Mapping between linked data vocabularies in ARIADNE

Ceri Binding & Douglas Tudhope
University of South Wales
douglas.tudhope@southwales.ac.uk

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

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 project, February 2013 → January 2017
  - 24 European partner organisations
  - Multiple languages, multiple controlled vocabularies
  - Thousands of metadata records
- Consolidating metadata does not make it more interoperable – adoption of common schema plus use of controlled vocabularies are the real key to interoperability

5 Star deployment scheme for Linked Open Data

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

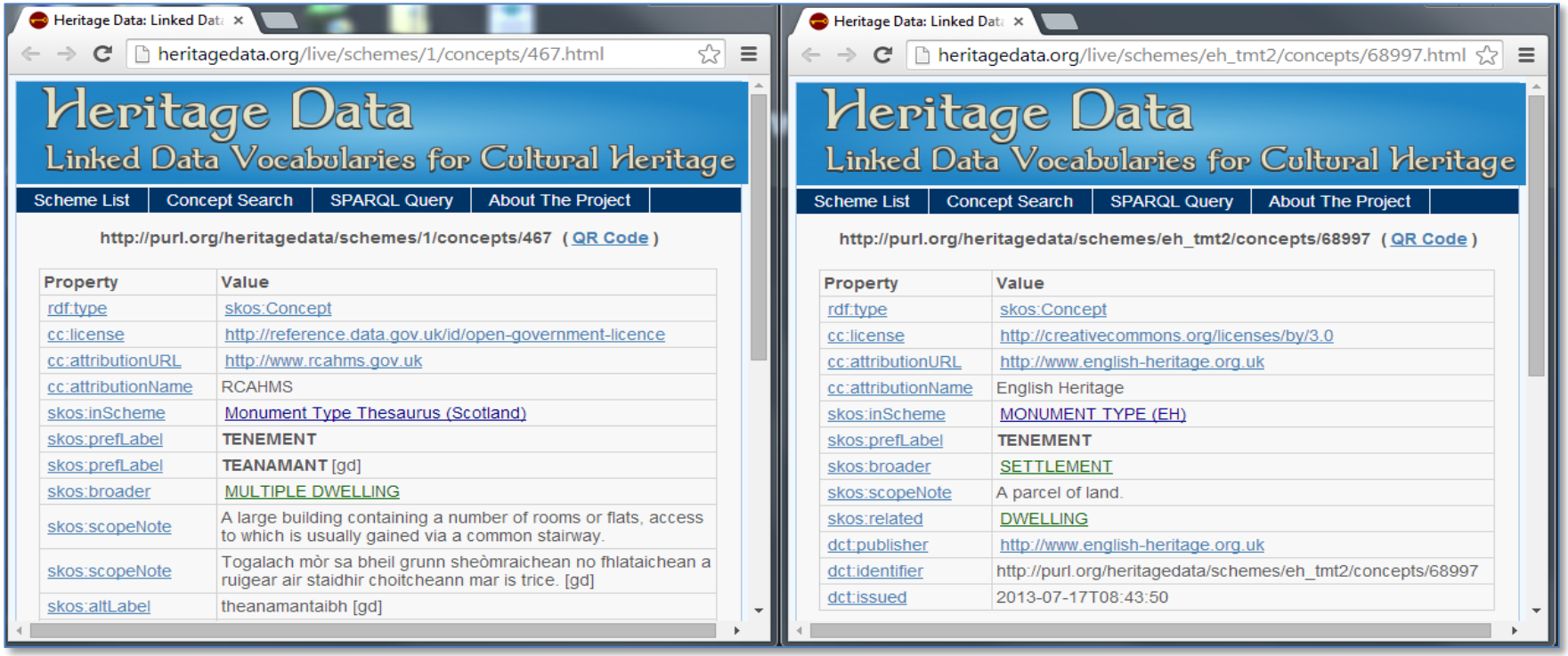
[http://www.w3.org/DesignIssues/LinkedData.html]

- The "5 Star" scheme therefore refers to data *format*, not data *quality*
- Much LOD emphasis to date has been on the *quantity* of data; seems to be less focus on the *quality*
- Difficult to locate information on exactly how links have been created
- The quality of links may vary – e.g. automatic links vs. manual links, the quality of the underlying data itself may also vary
- ISO 25964-2:2013 notes the need for caution in mapping (between thesauri), stating *"…it is better to have no mapping at all than to establish a misleading one"*

We should compare concepts, not just terms

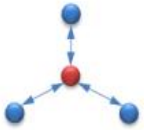
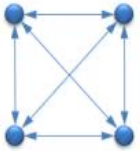
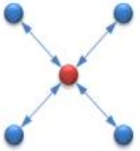
SENESCHAL project (www.heritagedata.org)

