

5 Star Data



Achieving the 5th Star

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

Outcomes of the SENESCHAL project

- 12 month AHRC funded project, in collaboration with English Heritage, RCAHMS, RCAHMW, ADS
- Project aims:
 - Widening access to key vocabulary resources
 - Facilitating improved consistency for existing and future metadata
- Project outcomes:
 - <http://www.heritagedata.org>
 - 14 vocabularies converted to SKOS format, made available online as Linked Open Data
 - Associated vocabulary web services and functional ‘widget’ user controls
 - Experimental alignment of legacy data sets to thesauri
 - Experimental inter-thesaurus concept alignment

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 out to other external data**

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

- The “5 Star” scheme refers to data format, not data quality
- SENESCHAL project achieved mainly “4 Star” data
- Maybe 4.1 Star? *Some* external links are present e.g. skos:broadMatch links (898) between RCAHMW Monuments concept URIs and DDC concept URIs
- Currently no inter-thesaurus concept links are exposed
- Would inter-thesaurus links count as “external”?

Quantity of links vs. quality

- Much LOD emphasis on the *quantity* of data; less focus on the *quality*. e.g. LOD cloud diagram:
 - Many presentations illustrate how the cloud has ‘grown’ over time
 - Circle sizes correspond to number of triples in datasets; arrow thicknesses correspond to number of links between datasets
 - Datasets qualify to be part of the diagram based on (arbitrary) numbers of triples and links
- Difficult to locate information on exactly how the links were created
- The quality of the 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, stating “...it is better to have no mapping at all than to establish a misleading one”
- Supplementary metadata is as important as the data itself – as a record of how, when and by whom the links were made

Comparing thesauri

- SENESCHAL project included SKOS conversion of:
 - EH Monument Types Thesaurus
 - RCAHMS Monument Types Thesaurus
 - RCAHMW Monument Types Thesaurus
- RCAHMS & RCAHMW thesauri both derived originally from EH thesaurus
- Better together?? Ideally shared conceptual knowledge about the domain would not be split along modern political boundaries
- As it is, at least there should be good potential for inter-thesaurus links?

SENESCHAL comparison 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 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 wrong...
 - Compared each preferred term from one thesaurus to all terms from another thesaurus – obtained best overall textual matches
 - Similarity threshold introduced to suppress low scoring matches. Levenshtein algorithm will *always* produce a match, even if it is a bad one!

Comparing terms between thesauri

Initial match suggestions - based on preferred terms

| RCAHMS concept | Best match | 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 EH monuments

| RCAHMS concept | Best match | 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 concept | Best match | 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 EH maritime

Exploring tools to establish concept links (1)

- OpenRefine (formerly Google Refine)
 - General tabular data cleansing / manipulation / conversion tool
 - ‘DBpedia Spotlight’ matched prefLabel with DBpedia terms. Not sure what to do next...
 - Optional RDF and Freebase extensions?

| re#broader | prefLabelNoLan | DBpedia Spotlight | |
|--------------------|------------------------|-----------------------------------|-------------------------------|
| | | | http://www.wikidata.org/wiki/ |
| | | | http://www.wikidata.org/wiki/ |
| obj/concepts/97593 | METAL WORKING DEBRIS | METAL WORKING Choose new match | http://www.wikidata.org/wiki/ |
| | | DEBRIS Choose new match | http://www.wikidata.org/wiki/ |
| obj/concepts/95070 | MINIATURE OBJECT | | http://www.wikidata.org/wiki/ |
| obj/concepts/96360 | FINIAL (ARCHITECTURAL) | FINIAL Choose new match | http://www.wikidata.org/wiki/ |
| | | ARCHITECTURAL Choose new match | http://www.wikidata.org/wiki/ |
| obj/concepts/97693 | LUCIFER MATCH | MATCH Choose new match | http://www.wikidata.org/wiki/ |
| obj/concepts/96067 | ROOF FINIAL | ROOF Choose new match | http://www.wikidata.org/wiki/ |
| | | FINIAL Choose new match | http://www.wikidata.org/wiki/ |
| obj/concepts/99629 | RELIQUARY | RELIQUARY Choose new match | http://www.wikidata.org/wiki/ |
| obj/concepts/97235 | | | http://www.wikidata.org/wiki/ |
| obj/concepts/97290 | MANUFACTURING DEBRIS | DEBRIS Choose new match | http://www.wikidata.org/wiki/ |

