires human oversight. Scope notes may require translation in the case of multilingual terminology.
- *Synonyms* – groups of alternate synonymous terms may help to reinforce the case for a match between two concepts.
- *Hierarchical context* – ancestors and descendants. If a top-down approach is employed there may be existing mappings higher up in the structure that can give additional contextual evidence to a potential match under consideration.

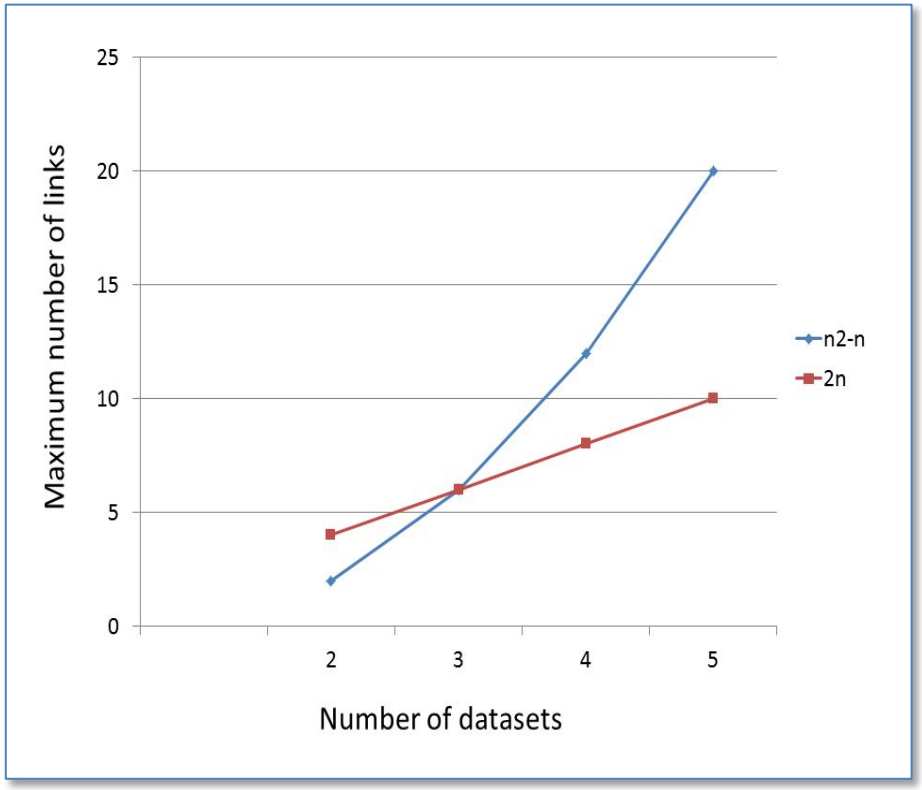
Vocabulary resources for ARIADNE (examples)

- Data Archiving and Networked Services (DANS) -monument types (Archeologische complextypen)
- FASTI Online - monument types
- Istituto Centrale per il Catalogo e la Documentazione (ICCD) - terminology for types of archaeological sites
- Historic England - Monument Types Thesaurus
- Deutsches Archäologisches Institut (DAI) -multilingual archaeological dictionary
- Should be significant conceptual overlap, but without inter-thesaurus mappings, all exist in isolation