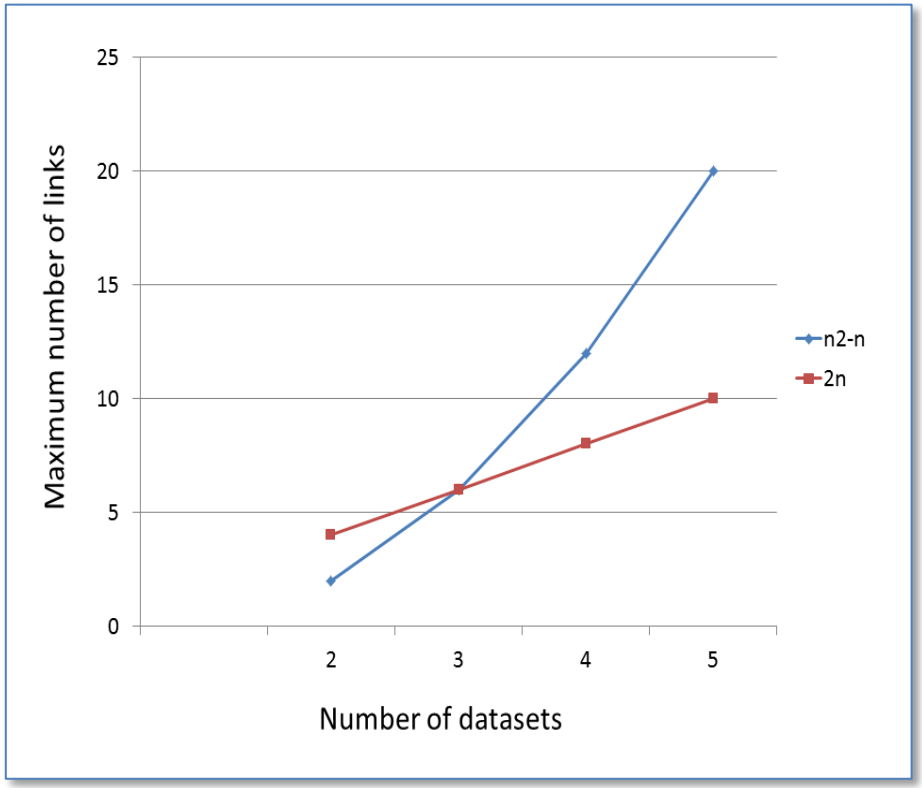
- Automated matching requires human checking and intervention
- Taking term matches at face value is an inadequate approach
- An exact match on a term is syntactic not semantic; does not mean an exact match on a concept
- Need to consider scope notes, synonyms and full hierarchical context



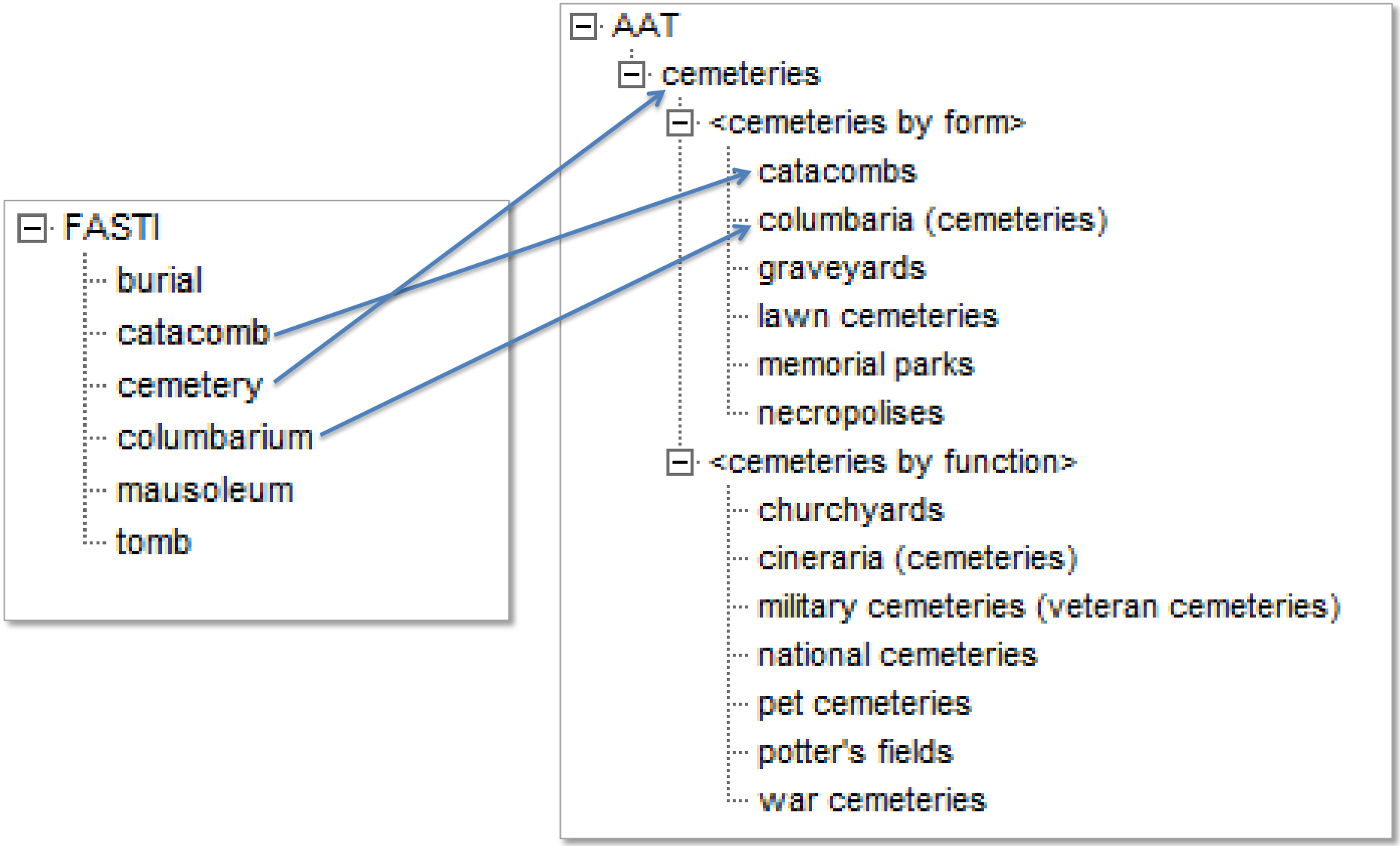
Rationale for a mapping hub (Getty AAT)