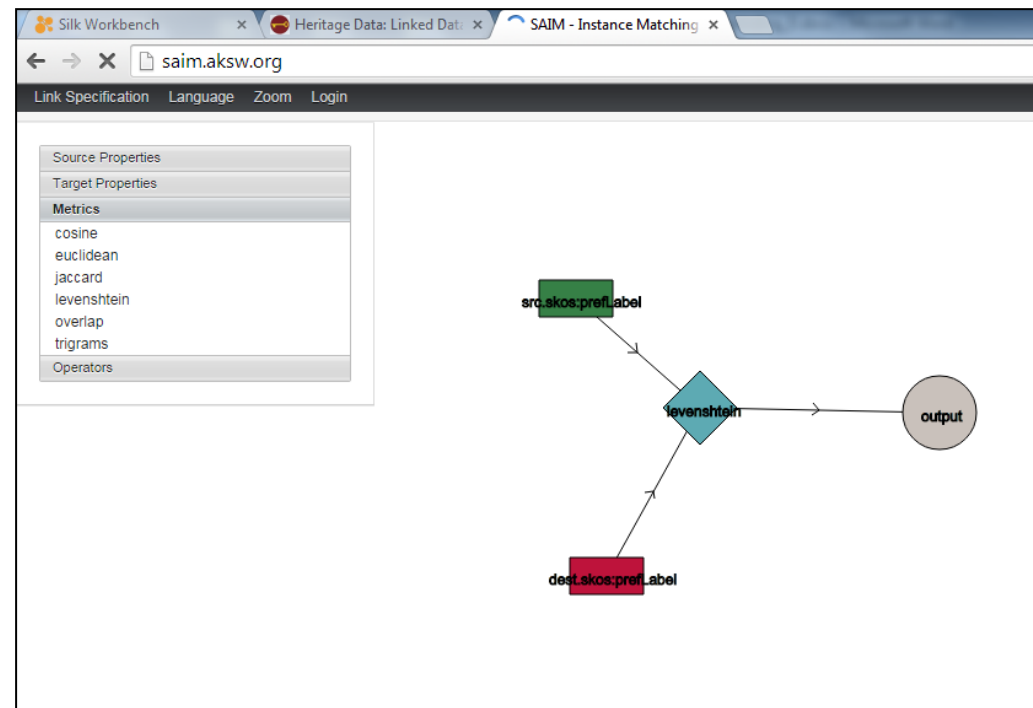
Exploring tools to establish concept links (2)

- LODRefine – specialised version of OpenRefine
 - Incorporates previously separate extensions
 - Reconciliation service, based on uploaded SKOS RDF files – compared prefLabels
 - Slow process – over 2 hours to compare RCAHMS monument types with EH monument types
 - Successfully suggested exact/partial matching of prefLabels, however selecting one just modifies the existing label – not link to underlying URI

| Show: 5 10 25 50 records | | |
|--------------------------|---|--------------------------------------|
| os/core#broader | <input type="text" value="http://www.w3.org/2004/02/sk"/> | <input type="text" value="http://"/> |
| /1/concepts/647 | "WINCH HOUSE"@en <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> WINCH HOUSE (0.688) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> WINCH (0.226) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> ELECTRIC WINCH (0.028) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match | http://purl. |
| | "TAIGH-UNNDAISE"@gd <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match | |
| /1/concepts/647 | "WOOD PROCESSING SITE"@en <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> WOOD PROCESSING SITE (0.8) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> METAL PROCESSING SITE (0.223) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> WOOD PRODUCT SITE (0.194) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match | http://purl. |
| | "LÀRACH GIULLACHD FIODHA"@gd <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match | |
| /1/concepts/1363 | "TIMBER PROCESSING SITE"@en <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> TIMBER PROCESSING SITE (0.815) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> FISH PROCESSING | http://purl. |

