Progress in semantic mapping

Antoine Isaac

aisaac@few.vu.nl

Europeana, Vrije Universiteit Amsterdam

W3C Library Linked Data incubator group

NKOS workshop, ECDL, Glasgow, Sept. 9th 2010

Semantic mapping?

- Connecting metadata structures

  Element sets, schemas, ontologies (in the canonical Semantic Web meaning)

- Connecting vocabularies of values

  KOS, knowledge bases

Connecting metadata structures

- Crosswalking, metadata profile matching…
- For example to port a dataset to Dublin Core elements

- People have a quite clear idea now on creating mappings to cope with interoperability

  Data integration, interoperability with complementary data
- This is about the structure of data, and that is (usually) closely tied to an application

  If you don't do it you just can't access data easily

Connecting metadata structures

- It is a tedious job but automatic assistance is not really needed
  Expertise is key
- Many projects doing this routinely
  Aggregators (such as Europeana) and their network,
- Even building tools for it
  Related to Europeana: ATHENA, PrestoPrime, EuropeanaConnect

- There are some issues which are not thoroughly explored yet
  - Co-existence between different metadata structures
  - Sharing and re-using mappings

Semantic mapping of KOSs

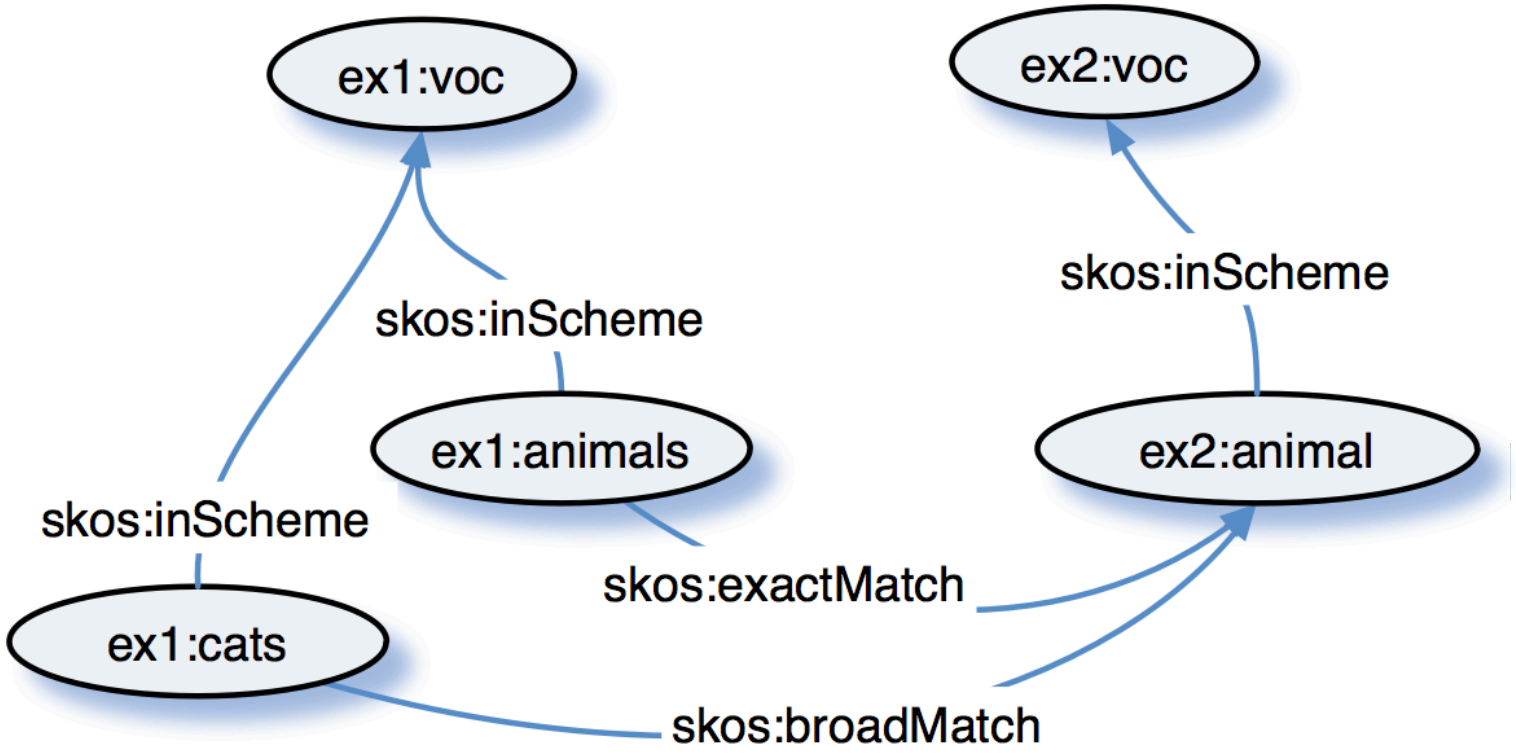
A wide area of “contextual resources”