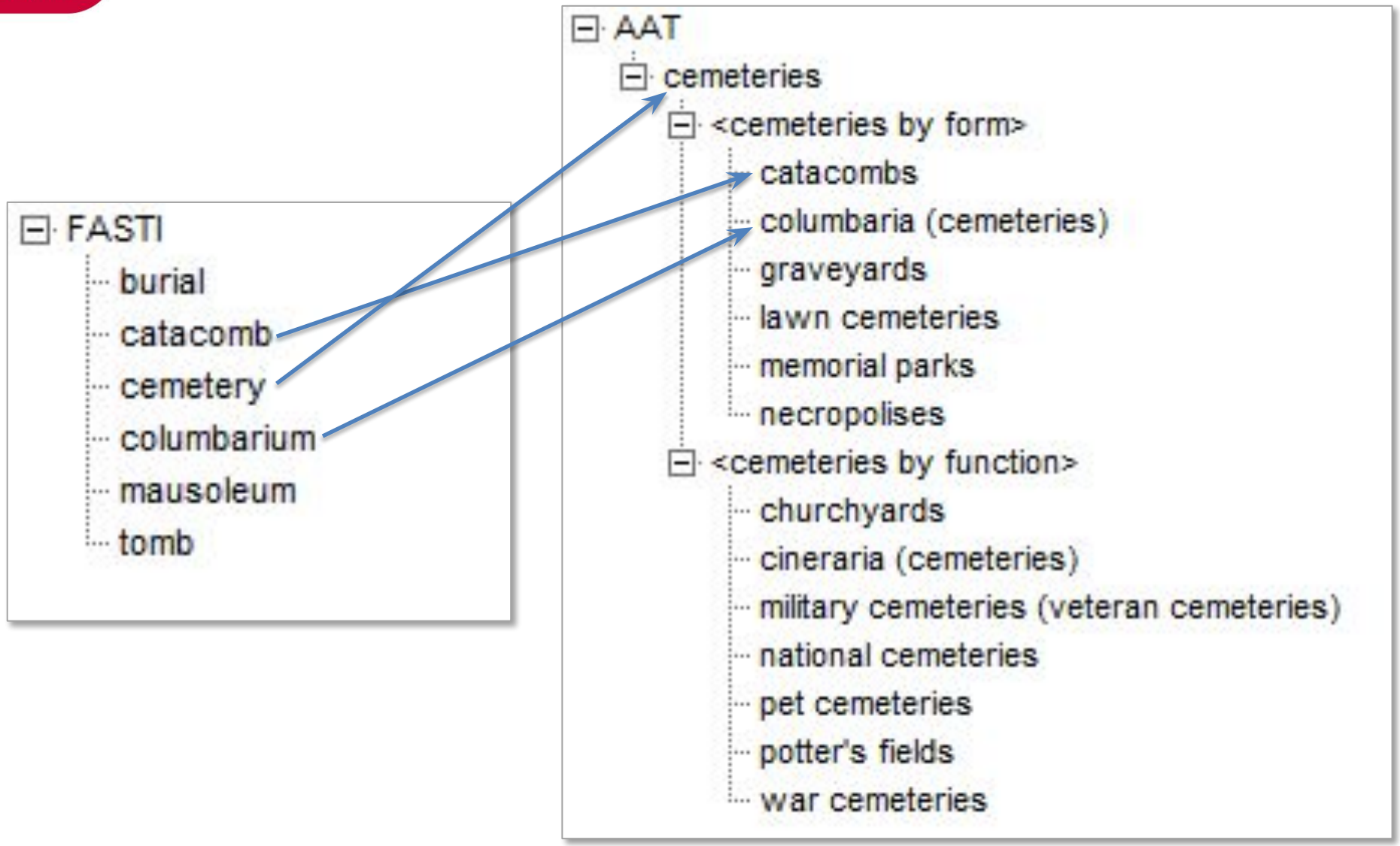
Links: Many-to-many vs. hub architecture

- Potential number of links produced when linking equivalent concepts from multiple thesauri

Datasets	M2M	Links (n^2-n)	HUB	Links (2n)
2		2		4
3		6		6
4		12		8
5		20		10



Mappings from source vocabularies to AAT



Using the created mappings (1)

- Using the SPARQL endpoint at http://vocab.getty.edu/sparql extracted the poly-hierarchical structure of the Getty AAT...

