Vocabularies as Linked Data:
SENESCHAL and HeritageData.org

University of Leicester and Historic England
Heritage Practice Training Course
12th April 2016

Keith May
@Keith_May
Historic England

Incorporating work by
Prof Doug Tudhope & Ceri Binding
University of South Wales

AHRC funded STAR, STELLAR and SENESCHAL Projects
http://hypermedia.research.southwales.ac.uk/kos/star/
http://hypermedia.research.southwales.ac.uk/kos/stellar/
http://hypermedia.research.southwales.ac.uk/kos/SENESCHAL/

Arts & Humanities Research Council | ads ARCHAEOLOGY DATA SERVICE | Historic England | University of South Wales Prifysgol De Cymru

Outline of Content

1. Overview of relevant Linked Data technologies
2. SENESCHAL project & Linked Data
3. LOD Vocabulary developments
4. HeritageData.org - Forum for Info Standards in Heritage (FISH)

Questions and Discussion - All

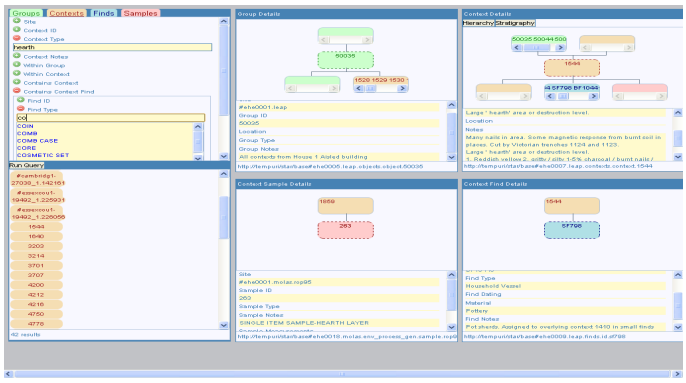
Linking - The Archaeological Archipelagos

Linked Data?
What's in it for Us & What do we need this for?

- Better shared understanding of *existing* information
- Enabling more complex and accurate Semantic Web searching by both Archaeologists & non-domain experts
- Wider Access and re-use of info by interested Public, Community Groups, Students, Researchers, *et al*
- Relating archaeology to other domains
  - E.g. Natural sciences, Biology, Anthropology, Environmental studies
- SKOS and W3C web standards enable standardisation & interoperability with other Linked Data online

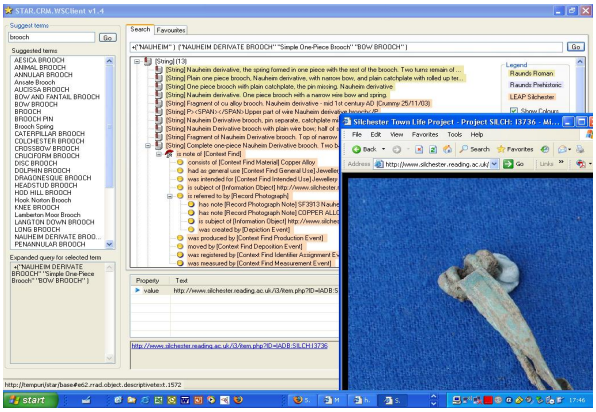


Background to *Vocabulary issues* that emerged in STAR project interface for cross-search of integrated data



Internet Archaeology Vol 30 (2011) http://intarch.ac.uk/journal/issue30/tudhope_index.html

Prototype Controlled Vocabulary searching



The SENESCHAL Project - Overview

- **seneschal** n. *Historical*
  "The steward or major-domo of a medieval great house"
- 12 month AHRC funded project: March 2013 → February 2014
- University of South Wales (formerly Glamorgan) and ADS with Project Partners including, RCAHMS, RCAHMW, EH/HE
- Knowledge Exchange based on enhanced vocabulary services
- Make it significantly easier for data providers to index their data with uniquely identified (machine readable) controlled terminology – ie semantically enriched and compatible with Linked Data.
- Make it easier for vocabulary providers to make their vocabularies available as Linked Data. HE Thesauri and RCAHMS/W thesauri as exemplar cases.