- Thesauri
- Classification systems
- Person authority files
- Gazetteers
- Wikipedia
- Wordnet(s)

…

A SKOS perspective

SKOS allows bridging across KOSs from different contexts



Some landmark SKOS implementations

- Swedish National Library's Libris catalogue and thesaurus http://libris.kb.se/
- Library of Congress' vocabularies, including LCSH http://id.loc.gov/
- DNB's Gemeinsame Normdatei (incl. SWD subject headings) http://d-nb.info/gnd/
  Documentation at https://wiki.d-nb.de/display/LDS
- BnF's RAMEAU subject headings http://stitch.cs.vu.nl/
- OCLC's DDC classification http://dewey.info/ and VIAF http://viaf.org/
- STW economy thesaurus http://zbw.eu/stw
- National Library of Hungary's catalogue and thesauri, e.g., http://oszkdk.oszk.hu/resource/DRJ/404
- Wikipedia categories through DBpedia http://dbpedia.org/
- New York Times subject headings http://data.nytimes.com/
- IVOA astronomy vocabularies http://www.ivoa.net/Documents/latest/Vocabularies.html
- GEMET environmental thesaurus http://eionet.europa.eu/gemet
- UMTHES
- Agrovoc http://aims.fao.org/
- Linked Life Data http://linkedlifedata.com/
- Taxonconcept http://www.taxonconcept.org/
- UK Public sector vocabularies http://standards.esd.org.uk/ (e.g., http://id.esd.org.uk/lifeEvent/7)

http://www.w3.org/2001/sw/wiki/SKOS/Datasets

KOS Alignments?

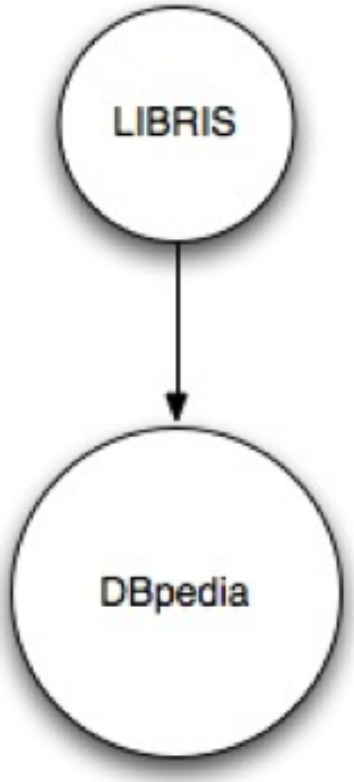
Quite many of them are linked to some other resource

- LCSH, SWD and RAMEAU interlinked through MACS mappings
- GND linked to DBpedia and VIAF
- Libris linked to LCSH
- Agrovoc to CAT, NAL, SWD, GEMET
- NYT to freebase, DBpedia, Geonames
- DBPedia links are overwhelming

  Hungary, STW, TaxonConcept, GND…

Order of magnitude: dozens of millions of concepts, millions of links – and growing