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX xl: <http://www.w3.org/2008/05/skos-xl#>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX aat: <http://vocab.getty.edu/aat/>

CONSTRUCT {?s  gvp:broader ?o; skos:prefLabel ?prefLabel}
WHERE {
   ?s skos:inScheme aat: ;
   (gvp:broaderGeneric | gvp:broaderPartitive) ?o .
   MINUS {?s a gvp:ObsoleteSubject} # don't need these
   MINUS {?o a gvp:ObsoleteSubject} # don't need these
   OPTIONAL { ?s skos:prefLabel ?prefLabel }
   OPTIONAL { ?s xl:prefLabel [xl:literalForm ?prefLabel] }
   FILTER(langMatches(lang(?prefLabel),"EN")) .
}
```

University of South Wales
Prifysgol De Cymru

Using the created mappings (2)

- Imported the extracted AAT structure plus all created mappings to a triple store
- (I used SPARQL GUI; a simple standalone tool for importing RDF and testing SPARQL queries)
  - https://bitbucket.org/dotnetrdf/dotnetrdf/wiki/UserGuide/Tools

```
@prefix fasti: <http://www.fastionline.org/concept/attribute/>
.
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix aat: <http://vocab.getty.edu/aat/> .

fasti:burial skos:closeMatch aat:300387004 .
fasti:catacomb skos:closeMatch aat:300000367 .
fasti:cemetery skos:closeMatch aat:300266755 .
fasti:columbarium skos:closeMatch aat:300000370 .
[etc.]
```

Using the created mappings (3)

- Query the combined data structure and mappings: e.g. find all concepts hierarchically related (via AAT structure) to FASTI “cemetery”

```
# SPARQL 1.1 to locate concepts related (via AAT structure) to FASTI "cemetery"
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX fasti: <http://www.fastionline.org/concept/attribute/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

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

Using the created mappings (4): results

concept	label
iccd:catacomba	catacomba
tmt:91386	catacomb (funerary)
fasti:catacomb	Catacomb
iccd:colombario	colombario
fasti:columbarium	Columbarium
dai:3736	Kolumbarium
dans:6a7482e5-2fd5-48fb-baf4-66ad3d4ed95e	kerkhof
dai:1947	Gräberfeld
iccd:necropoli	necropoli
dai:2485	Nekropole
tmt:70053	cemetery
tmt:70053	necropolis
dans:be95a643-da30-40b9-b509-eadfb00610c4	christelijk/joodse begraafplaats
dans:b935f9a9-7456-4669-91d0-2e9c0ff7d664	vlakgrafveld
tmt:100531	walled cemetery
tmt:92672	mixed cemetery
tmt:70060	inhumation cemetery
tmt:70056	cremation cemetery
tmt:70055	cairn cemetery
tmt:70054	barrow cemetery
iccd:cimitero	cimitero
dans:abb41cf1-30dc-4d55-8c18-d599ebba1bc2	rijengrafveld
fasti:cemetery	Cemetery
dai:1819	Friedhof