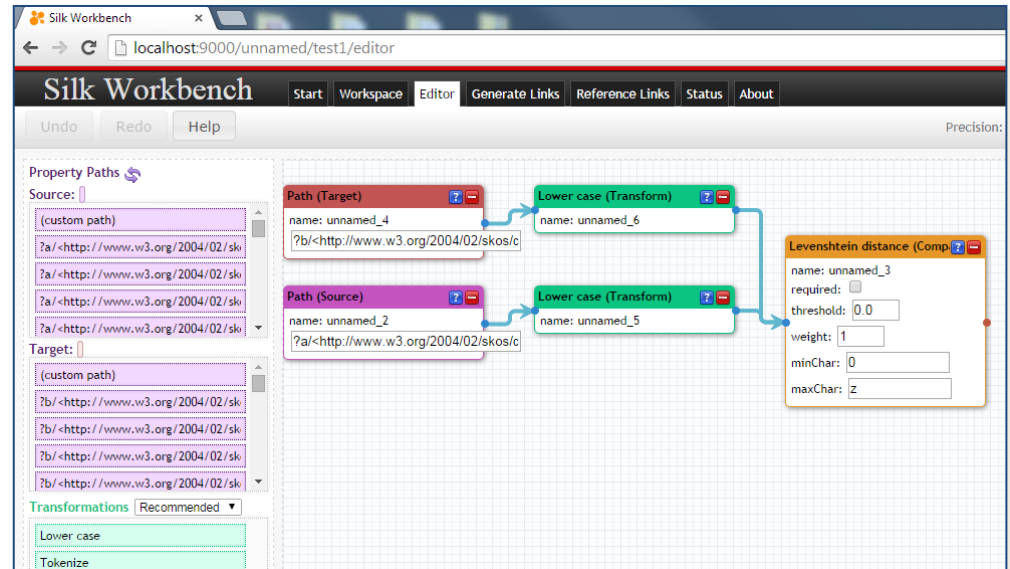
Exploring tools to establish concept links (3)

- SAIM
(<http://saim.aksw.org/>)
- Web interface to LIMES
- Link specification successfully set up (using Levenshtein comparison of preferred terms)
- Flashed up errors, no output. More configuration probably required



Exploring tools to establish concept links (4)

- Silk Workbench
 - Seems to have a lot of useful functionality
 - Link specification successfully set up
 - Achieved comparison of some preferred labels
 - Learning curve to do more and to use results produced



The screenshot shows the Silk Workbench Reference Links interface. The main area displays a comparison between two URIs:

| Source | Target | Score | Status | Correct |
|--|--|--------|---------|--------------------------|
| <code>http://purl.org/heritagedata/schemes/1/concepts/648</code> | <code>http://purl.org/heritagedata/schemes/eh_tmt2/concepts/91043</code> | 100.0% | correct | <input type="checkbox"/> |

Below the table, the comparison details are shown:

- Comparison: levenshteinDistance (unnamed_3) 100.0%
- Transform: lowerCase (unnamed_6) animal product site
 - Input: `?b/<http://www.w3.org/2004/02/skos/core#prefLabel>` (unnamed_4) ANIMAL PRODUCT SITE
- Transform: lowerCase (unnamed_5) animal product site
 - Input: `?a/<http://www.w3.org/2004/02/skos/core#prefLabel>` (unnamed_2) ANIMAL PRODUCT SITE

Compare concepts, not just terms

- Taking term matches at face value is an inadequate approach
- An exact match on a term does not mean an exact match on a concept
- Need to consider scope notes, synonyms and full hierarchical context
- Concept scope can change over time. We are only considering a snapshot when making a link, so need to produce associated metadata.