- Number of bidirectional links produced when linking equivalent concepts between multiple thesauri

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



Mapping from source vocabulary to AAT



Mapping issues

- Mapping tools (semi-automatic)
- Mapping guidelines for content providers (may be new to mapping work)
  Eg describing context / purpose of mappings
  Eg choosing SKOS mapping relationships
- Mapping metadata
  Eg mapping template

Mapping tools

- Mapping Tool for LD vocabularies
  http://heritagedata.org/vocabularyMatchingTool/
  https://github.com/cbinding/VocabularyMatchingTool
- AAT indexing browser based tool (if wanted at manual import)
  where no Partner subject indexing exists for a dataset
  http://heritagedata.org/vocabularyMatchingTool/indexingtool.html
- Spreadsheet mapping template if vocabulary not in LD
  plus XSL transform to RDF
- Future: multilingual archaeological dictionary as service ?

Mapping Data from partners (ongoing)

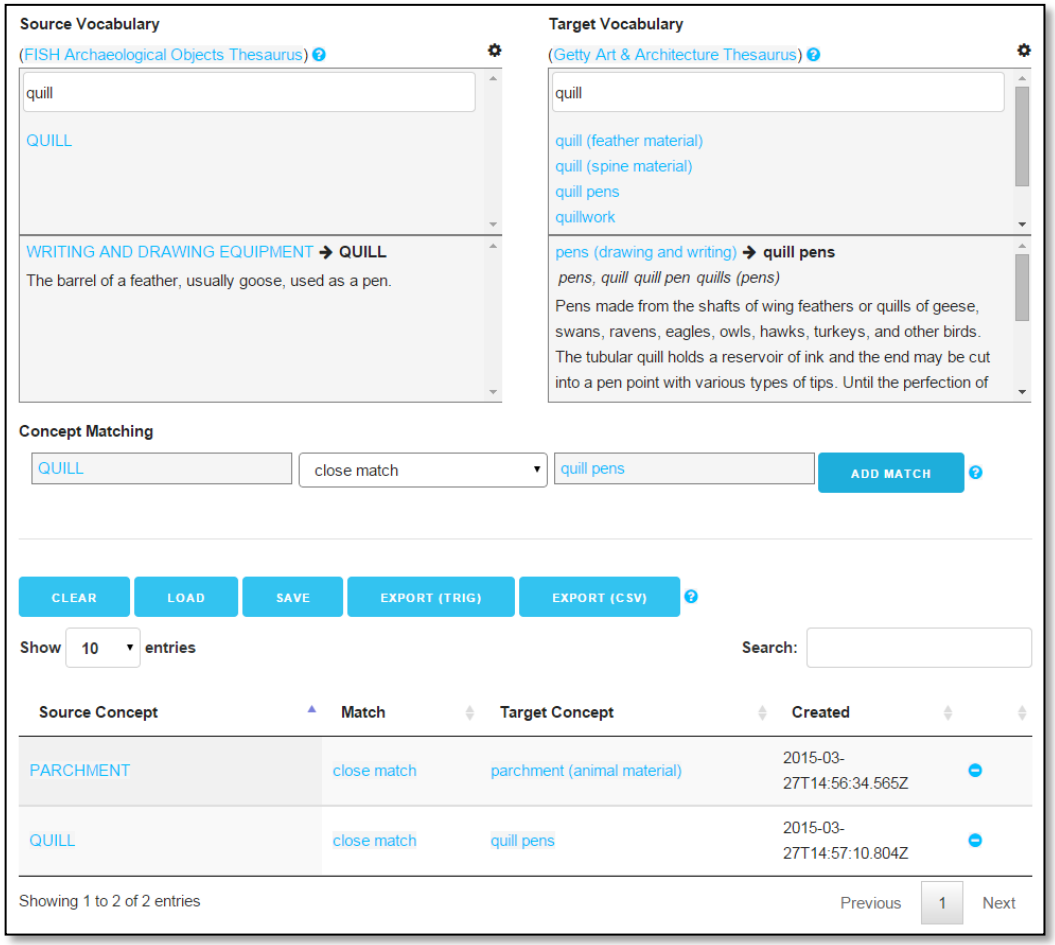
Source	Scheme mapped to AAT		No match	skos:exactMatch	skos:closeMatch	skos:broadMatch	skos:narrowMatch	skos:relatedMatch	Total
ADS	FISH Building Materials Thesaurus (subset)		0	4	8	0	0	0	**12**
ADS	Historic England Components Thesaurus (subset)		0	7	1	1	0	0	**9**
ADS	FISH Archaeological Objects Thesaurus (subset)		0	197	96	118	0	0	**411**
ADS	Historic England Maritime Craft Thesaurus (subset)		0	13	8	3	0	0	**24**
ADS	FISH Thesaurus of Monument Types (subset)		0	139	107	141	0	1	**388**
		Sub total	**0**	**360**	**220**	**263**	**0**	**1**	**844**
			0%	**43%**	**26%**	**31%**	**0%**	**0%**	**100%**
DANS	Archaeologische Artefacttypen		0	0	0	0	0	0	**0**
DANS	Archaeologische Complextypen		25	0	56	19	0	2	**102**
DANS	Archaeologische Perioden		54	0	10	1	0	0	**65**
		Sub total	**79**	**0**	**66**	**20**	**0**	**2**	**167**
			47%	**0%**	**40%**	**12%**	**0%**	**1%**	**100%**
FASTI	FASTI Monument Types		7	23	79	20	0	0	**129**
		Sub total	**7**	**23**	**79**	**20**	**0**	**0**	**129**
			5%	**18%**	**61%**	**16%**	**0%**	**0%**	**100%**
OEAW	UK Material Pool		0	3	0	0	0	0	**3**
OEAW	UK Thunau DB		0	3	1	0	0	0	**4**
OEAW	Franzhausen Kokoern DB		0	5	2	2	1	0	**10**
OEAW	DFMROE DB		0	2	0	0	1	0	**3**
		Sub total	**0**	**13**	**3**	**2**	**2**	**0**	**20**
