http://dbpedia.org/resource/Tim_Berners-Lee

http://dbpedia.org/resource/Spain

Electronics & Computer Science
University of Southampton

http://acm.rkbexplorer.com/id/resource-P112732

URI Disambiguation in the Context of Linked Data

http://sws.geonames.org/2510769

http://acm.rkbexplorer.com/id/person-282197

http://id.ecs.soton.ac.uk/person/7113

http://www.w3.org/People/Berners-Lee/card#i

Afraz Jaffri, Hugh Glaser, Ian Millard

ECS, University of Southampton

http://id.ecs.soton.ac.uk/person/21

http://www4.wiwiss.fu-berlin.de/dblp/resource/person/100007

http://citeseer.rkbexplorer.com/id/resource-CSP109020

http://southampton.rkbexplorer.com/id/person-00021

http://www4.wiwiss.fu-berlin.de/factbook/resource/Spain

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:

DBLP

- 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

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

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

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

DBLP Author Disambiguation

- 49 names - 10 most common English surnames with 5 common first names
- Authors disambiguated by looking at homepage, web publication, search engine results and institution
- 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

Electronics & Computer Science
University of Southampton

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/ftcs/SaeedLA91>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/ftrtft/LemosSA92>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/hybrid/AndersonLFS92>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/iccbss/AndersonFRR03>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/iciap/TruccoARI05>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/icnp/ElySWSA01>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/ifip/AndersonRR04>
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/seaai/AndersonH98>
is dc:creator of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/srds/Anderson86>
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/dt/ZorianASTI96>
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>
is dblp:editor of <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/sigcomm/2006>

[illegible] inc. USA

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://sw.cyc.com/2006/07/27/cyc/EthnicGroupOfWelsh> .
<http://sw.cyc.com/2006/07/27/cyc/Welsh-TheWord> .
<http://sw.cyc.com/2006/07/27/cyc/WelshLanguage> .
<http://sw.cyc.com/2006/07/27/cyc/Welshing-Cheating> .

<http://dbpedia.org/resource/H.P._Lovecraft> owl:sameAs
<http://sw.cyc.com/2006/07/27/cyc/HPLovecraft-Author> .
<http://zitgist.com/music/artist/8047a401-5ca7-48dd-9d7c-2d2b822e51e6> .

Possible Solutions

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