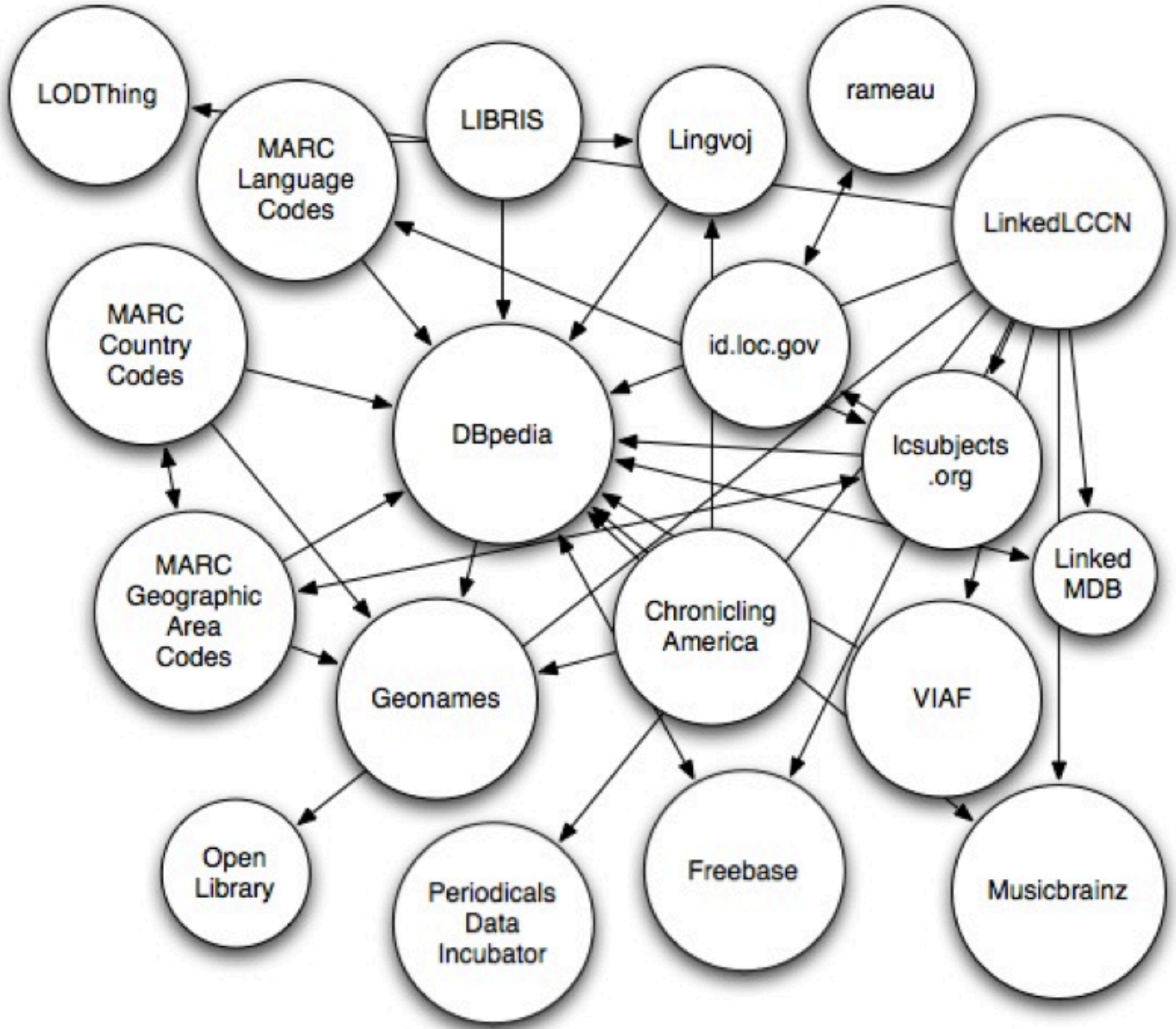
Example: Linked Library Cloud, beginning 2008



[Ross Singer, Code4Lib2010]

http://code4lib.org/conference/2010/singer

Example: Linked Library Cloud mid-2010



Plus:

- Germany NL
- Hungary NL
- STW
- GEMET
- NYT
- Agrovoc

[Ross Singer, Code4Lib2010] http://code4lib.org/conference/2010/singer

Is it good enough?