			0%	**65%**	**15%**	**10%**	**10%**	**0%**	**100%**
SND	Arkeologisk undersökningstyp		9	0	1	0	0	0	**10**
SND	FMIS		41	17	48	48	3	0	**157**
SND	SND Keywords - Archaeology & History		14	36	63	27	0	0	**140**
SND	SND Keywords - Time Periods		22	17	6	20	0	0	**65**
		Sub total	**86**	**70**	**118**	**95**	**3**	**0**	**372**
			23%	**19%**	**32%**	**26%**	**1%**	**0%**	**100%**

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

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 data – Javascript components using external SPARQL endpoints (no back-end server or DB)
- Side by side comparison of concepts, with contextual details (labels, scope notes, linked concepts)
- Multilingual - French, German, Spanish, English, Dutch AAT concept details (fall back to English if chosen language not available)
- Export created mappings to JSON, CSV, RDF
- Creative Commons (CC0) open source (warts and all!). see https://github.com/cbinding/VocabularyMatchingTool/

Data received from partners (ongoing)

		Match type						
Source	Scheme mapped to AAT	No match	skos:exactMatch	skos:closeMatch	skos:broadMatch	skos:narrowMatch	skos:relatedMatch	Total
ADS	FISH Building Materials Thesaurus (subset)	0	4	8	0	0	0	12
ADS	Historic England Components Thesaurus (subset)	0	7	1	1	0	0	9
ADS	FISH Archaeological Objects Thesaurus (subset)	0	197	96	118	0	0	411
ADS	Historic England Maritime Craft Thesaurus (subset)	0	13	8	3	0	0	24
ADS	FISH Thesaurus of Monument Types (subset)	0	139	107	141	0	1	388
	Sub total	**0**	**360**	**220**	**263**	**0**	**1**	**844**
		0%	**43%**	**26%**	**31%**	**0%**	**0%**	**100%**
DANS	Archaeologische Artefacttypen	0	0	0	0	0	0	0
DANS	Archaeologische Complextypen	25	0	56	19	0	2	102
DANS	Archaeologische Perioden	54	0	10	1	0	0	65
	Sub total	**79**	**0**	**66**	**20**	**0**	**2**	**167**
		47%	**0%**	**40%**	**12%**	**0%**	**1%**	**100%**
FASTI	FASTI Monument Types	7	23	79	20	0	0	129
	Sub total	**7**	**23**	**79**	**20**	**0**	**0**	**129**
		5%	**18%**	**61%**	**16%**	**0%**	**0%**	**100%**
OEAW	UK Material Pool	0	3	0	0	0	0	3
OEAW	UK Thunau DB	0	3	1	0	0	0	4
OEAW	Franzhausen Kokoern DB	0	5	2	2	1	0	10
OEAW	DFMROE DB	0	2	0	0	1	0	3
	Sub total	**0**	**13**	**3**	**2**	**2**	**0**	**20**
		0%	**65%**	**15%**	**10%**	**10%**	**0%**	**100%**

