URI Disambiguation in the Context of Linked Data

Afraz Jaffri, Hugh Glaser, Ian Millard

ECS, University of Southampton
Presentation Outline

- Linked Data Repositories
- Coreference on the Semantic Web
- Author Disambiguation
- DBLP Linked Data
- DBLP Author Disambiguation
- Disambiguation Results
- DBpedia
- Possible Solutions
- Summary
RKBexplorer.com

- Contains URIs for more than 10 million entities
- Over 25 Linked Data sites, including:
  - Data relating to people, projects, papers and institutions
  - A single entity has a number of URIs (even within the same repository)
  - Entities are linked using CRSes

DBLP
Linked Data Repositories

- Existing databases on the Web are being exposed as Linked Data (D2R, Virtuoso)
- Databases contain inconsistencies and require constant curation
- Datasets such as Wikipedia are being continually checked and updated, especially in the case of disambiguation (WikiProject_Disambiguation)
- Linked Data repositories should also provide consistent data
Disambiguation on the Semantic Web

- Coreference on the Semantic Web is defined as being the situation where two or more URIs are used for a single non-information resource.

- URI usage can change with context.

- Non-Information resource equality is hard to define precisely.

Examples:
- ‘Hugh Glaser’ at Southampton vs. ‘Hugh Glaser’ at Imperial.

- ‘Harry Potter and the Order of the Phoenix’ in Hardback vs. Softback

  ISBN: 978-0747561071 978-0747551003
URI Multiplicity

Å URIs for ‘Spain’:
Å http://dbpedia.org/resource/Spain
Å http://ww4.wiwiss.fu-berlin.de/factbook/resource/Spain
Å http://sws.geonames.org/2510769
Å http://www4.wiwiss.fu-berlin.de/eurostat/resource/countries/Espa%C3%B1a

Å URIs for ‘Hugh Glaser’:
Å http://acm.rkbexplorer.com/id/resource-P112732
Å http://citeseer.rkbexplorer.com/id/resource-CSP109020
Å http://citeseer.rkbexplorer.com/id/resource-CSP109013
Å http://citeseer.rkbexplorer.com/id/resource-CSP109011
Å http://citeseer.rkbexplorer.com/id/resource-CSP109002
Å http://dblp.rkbexplorer.com/id/resource-27de9959
Å http://europa.eu/People/#person-0ff816fa
Å http://resist.ecs.soton.ac.uk/wiki/User:hugh_glaser
Å http://id.ecs.soton.ac.uk/people/21
Author Disambiguation

- A known problem in the Information Science field
- How to determine: Hugh Glaser/H. Glaser/Glaser, H. are the same person?
- How to determine: Tom Anderson – Newcastle University Tom Anderson – University of Washington are different people?
Existing Approaches

- **String Metrics**
  - Name Equivalence identification
  - Record Linkage
  - Citation Matching

- **Web Assisted**
  - Look up publications on author’s home page
  - Use search engine results on publication title

- **Machine Learning**
  - k-way spectral clustering
  - Use author name, co-author frequency and publication venue
DBLP Linked Data

- Converted from an XML dump of DBLP database
- 950 000 Publications
- 540 000 Authors
- 28 million triples
- Updated Weekly
- Linked to other datasets including RDF Book Mashup and RKBExplorer.com

LDOW2008 - Beijing, China
DBLP Author Disambiguation

å 49 names - 10 most common English surnames with 5 common first names
å Authors disambiguated by looking at homepage, web publication, search engine results and institution
å When in doubt, authors assumed to be the same if:
  - The co-authors of any publication are the same
  - The publication venue was the same
  - The area of research was the same
It’s all about Identity

Tom Anderson – http://www4.wiwiss.fu-berlin.de/dblp/resource/person/109074

Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/dac/MorettiHNCKABDF01>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/sc/BorchersASW95>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/words/AndersonFRR05>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/bell/LiuBFSRA04>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/cj/LemosSA92>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/dt/Anderson01>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/dt/Anderson03>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/software/LemosSA95>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/ton/SavageWKA01>
Is [dc:creator] of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/journals/tse/AndersonBHM85>
DBLP Author Disambiguation Results

- 92% of authors with common names had publications incorrectly merged
- Worst case - 15 different authors with 1 URI
- Many authors who are the same have publications under different names (Cliff Jones, C.B. Jones)
- Inconsistency in data means inconsistency with linked data
- It is incorrect to use owl:sameAs to link different authors who have the same URI
DBpedia

- DBpedia 3.0 improves disambiguation management by including the ‘disambiguates’ property
- owl:sameAs linkage still inconsistent:
  
  <http://dbpedia.org/resource/Welsh>       owl:sameAs

  <http://dbpedia.org/resource/H.P._Lovecraft>       owl:sameAs
Possible Solutions

Å CRS: Consistent Reference Service
- Groups similar URIs into ‘bundles’
- Bundles can be made according to context
- Each KB can have one or more CRSes

Å OKKAM
- Coming up soon!
Summary

- Linked Data providers need to think about data consistency in the same way as database providers.
- Failure to manage coreference within datasets leads to incorrect linkage with other datasets.
- The network effect of the Web of Data means coreference needs to be even more carefully managed than in the Web of Documents.
- Systems are being developed to help manage coreference, the community needs to decide how to handle the problem.
Questions?

Further questions:

a.o.jaffri
hg @ecs.soton.ac.uk
icm