- Mapping management and publication
- Mapping links
- [Sparseness of linkage]

1. Mapping registries?

- There are good examples of vocabulary registries, both for MD elements and KOSs

  metadataregistry.org, work at FAO, OCLC, Glamorgan…
- They keep track of mappings, but these are not first-class citizens
- Mapping-oriented registries clearly marked as mapping-oriented are scarce and mostly by and for researchers

  NeOn project, CATCH, FinnONTO vocabulary services
- It can be very useful
  - Sharing and re-using mappings
  - Combining mappings

MD element mapping on the semantic web

- Direct re-use **and extension** is (quite) common
  - Some ontologies have been massively re-used: DC, FOAF

- With SW representation techniques, mappings are tied to schemas, which are tied to data
  - Best practice: URIs of ontology elements lead to formal definitions which include mappings
  - http://www.w3.org/2004/02/skos/core#altLabel rdfs:subPropertyOf http://www.w3.org/2000/01/rdf-schema#label
  - *"follow your nose" approach to crosswalking*

- Ensures a minimal but global service for (simple) mappings

MD element mapping on the semantic web

The main issues are higher-level

- Many extensions are produced now that more players come into play

  E.g., PND, VIAF could not use SKOS and FOAF alone

- There is a danger of continuously re-inventing the wheel

  Several vocabularies for FRBR!

- Yet the solution lies probably in good vocabulary repositories to find the right vocabulary to re-use

Repositories of crosswalks would be needed for complex mappings and for specific applications

- support for mapping creation

KOSs?

The situation is similar for KOSs

- With SKOS mappings are easy to publish and access

Organizational issues are more important

- The wheel has already been re-invented many times
- Many mappings will probably created by actors who don't own (and publish) the KOS data itself

Here (simple) repositories have a global role to play

- Cf. sameAs.org

What has changed in the last, say, 10 years?

Warning – own intuition ahead!

- Steady progress
- Data publication practices provide a great deal of the needed functionality
- The rest is work in progress
- But it is driven by well-identified (niche) applications

2. Mapping links?

For KOSs there are many links, to mention SW stuff alone

- owl:sameAs
- skos:exactMatch
- skos:closeMatch
- rdfs:seeAlso
- umbel:isLike
- skos:broadMatch, narrowMatch & relatedMatch

Mapping link confusion

- Various semantics, allowing to pick what corresponds best to:
  - the detected similarity
  - how the mapping should be used

- The problem is that they're not applied consistently

  Cf. owl:sameAs issue

- But does that really matter?
  - Cf sameAs.org approach
  - A catch-all attempt at co-reference resolution can already help solving many problems

More precision?

Not easy, cf. the fate of skos:AND, OR and NOT

- A draft for mapping combinations of concepts didn't go through due to lack of evidence for (coherent) practice
- AND sometimes looked like boolean operators ("+") for full text search engines or post-coordination-aware system, sometimes like pre-coordination for synthesized classes or subject heading strings ("--")
- For Semantic Webers it also looks like formal OWL constructs for combining classes and properties

More precision?

- “Coordination” appears to be too application- and/or KOS-specific
  - Or at least it appeared to us, compared to other KOS features
- It is also quite complex, not for Simple-KOS!
- Now that BS8723 is out, and the new ISO is on its way, situation can be clearer wrt. “+”
  - A future SKOS extension?

And then we could also discuss “(major/minor) overlap”, and what Stella will present next…

NKOS contribution?

- NKOS community has an impressive record of practical and theoretical work

  DESIRE, CARMEN, Renardus, AQUARELLE, LIMBER, MACS, SWAD-Europe, HILT, MSAC, Crisscross, KoMoHe, FAO, OCLC, BS8723…

- Problem: too much focus on types of KOSs?

  NKOS workshops are full of applications but the general typology work is less focused on mapping applications

- Need for consolidating work on
  - Gathering scenarios (cf. SKOS Use Cases effort)
  - Doing application-specific alignment and evaluation

Identify applications of KOS mappings

- Something like categories from the STITCH project?
  - Concept-based search
  - Book re-indexing
  - Integration of one thesaurus into the other
  - Thesaurus merging
  - Free-text search
  - Navigation