The SENESCHAL Project – Deliverables

- Controlled vocabularies online
  - Vocabularies from HE, RCAHMS, RCAHMW
  - Conversion to a common standard format (SKOS)
  - Persistent globally unique identifiers for every concept
  - Made available online as Linked Open Data
  - Also downloadable data files and listings
- Web services
  - Facilitate concept searching, browsing, suggestion, validation
- Tools to use controlled vocabularies
  - Browser-based 'widget' user interface controls
  - Search, browse, suggest, select concepts
- Case studies
  - Legacy data to thesaurus alignment
  - Thesaurus to thesaurus alignment
  - Third party use of project outcomes

Problem: Semi-controlled vocabularies…

Deposit Colour			Deposit Texture	Deposit Compaction
(Reddy) Brown	Dark orange/brown	Orangy brown, very light	Firm	Plastic
9Reddy) brow				cky
Brown				y (wet)
Brown red				y/firm
Brown/redd				s
Dark brown				
Dark brown				
Dark grey b				
Dark orange				
Dark orange				
darker patch				
Dark orange loam		Yellow brown		
		Yellow/orange brown		

"…another of my examples has something about some flint that is 'snuff coloured' & I don't know if I've ever seen snuff, let alone know what colour it is, or might have been over 150 years ago, and I would think it would make sense to take some kind of integrated approach from the outset,…." [G. Carver]

For data entry: Semi-controlled vocabularies represent a useful compromise somewhere between descriptive & controlled vocabularies, ***the best of both worlds!***

For data retrieval: ***The worst of all worlds*** *(Re. find all the iron age post holes)*

This problem arises from trying to do two different things within a single input field. Should do both, but separately ***– 1) describe using free text description fields,*** *and* ***2) index using controlled index fields***

Try using CONTROLLED Vocabularies online

Vocabularies from Historic England
- Archaeological Sciences
- Building Materials
- Components
- Event Type
- Evidence
- FISH Archaeological Objects
- Maritime Craft Type
- Monument Type
- Periods

Vocabularies from RCAHMS
- •Archaeological Objects Thesaurus (Adapted version of the FISH Archaeological Objects Thesaurus)
- •Maritime Craft Thesaurus
- •Monument Type Thesaurus (Multilingual - includes Scottish Gaelic translations)

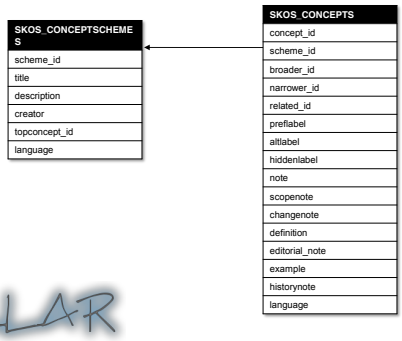
Vocabularies from RCAHMW
- •Monument Type Thesaurus
- •Period

Moving from term based towards concept based indexing
- •Start to create links between concepts… between vocabularies… between datasets… between sites… between countries
- •Alignment from legacy data to persistent concept identifiers
- •Alignment between thesauri
- •True interoperability of (multilingual) cultural heritage resources

STELLAR Project Tools - SKOS Template

SKOS = Simple Knowledge Organisation System

Using SKOS - W3C standard for Web-based Terminologies



STELLAR

RDF – Resource Description Framework

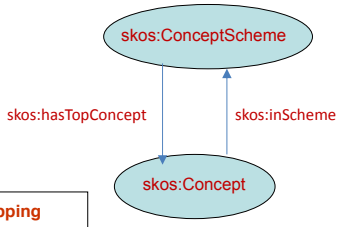
- Data exported to an RDF Triple Store (big database)
- RDF triples in the form of:
- Subject – Predicate – Object