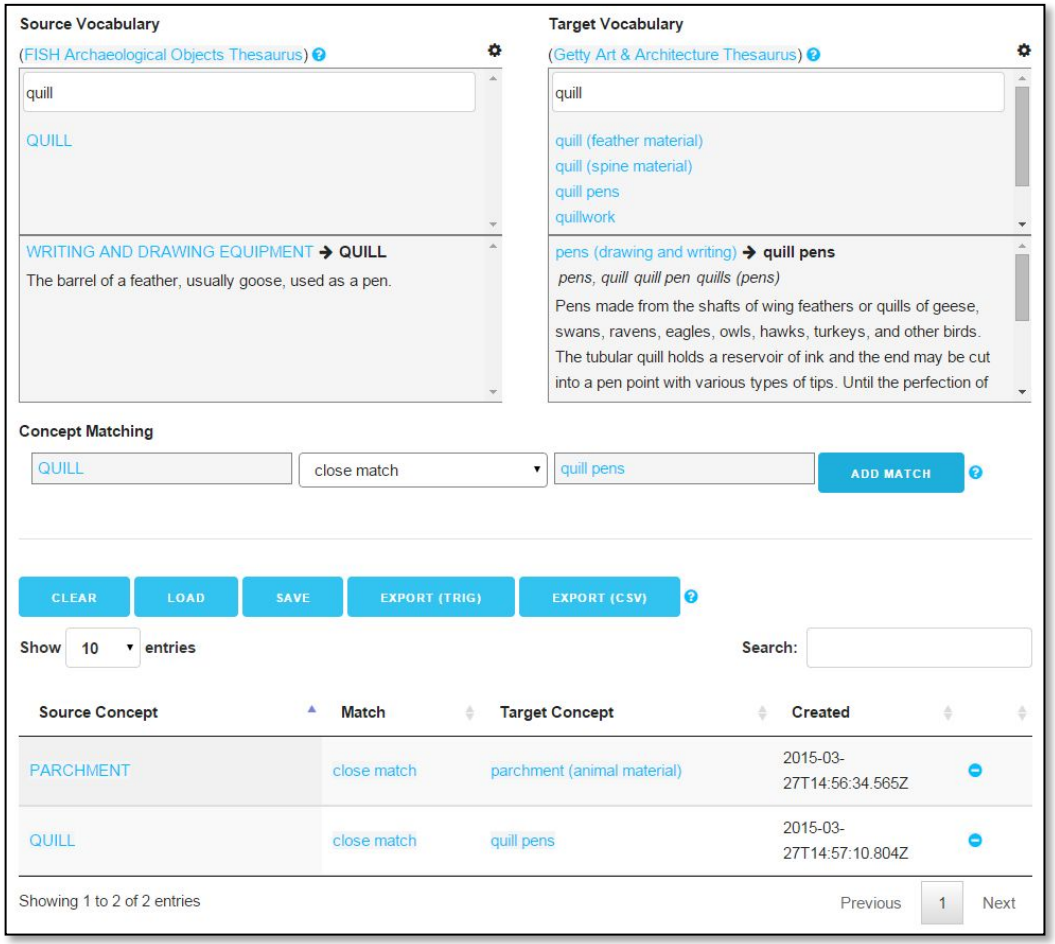
Vocabulary matching tool – requirements

- Creating concept □ concept links, not just term □ term – so utilise more contextual data when matching – labels, scope notes, relationships to other concepts
- Work interactively and allow manual matching. Matching concepts requires human judgement
- Facilitate simple side by side comparison of concepts, with useful accompanying contextual information
- Provide list of possible link types to choose from
- Generate associated metadata, export matches in a suitable serialisation format
- Aim is to facilitate creation of higher quality matches

University of South Wales
Prifysgol De Cymru

Vocabulary matching tool - implementation

See http://heritagedata.org/vocabularyMatchingTool/



Creative Commons zero (CC0) open source code, available from https://github.com/cbinding/VocabularyMatchingTool/

Vocabulary matching tool - features

- Manually matching vocabulary concepts to Getty Art & Architecture Thesaurus (AAT) concepts
- Usage of Linked Open Data resources – accesses external SPARQL endpoints (tool has no server component or database)
- Facilitates side by side comparison of concepts, with contextual details (labels, scope notes, linked concepts)
- Multilingual - French, German, Spanish, English, Dutch AAT concept details (falls back to English if chosen language is not available)
- Export created mappings to JSON, CSV or RDF
- Creative Commons (CC0) open source - see https://github.com/cbinding/VocabularyMatchingTool/

New vocabulary matching exercise

- ADS ArchSearch: 1.4 million records identified for upload to ARIADNE registry
- Subject term metadata extracted. ADS had previously established links from subject terms to HeritageData.org concept URIs
- Matching to Getty AAT concepts undertaken by ADS staff using vocabulary matching tool, for only terms actually used in indexing
- Produced 844 matches to Getty AAT concepts, for source terms originating from 5 thesauri
- Data to be uploaded to ARIADNE registry to assist subsequent search operations

University of South Wales
Prifysgol De Cymru

Conclusions

- Compare concepts, not just terms
- Automated matching requires human review of results. Manual matching is more time consuming, but it only needs to be done once.
- The vocabulary mappings facilitate multilingual cross search over multiple datasets
- The KOS structure supports hierarchical semantic expansion.
- Re-use of existing data, supplemented with new mappings.

University of South Wales
Prifysgol De Cymru

Vocabulary Linked Data Publication and Matching

Ceri Binding, Douglas Tudhope
University of South Wales
ceri.binding@southwales.ac.uk