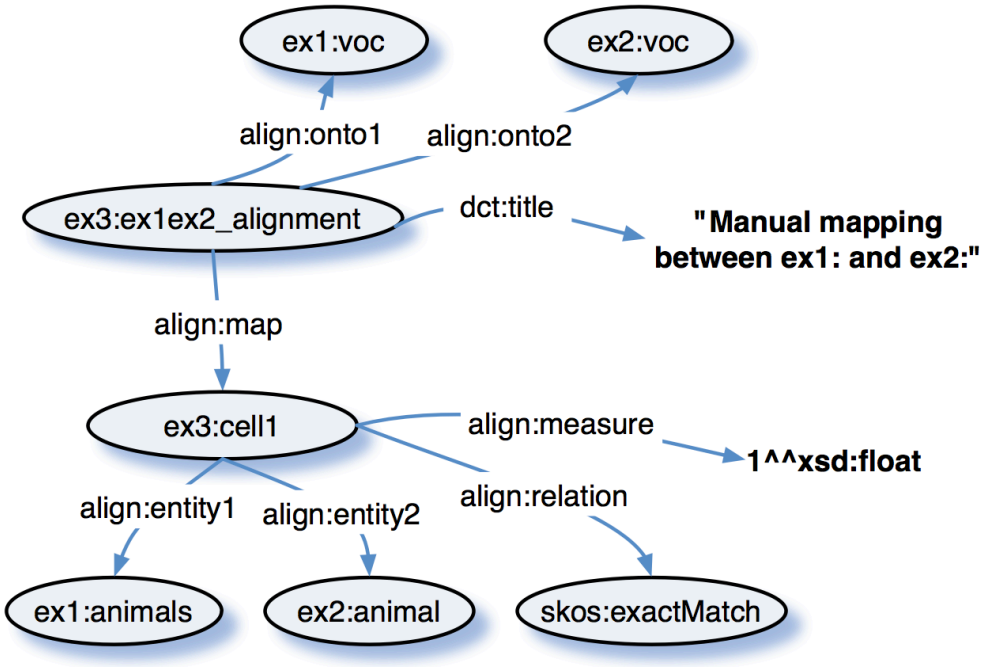
- A bit more precise than, say, IFLA categories

  Find, identify, obtain, select…

  hint at how mappings are used, which info they should provide

Not really a technical issue

- For specific needs, one can represent fine-grained mappings in RDF, in a flexible way
- That doesn't replace community agreement on mapping links, mapping techniques, mapping applications



http://alignapi.gforge.inria.fr/edoal.html

What has changed in the last 10 years?

Warning – own intuition ahead!

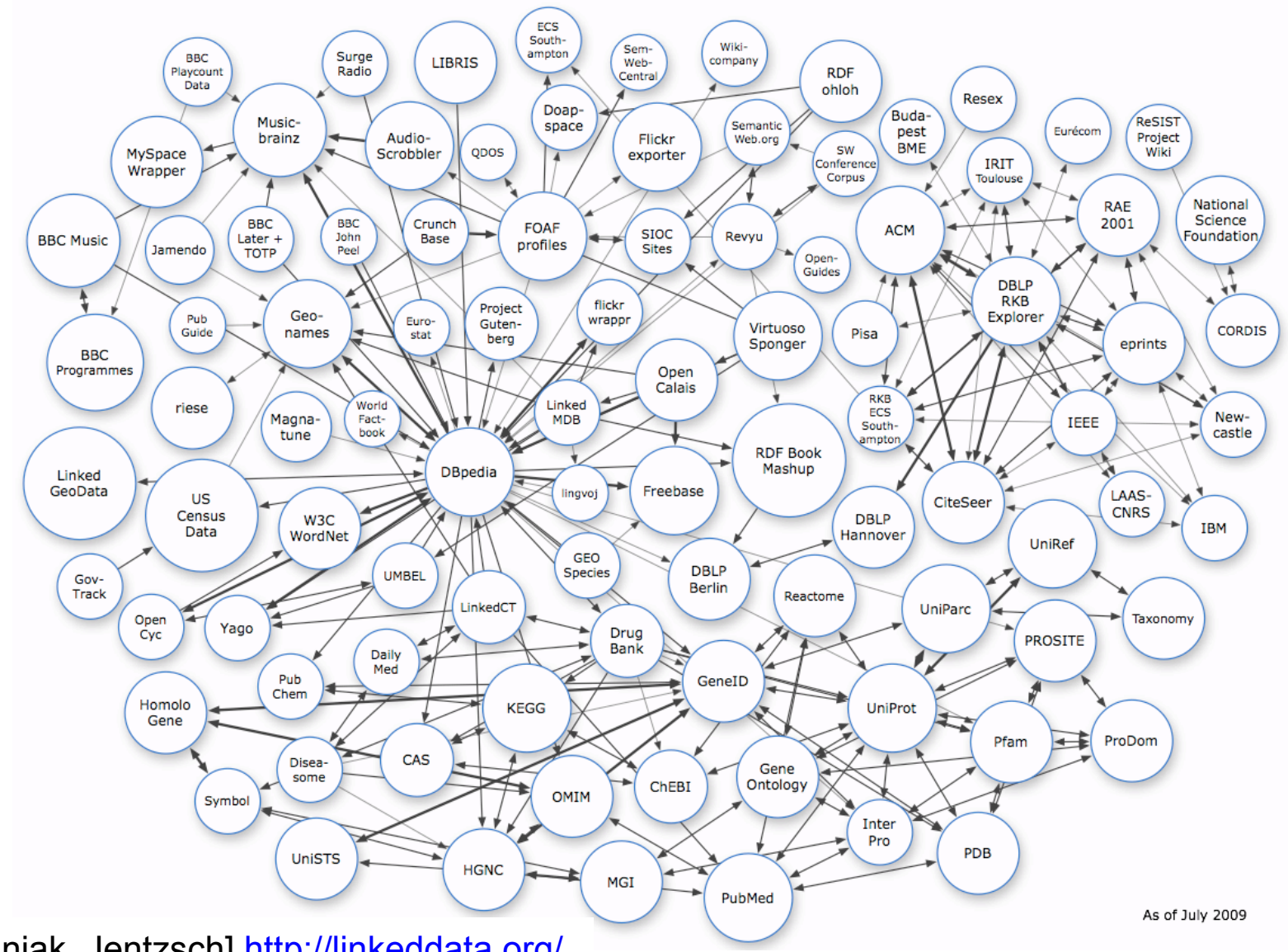
- Semantics have made little progress
  - What is precise (owl:sameAs) is badly used, what is loose is widely used
- Better connection needs to be made between application scenarios and mapping requirements
- *Real* applications remain rare
- Large amounts of data are accessible!