- Entity – Relationship – Entity
- Class – Property – Class
- SKOS is W3C standard format for data representation & Exchange
- The boxes in the diagram show each Entity that is joined to another Entity by a Relationship i.e. forms a Triple

SKOS Concepts v Term Hierarchies

Semantic
- broader
- narrower
- related

Documentation
- note
- changeNote
- Definition
- editorialNote
- Example
- historyNote
- scopeNote



Mapping
- broadMatch
- narrowMatch
- relatedMatch
- closeMatch
- exactMatch

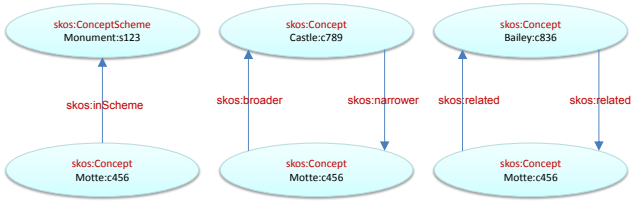
Labelling
- prefLabel
- altLabel
- hiddenLabel

SKOS_CONCEPTS – scheme_id, broader_id, related_id

concept_id	scheme_id
c456	s123

concept_id	broader_id
c456	c789

concept_id	related_id
c456	c836




```
<skos:Concept rdf:about="http://stellar/c456"/>
<rdf:Description rdf:about="http://stellar/c456">
  <skos:inScheme rdf:resource="http://stellar/s123"/>
</rdf:Description>
```

```
<skos:Concept rdf:about="http://stellar/c456"/>
<rdf:Description rdf:about="http://stellar/c456">
  <skos:broader rdf:resource="http://stellar/c789"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/c789">
  <skos:narrower rdf:resource="http://stellar/c456"/>
</rdf:Description>
```

```
<skos:Concept rdf:about="http://stellar/c456"/>
<rdf:Description rdf:about="http://stellar/c456">
  <skos:related rdf:resource="http://stellar/c789"/>
</rdf:Description>
<rdf:Description rdf:about="http://stellar/c789">
  <skos:related rdf:resource="http://stellar/c456"/>
</rdf:Description>
```

Concepts: Accommodating colloquial terms

Dr. Johnson: (proudly) "*Here it is sir, the very cornerstone of English scholarship. This book contains every word in our beloved language.*"

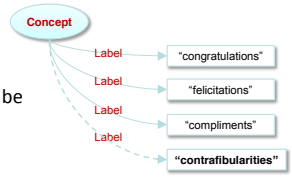
Blackadder: "*every single one sir? [..] In that case I hope you will not object if I also offer my most enthusiastic ... **contrafibularities**.*"

Dr. Johnson: "*What?*"

Blackadder: "***contrafibularities** sir – it is a common word down our way.*"

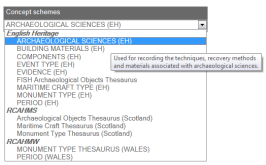
Dr. Johnson: (flustered and scribbling) "*Damn…*"

Blackadder's mischievous suggestion may be a new *term,* but it is not a new *concept*.
It fits into the existing concept structure, further enriching the entry vocabulary.



Thanks to Ceri Binding for this slide – and others

Voacabulary Widgets – e.g. for OASIS

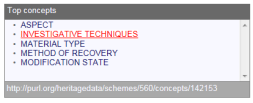


(scheme list)

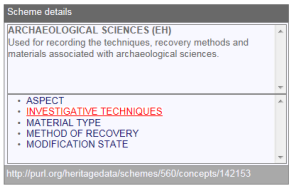
- Scheme list
- Scheme details
- Top concepts
- Composite control

More Widget details on HeritageData.org

(scheme details)