Data received from partners (ongoing)

	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 eller
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

Spreadsheets containing local vocabulary → AAT mappings

Transformation of vocabulary mappings

Spreadsheet data saved to tab-delimited text:

stenkammargrav	skos:broadMatch	chamber tombs	http://vocab.getty.edu/aat/300005935
stenkistgrav	skos:exactMatch	cist graves	http://vocab.getty.edu/aat/300005941
stenkrets	skos:broadMatch	burial sites	http://vocab.getty.edu/aat/300387004

XSL Transformation

RDF (NTriples):

<http://tempuri/SND/stenkammargrav> <http://www.w3.org/2004/02/skos/core#broadMatch> <http://vocab.getty.edu/aat/300005935> .
<http://tempuri/SND/stenkistgrav> <http://www.w3.org/2004/02/skos/core#exactMatch> <http://vocab.getty.edu/aat/300005941> .
<http://tempuri/SND/stenkrets> <http://www.w3.org/2004/02/skos/core#broadMatch> <http://vocab.getty.edu/aat/300387004> .

<http://tempuri/SND/stenkammargrav> <http://www.w3.org/2004/02/skos/core#prefLabel> "stenkammargrav"@sv .
<http://tempuri/SND/stenkistgrav> <http://www.w3.org/2004/02/skos/core#prefLabel> "stenkistgrav"@sv .
<http://tempuri/SND/stenkrets> <http://www.w3.org/2004/02/skos/core#prefLabel> "stenkrets"@sv .