3. Sparseness of linkage?

Cf. previous wet finger estimation

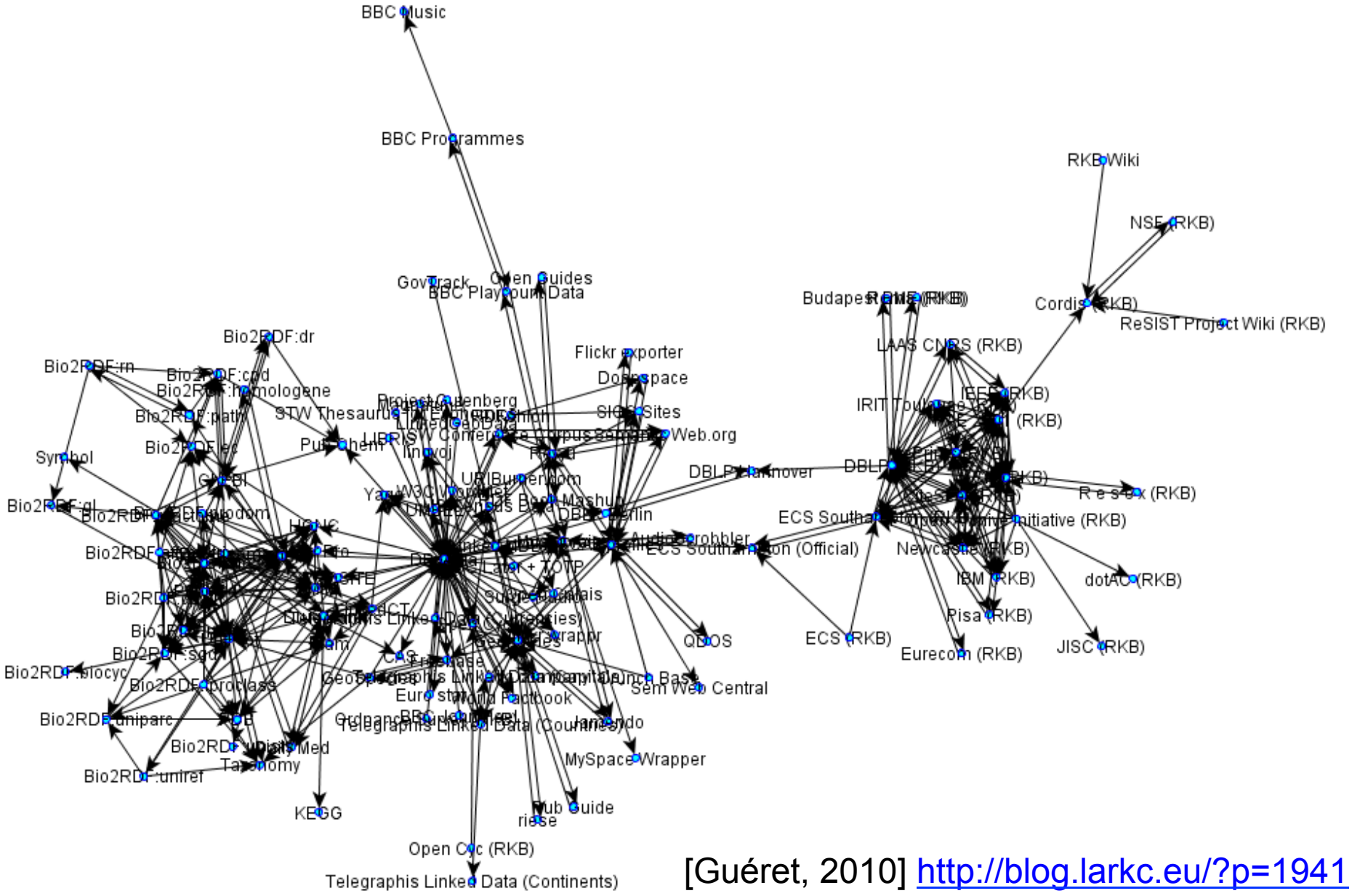
> “Dozens of millions of concepts, millions of links”