Heritage Data: Linked Dat: x

heritagedata.org/live/schemes/1/concepts/467.html

Heritage Data

Linked Data Vocabularies for Cultural Heritage

Scheme List | Concept Search | SPARQL Query | About The Project

<http://purl.org/heritagedata/schemes/1/concepts/467> (QR Code)

| Property | Value |
|------------------------------------|---|
| rdf:type | skos:Concept |
| cc:license | http://reference.data.gov.uk/id/open-government-licence |
| cc:attributionURL | http://www.rcahms.gov.uk |
| cc:attributionName | RCAHMS |
| skos:inScheme | Monument Type Thesaurus (Scotland) |
| skos:prefLabel | TENEMENT |
| skos:prefLabel | TEANAMANT [gd] |
| skos:broader | MULTIPLE DWELLING |
| skos:scopeNote | A large building containing a number of rooms or flats, access to which is usually gained via a common stairway. |
| skos:scopeNote | Togalach mòr sa bheil grunn sheòraichean no fhlàtaichean a ruigear air staidhir choitcheann mar is trice. [gd] |
| skos:altLabel | theanamantaibh [gd] |

Heritage Data: Linked Dat: x

heritagedata.org/live/schemes/eh_tmt2/concepts/68997.html

Heritage Data

Linked Data Vocabularies for Cultural Heritage

Scheme List | Concept Search | SPARQL Query | About The Project

http://purl.org/heritagedata/schemes/eh_tmt2/concepts/68997 (QR Code)

| Property | Value |
|------------------------------------|---|
| rdf:type | skos:Concept |
| cc:license | http://creativecommons.org/licenses/by/3.0 |
| cc:attributionURL | http://www.english-heritage.org.uk |
| cc:attributionName | English Heritage |
| skos:inScheme | MONUMENT TYPE (EH) |
| skos:prefLabel | TENEMENT |
| skos:broader | SETTLEMENT |
| skos:scopeNote | A parcel of land. |
| skos:related | DWELLING |
| dct:publisher | http://www.english-heritage.org.uk |
| dct:identifier | http://purl.org/heritagedata/schemes/eh_tmt2/concepts/68997 |
| dct:issued | 2013-07-17T08:43:50 |

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 translation in the case of multilingual terminology.
- *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.

Determine relationships between concepts

| EH monument types thesaurus | RCAHMS monument types thesaurus - suggested match (a) | RCAHMS monument types thesaurus - alternative match (b) |
|---|---|--|
| <p>PT : CUP AND RING MARKED STONE</p> <p>BT : ROCK CARVING</p> <p>RT : CUP MARKED STONE</p> <p>RT : CARVED STONE</p> <p>SN : <i>A stone, either in situ or part of a monument, bearing one or more small, roughly hemispherical depressions surrounded by a concentric arrangement of annular or pennanular grooves. More complex designs may also occur</i></p> | <p>PT: CUP AND RING MARKED STONE</p> <p>TT : RELIGIOUS RITUAL AND FUNERARY</p> <p>TT : MONUMENT (BY FORM)</p> <p>BT : CARVED STONE</p> <p>BT : CUP AND RING MARKINGS</p> <p>RT : CUP MARKED STONE</p> <p>RT : RING MARKED STONE</p> <p>RT : CUP AND RING MARKED ROCK</p> <p>SN : <i>A stone bearing one or more small, roughly hemispherical depressions surrounded by a concentric arrangement of annular or penannular grooves. More complex designs may also occur.</i></p> | <p>PT : CUP AND RING MARKED ROCK</p> <p>TT : RELIGIOUS RITUAL AND FUNERARY</p> <p>TT: MONUMENT (BY FORM)</p> <p>BT : CUP AND RING MARKINGS</p> <p>BT : ROCK CARVING</p> <p>RT : CUP MARKED ROCK</p> <p>RT : RING MARKED ROCK</p> <p>RT : CUP AND RING MARKED STONE</p> <p>RT : CUP MARKED STONE</p> <p>RT : RING MARKED STONE</p> <p>SN: <i>One or more small, roughly hemispherical depressions surrounded by a concentric arrangement of annular or penannular grooves carved on natural rock outcrop. More complex designs may also occur.</i></p> |





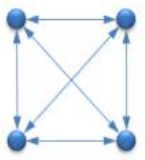



Suggested match (a) shows exact match on PT, plus SN and RT initially suggests substantial similarity - but an alternative concept (b) exists. Does the EH scope note cover both the RCAHMS concepts? What is the relationship?