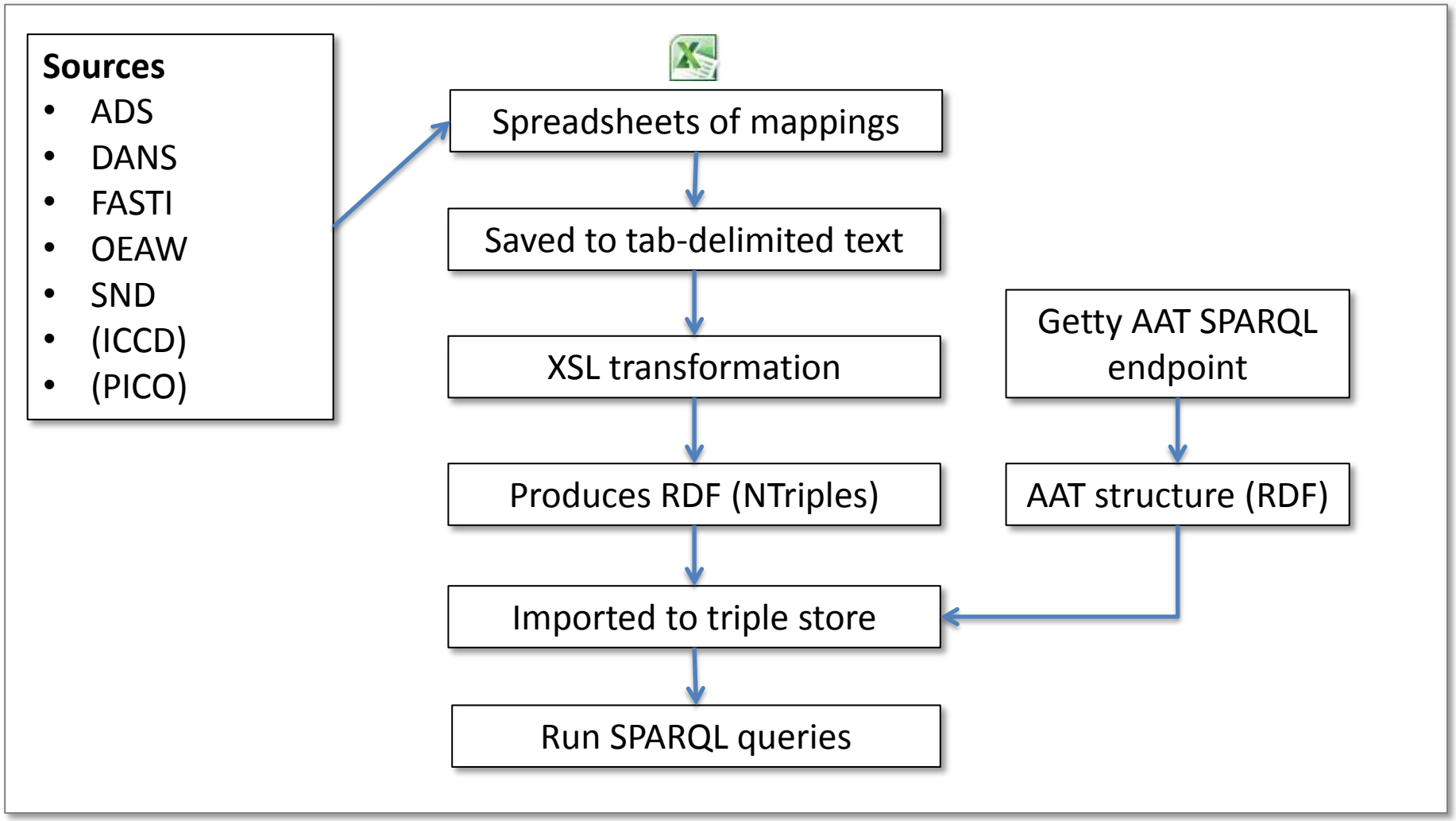
Obtaining the Getty AAT structure

Using the SPARQL endpoint at http://vocab.getty.edu/sparql extract 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")) .
}
```

Converting the vocabulary mappings

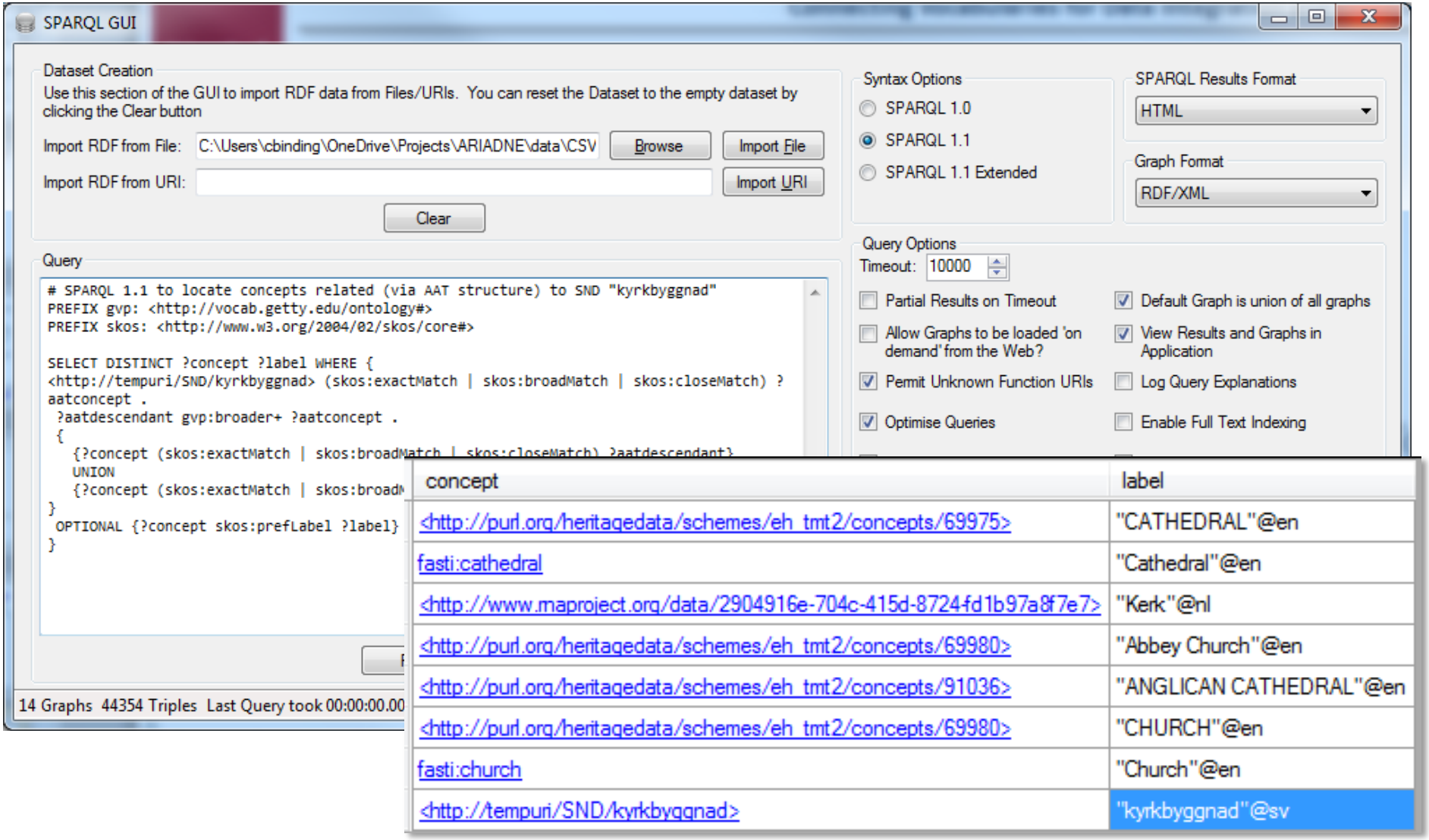


Consolidating the mappings