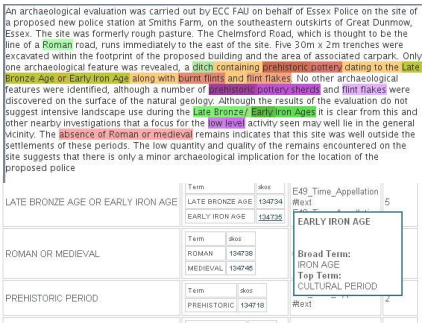
(top concepts)



(composite control)

STAR - Semantic Technologies for Archaeological Resources

Natural Language Processing (NLP)



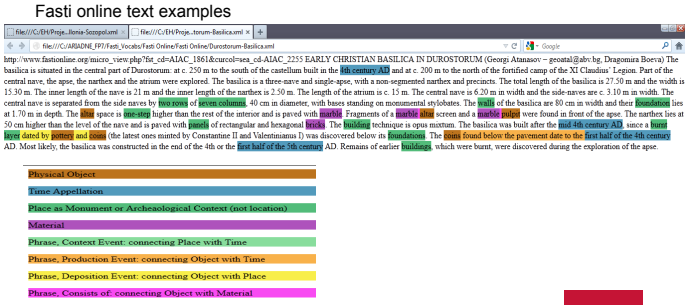
- **NLP Information Extraction (IE) of Concepts from OASIS GL Reports such as**
- **Place**
- **Period**
- **Object**
- **Utilise semantic annotation XML files**
- **Using SKOS RDF versions of thesauri concepts.**

With thanks to Andreas Vlachidis

University of South Wales
Prifysgol De Cymru

ARIADNE FP7 project R&D work
NLP using SKOS vocabularies

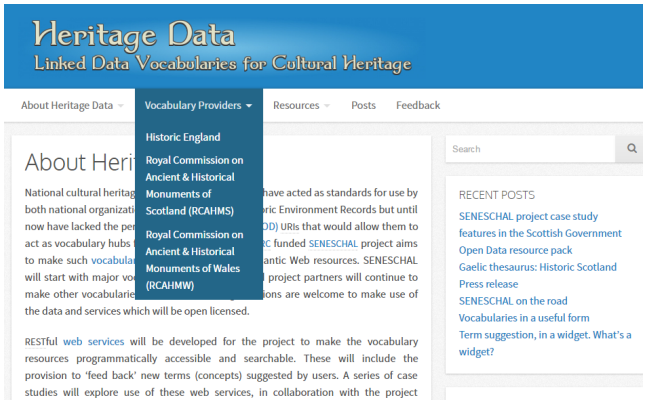
Fasti online text examples



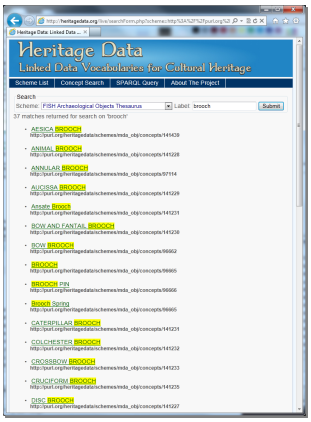
With thanks to Andreas Vlachidis

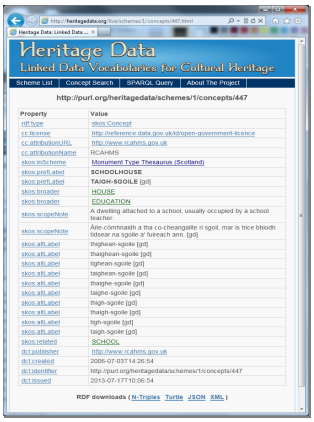
University of South Wales
Prifysgol De Cymru