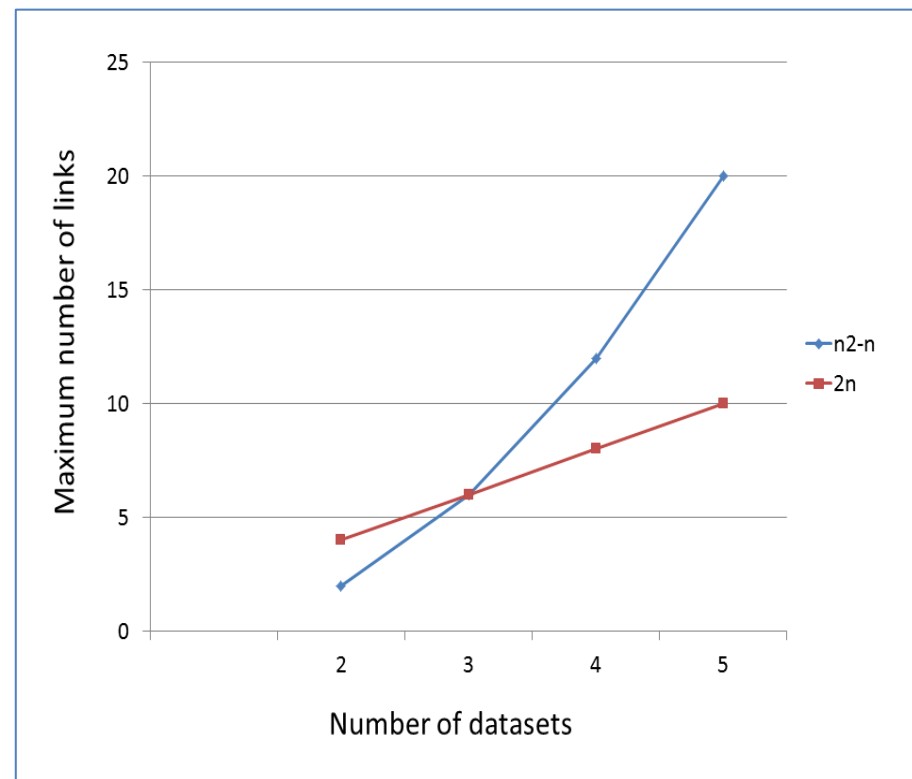
Link types

- The type of link is not suggested by the tools, they just establish a degree of similarity based on text matching
- Cannot use owl:sameAs because:
 - <A> skos:inScheme <X> ; skos:prefLabel “axe”@en .
 - <skos:inScheme <Y> ; skos:prefLabel “hatchet”@en .
 - <A> owl:sameAs . → wrong, would mean <A> and have 2 English preferred labels each - inconsistent
- SKOS instead provides inter-thesaurus concept mapping properties:
 - exactMatch
 - closeMatch
 - broadMatch – hierarchical mapping
 - narrowMatch – hierarchical mapping
 - relatedMatch – associative mapping
- (No compound mapping properties in SKOS)

Links: Many-to-many vs. hub architecture

- Number of bidirectional links when linking between multiple thesauri

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



Requirements?

- Creating concept → concept links, not term → term – so utilise more of the contextual data
- Needed more understanding of existing tools prior to use; tools trialled had a broader scope (general linked data)
- Any other tool suggestions? Need to:
 - Work top down through concept hierarchy, taking account of existing higher level mappings when suggesting matches
 - Work interactively and allow manual intervention. Automatically suggested matches really require human judgement
 - Employ a combination of similarity measures involving more than just preferred label matching
 - Facilitate simple side by side comparison of best matching concepts, with useful accompanying contextual information
 - Provide list of possible link types to choose from
 - Generate associated metadata to describe both the link and the process used to establish it, export in a chosen suitable serialisation format

Conclusions

- Need quality of data, not quantity
- Comparing concepts, not just terms
- Provide supplementary metadata describing how the links were created
- Automated mapping tools can assist, but results require manual assessment
- Achieving the ‘5th star’ is deceptively hard work

5 Star Data



Achieving the 5th Star

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