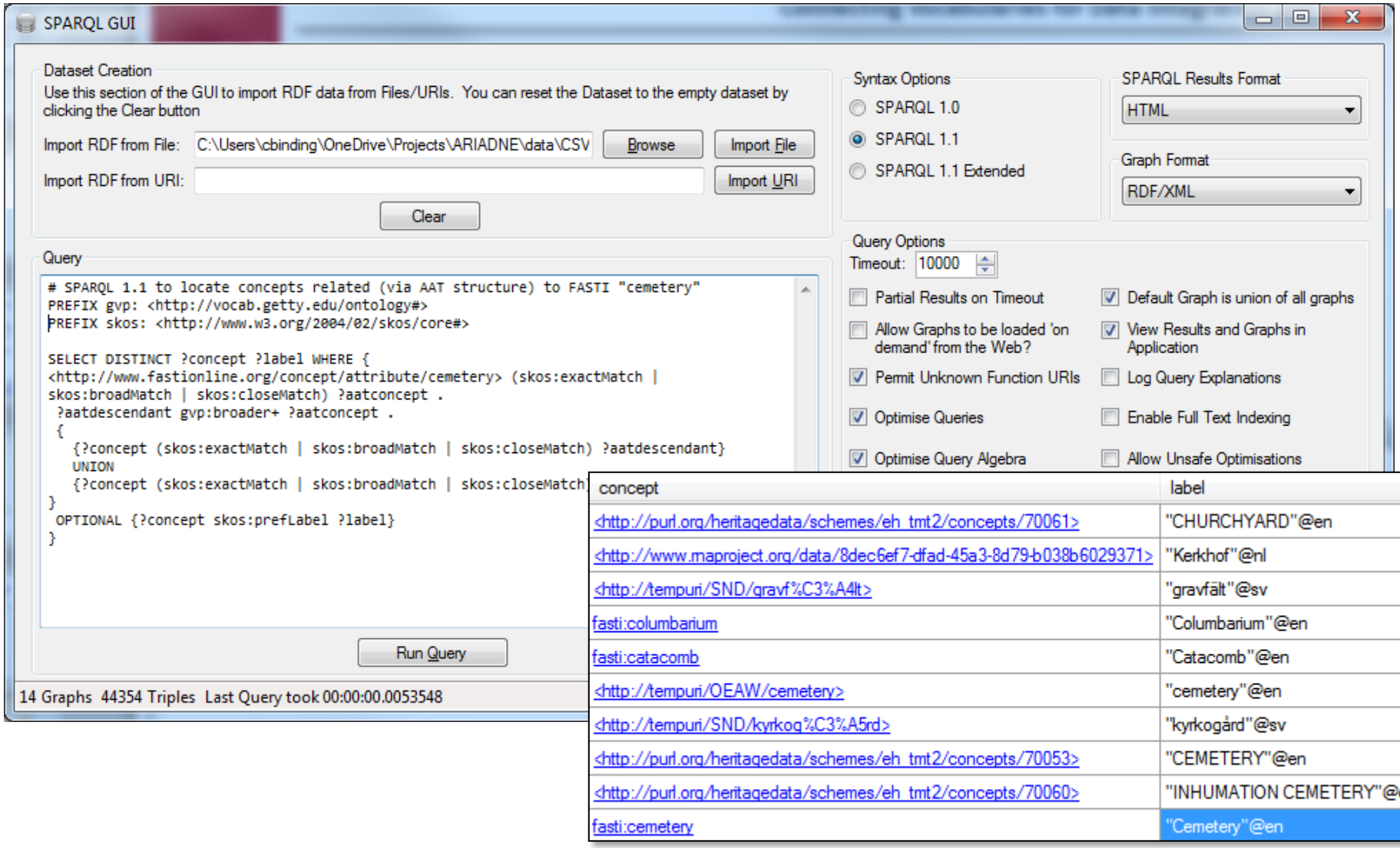
- Import the extracted AAT structure to a triple store
- (For the examples we used SPARQL GUI; a simple standalone tool for importing RDF and testing of SPARQL queries)
  - https://bitbucket.org/dotnetrdf/dotnetrdf/wiki/User Guide/Tools
- Import all the converted mappings to the triple store