LOD Heritage Vocabularies: http://heritagedata.org



Thesaurus searching and browsing





Typical alignment problems encountered

- Simple spelling errors
  - POSTHLOLE", "CESS PITT", "FURRROWS", FLINT SCRAPPER"
- Alternate word forms
  - "BOUNDARY"/"BOUNDARIES", "GULLEY"/"GULLIES"
- Prefixes / suffixes
  - "RED HILL (POSSIBLE)", "TRACKWAY (COBBLED)", "CROFT?", "CAIRN (POSSIBLE)", "PORTAL DOLMEN (RE-ERECTED)"
- Nested delimiters
  - "POTTERY, CERAMIC TILE, IRON OBJECTS, GLASS"
- Terms not intended for indexing
  - "NONE", "UNIDENTIFIED OBJECT", "N/A", "NA", "INCOHERENT"
- Terms that would not be in (any) thesauri
  - "WOTSITS PACKET", "CHARLES 2ND COIN", "ROMAN STRUCTURE POSSIBLY A VILLA", "ST GUTHLACS BENEDICTINE PRIORY", "WORCESTER-BIRMINGHAM CANAL", "KUNGLIGA SLOTTET", "SUB-FOSSIL BEETLES"
- More specific phrases
  - "SIDE WALL OF POT WITH LUG", "BRICK-LINED INDUSTRIAL WELL OR MINE SHAFT", "ALIGNMENT OF PLATFORMS AND STONES"

Data alignment - R&D approach

- *Levenshtein* edit distance algorithm
  - Measures optimal number of character edits required to change one string into another
  - Accommodates small spelling differences/errors
- Bulk alignment process
  - Compares each value to all terms from specified thesaurus – obtain best textual match
  - Similarity threshold introduced to suppress low scoring matches. Levenshtein algorithm will always produce a match, even if it is a bad one!
  - Periods require an additional approach due to mixed formats (named periods, numeric ranges etc.)

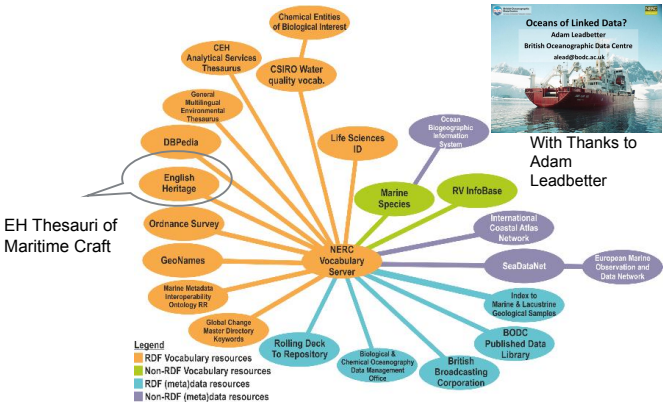
Data Alignment Results – Monument Types

Data value	Highest scoring match	Score
ABBEY FOUNDATIONS	Foundation	74%
AXE FACOTRY	Axe Factory	90%
BOUNDARIES	BOUNDARY	77%
BOUNDARY	BOUNDARY	100%
BUIED SOIL HORIZON	BURIED SOIL HORIZON	97%
CAIRN	CAIRN	100%
CAIRN (POSSIBLE)	CAIRN	100%
CAIRNN	CAIRN	90%
CESS PITT	CESS PIT	94%
CHAMBERED TOM	CHAMBERED TOMB	96%
COMERCIAL	COMMERCIAL	94%
CROFT?	CROFT	90%
CUP-MARKED STONE	CUP MARKED STONE	93%
DICTH	DITCH	80%
ENCLSOURE	ENCLOSURE	88%
EXTRACTION PIT	EXTRACTIVE PIT	85%
EXTRACTIVE PIT	EXTRACTIVE PIT	100%

