Co-reference and sameAs.org

DevCSI Linked Data Hackdays
ILRT, Bristol, UK, 14th January 2011

Hugh Glaser
Seme4 Limited and University of Southampton, UK

http://eprints.ecs.soton.ac.uk/id/eprint/21897
The LOD “cloud”

Co-Reference

• Lots of resources with multiple URIs
  – This is the nature of the Beast
  – Legal, Sociological and Technical reasons

• This is a Big Problem
  – Everything is a URI (not title, name, number…)
  – Identifying multiple URIs for one resource
  – Rejecting incorrectly conflated resources
  – Publishing
  – Using

• Solution
  – Co-reference is just Knowledge, but quite special, and crucial, knowledge
  – Generating a new URI is not the solution to already having too many
  – The web is anarchic – don’t try to impose impossible structure on it
  – Embrace the anarchy and multiple URIs in your systems ab initio
Co-Reference Service (CRS)

- **CRS Subsystem**
  - Find co-references
  - Store them
  - Publish them
    - Essentially:
      - $\text{URI}_i \rightarrow \{ \text{URI}_1, \ldots, \text{URI}_i, \ldots, \text{URI}_n \}$
    - Recommend a “Canon”

- **Published by the Data Publisher**
  - And possibly others

- **Middleware aggregates co-references from recognised CRSes**
CRS continued

- **CRS Policies are defined by context**
  - Often one per Triplestore
  - Can be many per Triplestore for different purposes
  - May not be associated with a particular Triplestore

- **Maintenance**
  - Provenance
  - Rollback

- **Can be used to infer owl:sameAs**

- **Eg OAI CRS has**
  - 7531045 different URIs
  - in
  - 2544955 bundles
Look up Seungwoo

kisti.rkbexplorer.com

Canon

http://kisti.rkbexplorer.com/id/PER_00000000000000131417

Duplicate URIs

http://acm.rkbexplorer.com/id/person-407157-c23d3b44f74d68c8ce281e3f4795bf4c
http://citeseer.rkbexplorer.com/id/resource-CSP180445-4e6159a7e2af5838836a83e27748b7ac
http://dblp.rkbexplorer.com/id/people-1ec5a60299222d6374695ef5214f05-3aac82bb29cfbe2e83d8b23349445
http://kisti.rkbexplorer.com/id/PER_00000000000000131417

Bundle instantiated at
2009-12-20 02:39:10

Alternative representations

RDF/XML
sameAs list in N3
image/png
Co-Reference Closure

Complete Co-Reference Information

This service computes the equivalence class within the known URIs for a specified URI by consulting all relevant CRS knowledge bases.

http://acm.kbexplorer.com/id/doi-10.1007/978-94-007-4866-9

Display all related triples? ○
Display connectivity image? ○ None ○ JPEG ○ PNG ○ GIF ○ Postscript ○ SVG

Equivalent URIs...

1. (Canon) http://acm.kbexplorer.com/id/person-407157-c23d3b4f7466c4281e3f7775af6c
2. http://classier.kbexplorer.com/id/resource-CLASSIFIER-10445-4f8f7a57c3f838baa2a2f77407874c
3. http://classier.kbexplorer.com/id/resource-CLASSIFIER-10445-4f8f7a57c3f838baa2a2f77407874c
4. http://dblp.kbexplorer.com/id/people-1e5a602092023d8374665f32140405-5c1471358870e3d31a4714469e60b41
5. http://dblp.kbexplorer.com/id/people-1e5a602092023d8374665f32140405-5c1471358870e3d31a4714469e60b41

The following diagram shows the interconnectivity between the CRS knowledge bases which maintain the context-dependent representation of coreference for each of the KBKBExplorer domains.

http://acm.kbexplorer.com/id/person-407157-c23d3b4f7466c4281e3f7775af6c

Showing information queried from all repositories...