```
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.]
```

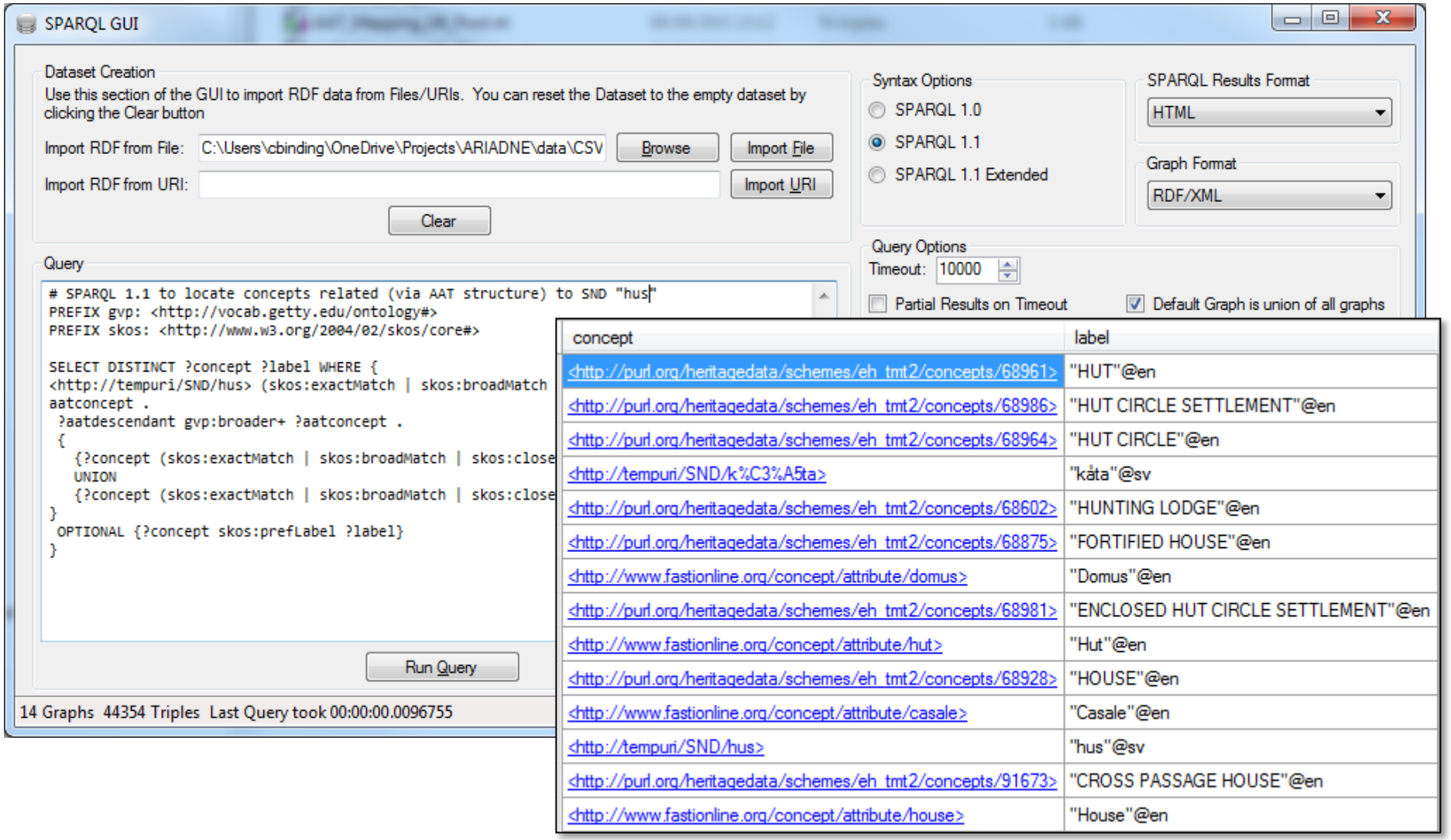
Utilizing the vocabulary mappings (1)



Utilizing the vocabulary mappings (2)



Utilizing the vocabulary mappings (3)



Conclusions

- Compare concept not just terms
- The vocabulary mappings facilitate multilingual cross search over multiple datasets
- Integration of semantic structure can improve recall AND precision of search
- The spine structure supports hierarchical semantic expansion
- Supports semantic browsing (more like this)
- Can be used in addition to free text searching
- Quality mappings require ‘expert’ review of results. Manual involvement is more time consuming, but can be supported by semi-automated tools. Only needs to be done once and can support various applications.

Mapping issues

- Mapping tools (semi-automatic)

- Mapping guidelines for content providers (may be new to mapping work)
  Eg describing context / purpose of mappings
  Eg choosing SKOS mapping relationships
  -- implications for retrieval?

- Mapping metadata
  Eg mapping template
  Typology of mapping methods?

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.