Data value	Highest scoring match	Score
FEATURE – COBBLED SURFACE	Cobbled Surface	75%
GULLEY	GULLY	90%
GULLIES	GULLY	66%
HILL FORT	HILLFORT	94%
HILLFORT	HILLFORT	100%
IINEAR SYSTEM	LINEAR SYSTEM	92%
MEDIEVAL CASTLE / FORTIFIED MANOR RUINS	FORTIFIED MANOR HOUSE	60%
PARIS CHURCH	PARISH CHURCH	96%
PASSAGE GRACE	PASSAGE GRAVE	92%
PORTAL DOLMEN (RE-ERECTED)	PORTAL DOLMEN	100%
POSTHLOLE	POST HOLE	88%
PRIORY? WALL	Priory Wall	95%
RED HILL (POSSIBLE)	RED HILL	100%
ROMAN STRUCTURE POSSIBLY A VILLA	TRAINING STRUCTURE	52%
SOIL FILLED PIT	RIFLE PIT	66%
ST GUTHLACS BENEDICTINE PRIORY	Benedictine Priory	75%
STONE ALIGMENT	STONE ALIGNMENT	96%
TRACKWAY (COBBLED)	TRACKWAY	100%
WORCESTER-BIRMINGHAM CANAL	ORNAMENTAL CANAL	52%

Opportunities

E.g. British Oceanographic Data Centre - LOD

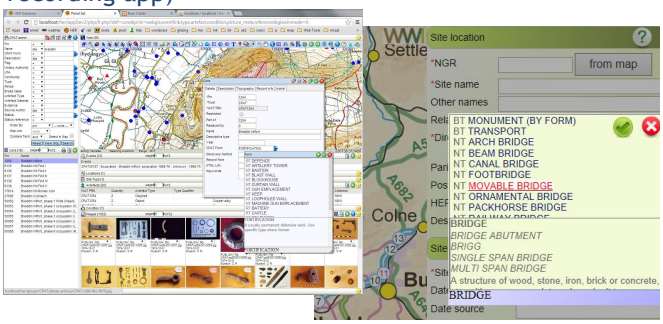


With Thanks to Adam Leadbetter

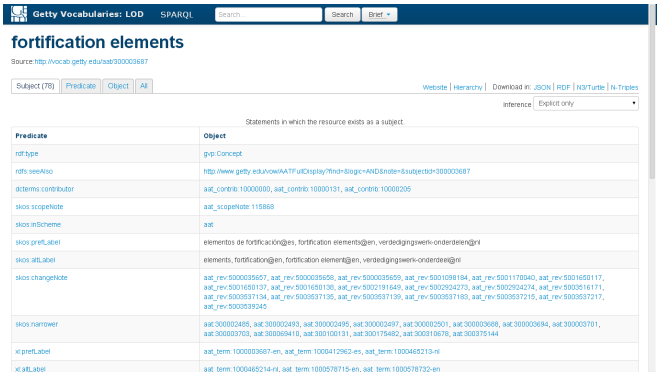
EH Thesauri of Maritime Craft

Opportunities

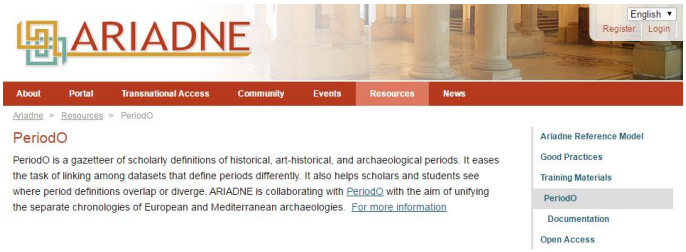
Clwyd-Powys (Wales) Archaeological Trust (SENESCHAL widgets embedded into HER application and mobile field recording app)



Vocab Cross-ref Opportunities!
Getty A&AT Vocab as LOD



ARIADNE project using Getty A&AT LOD as “Hub”