<table>
<thead>
<tr>
<th>Subject</th>
<th>Property</th>
<th>Object/Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seungwoo Lee</td>
<td>akr.full-name</td>
<td>Seungwoo Lee [Explore]</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>akr.full-name</td>
<td>Seungwoo Lee [Explore]</td>
<td>acm-proceedings.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>akr.full-name</td>
<td>Seungwoo Lee [Explore]</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>akr.has-affiliation</td>
<td>Electrical and Computer Engineering Division, Pohang University of Science &amp; Technology (POSTECH), Pohang, South Korea</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>akr.has-affiliation</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>kisti:angNameOfPerson</td>
<td>Seungwoo Lee [Explore]</td>
<td>datasetproperties.ttl</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>akr:Affiliated-Person</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>akr:Affiliated-Person</td>
<td>acm-proceedings.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>Generic Agent</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>Generic Agent</td>
<td>acm-proceedings.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>Generic Agent</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>akr:Person</td>
<td>acm-periodicals.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>akr:Person</td>
<td>acm-proceedings.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>akr:Person</td>
<td>acm-proceedings.rdf</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>PER_char207**</td>
<td>datasetproperties.ttl</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>PER_char207**</td>
<td>objectproperties.ttl</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdf.type</td>
<td>PER_char207**</td>
<td>resources.ttl</td>
</tr>
</tbody>
</table>

Seme4.com
<sameAs>
interlinking the Web of Data

The Web of Data has many equivalent URIs. This service helps you to find co-references between different data sets. Enter a known URI, or use Sindice to search first.

Why not try searching for the string “Southampton”, (which we will look up for you on Sindice first) or finding other equivalent identifiers for http://transport.data.gov.uk/id/local-authority/1755?
<sameAs>
interlinking the Web of Data

The Web of Data has many equivalent URIs. This service helps you to find co-references between different data sets. Enter a known URI, or use Sindice to search first.

Equivalent URIs for http://acm.rkexplorer.com/id/person-344165-529a238779b9822d610a72843a3980c –
2. http://acm.rkexplorer.com/id/person-344165-529a238779b9822d610a72843a3980c
3. http://acm.rkexplorer.com/id/person-344165-be4c940afa0b483a2d1675df9fa0bab6f
4. http://acm.rkexplorer.com/id/person-344165-c23ba3f792faeca763400f197ba9adb1
5. http://acm.rkexplorer.com/id/person-344165-e0b9ee0f283efab1c5ae86a8b1f111
7. http://dblp.rkexplorer.com/id/people-2ceb813d69a235014ed3d46c881a1...86ce3c9dc7
8. http://dblp.rkexplorer.com/id/people-2ce8813d69a235014ed3d46c881a1...86ce3c9dc7
10. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...88ed85ab7e
11. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...601d0b0236
12. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...939b59e9f
13. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...872498927e
14. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...411b54f75
15. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...3621e49944
16. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...3a2580a589
17. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...b88af8f4e0
18. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...69229368c
19. http://dblp.rkexplorer.com/id/people-c9d903eb4b98a73185165b5f9...a42f7e7ace
20. http://citeseer.rkexplorer.com/id/resource- CSP162937-12a5a2893245...80ce1a874c
21. http://citeseer.rkexplorer.com/id/resource-CSP162937-2887e967049a...00d1d4c62c
22. http://citeseer.rkexplorer.com/id/resource-CSP162937-4bfe00a2ed7f9_90425f22
23. http://citeseer.rkexplorer.com/id/resource-CSP162937-64595e407c94...7867a8505d
24. http://citeseer.rkexplorer.com/id/resource-CSP162937-8b4779db34ec9...af2d24ddc4
25. http://citeseer.rkexplorer.com/id/resource-CSP162937-7866464de27c6...899db880b
What is Southampton?
Concluding Remarks

- RKBExplorer uses CRS services successfully
  - Often 100s of URIs per person

- British Museum uses a CRS service for Data Fusion on its web site

- Organisations can go ahead and build systems, then do the identity management
  - Both internally and externally
  - Doing identity management first requires cross-organisation agreement, a delay of months to years
  - It then freezes the business processes to conform to the external requirement

- sameAs.org turns out to be pretty useful
  - > 10K hits/day

- Build systems that accept the way the world is, not what you would like it to be
You are never alone

• Ian Millard
• ReSIST Project
• AKT Project
• EnAKTing Project
• 10 years of collaborators

http://eprints.ecs.soton.ac.uk/id/eprint/21897