Sparse linkage: the LD cloud



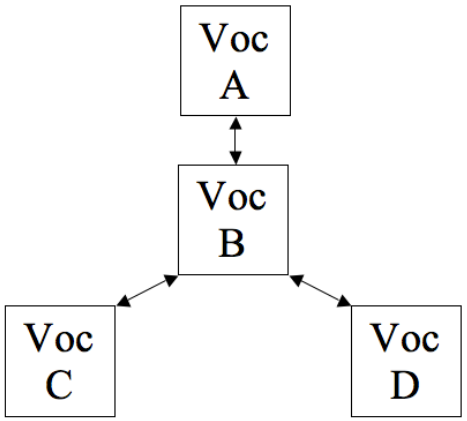
[Cyganiak, Jentzsch] http://linkeddata.org/

Sparse linkage: another view



[Guéret, 2010] http://blog.larkc.eu/?p=1941

Pivot datasets in the cloud?



- Looks like BS8723 backbone approach
- Hypothesis: a NKOS cloud would look the same

  Not with strict pivots, but still high centrality
- There are dangers compared to a direct mapping pattern
  - Longer paths between concepts
  - Dependence on the pivot

Do we have the means to do better?

- Manual mapping is still a bottleneck and lacks good assistance tools
  - SW community has focused on ontology matching
- Linked Data is changing the focus
  - E.g., SILK http://www4.wiwiss.fu-berlin.de/bizer/silk/
- Still, automatic tools take time to build and deploy for specific case
  - EuropeanaConnect project will follow a (loose) pivot approach for Europeana's "semantic layer"

What has changed in the last 10 years?

Warning – own intuition ahead!

- Large amounts of data are accessible! Yes.
- Minds have changed
- Means have not changed much
- But they are changing

  Project money is switching towards establishing linked data-style connections, in Europeana network and elsewhere

Conclusion

A NKOS view on Mike Uschold's "semantic elephants" for linked data?

- Versioning and URIs
- Overloading owl:sameAs
- Proliferation of URIs, Managing Co-reference

http://lists.w3.org/Archives/Public/public-lod/2010May/0012.html

Thanks!

Wanna participate?

SKOS mailing list	public-esw-thes@w3.org
SKOS wiki	http://www.w3.org/2001/sw/wiki/SKOS
Library Linked Data community list	public-lld@w3.org
LLD wiki	http://www.w3.org/2005/Incubator/lld/wiki
Linked Open Data community	http://linkeddata.org
LOD mailing list	public-lod@w3.org