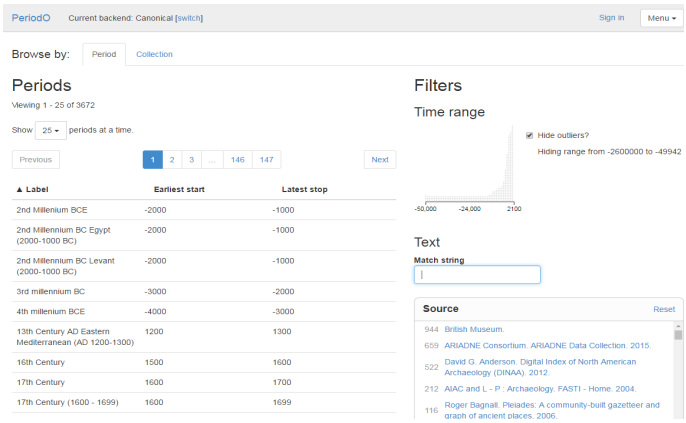


http://www.ariadne-infrastructure.eu/

PeriodO

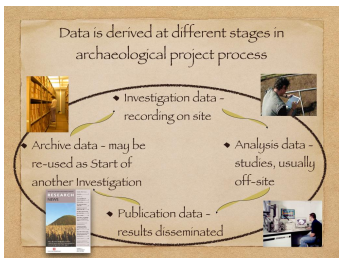
A gazetteer of period definitions for linking and visualizing data.

http://perio.do/



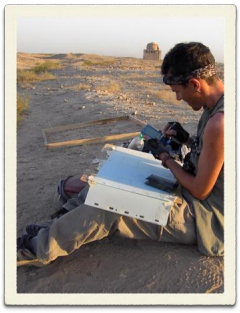
Stages for making Data Open

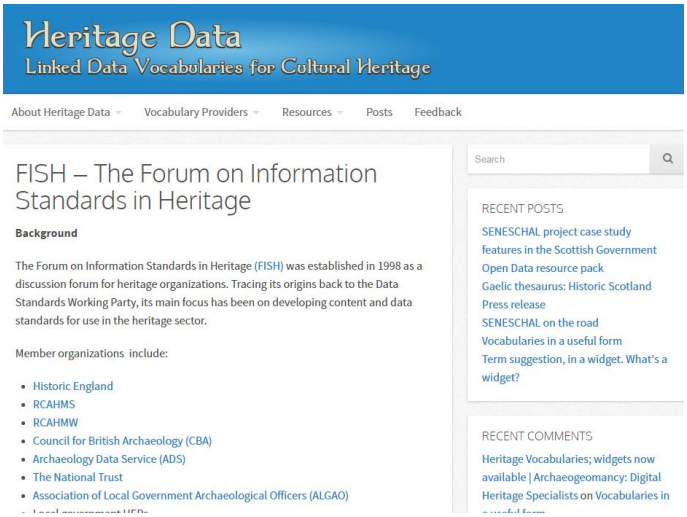
- LOD may blur existing boundaries as (Big) data integration becomes more dynamic
- STAR outcomes suggest still 4 key stages for coherent data integration in the Archaeological Research Cycle.
- Excavation archive stage
- Results of Analysis
- "Final" Publication
- Integrated Archive for new Research



Open Archaeological Data somewhere on/over the horizon?

- Different archaeological recording systems share common conceptual frameworks and semantic relationships
- By conceptualising common relationships in our different data sets at a broad level and ***aligning*** vocabularies of shared reference terms we can cross-search data for patterns and broader answers to related research questions
- The technologies are being developed in other domains (e.g. Biology) but is there a common will for ***sharing*** archaeological data ***Openly*** for ***re-use*** in the interests of improving research methods?





Heritage Data
Linked Data Vocabularies for Cultural Heritage

http://www.heritagedata.org/

Ceri Binding, Doug Tudhope, Andreas Vlachidis
University of South Wales
ceri.binding@southwales.ac.uk
douglas.tudhope@southwales.ac.uk
andreas.vlachidis@southwales.ac.uk

Keith May
Historic England & University of South Wales

Keith.May@historicengland.org.uk

Arts & Humanities Research Council
Historic England
Royal Commission on the Ancient and Historical Monuments of Scotland
COMISIWN BRENHINOL HENEBION CYMRU ROYAL COMMISSION ON THE ANCIENT AND HISTORICAL MONUMENTS OF WALES
ADS ARCHAEOLOGY DATA SERVICE