RKB, sameAs and dotAC

at
“2009: Beyond the Repository Fringe”

Edinburgh
30-31 July 2009

Hugh Glaser & Ian Millard

<sameAs>
interlinking the Web of Data
Linked Data

- Tim Berners-Lee
  - “the Semantic Web done right, and the Web done right”
  - http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
Design Issues - Linked Data

Tim Berners-Lee

Date: 2006-07-27, last change: $Date: 2007/05/02 14:30:56$

Status: personal view only. Editing status: imperfect but published.

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

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information.
- Include links to other URIs, so that they can discover more things.
RDF

• `<subject-uri> <predicate-uri> <object-uri>`

  or

• `<subject-uri> <predicate-uri> “String”`

• Eg

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```
4.5 billion triples
180 million data links
We are pleased to acknowledge the help and work provided by our ReSIST Partners and many others.

This work was supported by the ReSIST Network of Excellence, which is sponsored by the Information Society Technology (IST) priority under contract number IST-4-02026764 NOE.
Some Underlying Sources

<table>
<thead>
<tr>
<th>Website</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>acm.rkbexplorer.com</td>
<td>italy.rkbexplorer.com</td>
</tr>
<tr>
<td>budapest.rkbexplorer.com</td>
<td>kaunas.rkbexplorer.com</td>
</tr>
<tr>
<td>citeseer.rkbexplorer.com</td>
<td>kisti.rkbexplorer.com</td>
</tr>
<tr>
<td>cordis.rkbexplorer.com</td>
<td>laas.rkbexplorer.com</td>
</tr>
<tr>
<td>courseware.rkbexplorer.com</td>
<td>lisbon.rkbexplorer.com</td>
</tr>
<tr>
<td>darmstadt.rkbexplorer.com</td>
<td>newcastle.rkbexplorer.com</td>
</tr>
<tr>
<td>dblp.rkbexplorer.com</td>
<td>nsf.rkbexplorer.com</td>
</tr>
<tr>
<td>dbpedia.org</td>
<td>pisa.rkbexplorer.com</td>
</tr>
<tr>
<td>deepblue.rkbexplorer.com</td>
<td>rae2001.rkbexplorer.com</td>
</tr>
<tr>
<td>deploy.rkbexplorer.com</td>
<td>resex.rkbexplorer.com</td>
</tr>
<tr>
<td>epsrc.rkbexplorer.com</td>
<td>roma.rkbexplorer.com</td>
</tr>
<tr>
<td>eurecom.rkbexplorer.com</td>
<td>southampton.rkbexplorer.com</td>
</tr>
<tr>
<td>ft.rkbexplorer.com</td>
<td>ulm.rkbexplorer.com</td>
</tr>
<tr>
<td>ibm.rkbexplorer.com</td>
<td>unlocode.rkbexplorer.com</td>
</tr>
<tr>
<td>ieee.rkbexplorer.com</td>
<td>wiki.rkbexplorer.com</td>
</tr>
<tr>
<td>irit.rkbexplorer.com</td>
<td>xxx.yyy.zzz</td>
</tr>
</tbody>
</table>

Range from a few 100 to more than 10,000,000 “facts”
Knowledge Sources

• Partners
• Publications
• Funding Agencies
• Project Wiki
• Courseware
• Resilient-Explicit Computing

• Wide range, don’t just look where you expect to find
For example

- Statistics for repository kisti.rkbexplorer.com
  - Last data assertion 2008-09-18 17:16:41
  - Number of triples 12815162
  - Number of symbols 3239105
  - Size of RDF dataset 671M
Co-Reference

- Repositories have people, publications, etc. from other institutions who also have records there and elsewhere

- And vice versa

- Co-Reference is a Big Problem
  - Everything is a URI (not title, name, number...)
  - Identifying multiple URIs for one resource
  - Rejecting incorrectly conflated resources
  - Publishing
  - Using

- Coldstart
  - A serious problem
  - Nothing is linked to anything
Co-Reference Service (CRS)

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

- Published by the Data Publisher
  - And possibly others

- Middleware aggregates co-references from recognised CRSes
CRS – Consistent Reference Service

- A service to manage and publish co-referent information
- Identify co-referent pairs using a set of tools
- Assert into the CRS
- Query the CRS
  - URI_i -> { URI_1, ..., URI_i, ..., URI_n }
- Recommend a Canon
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
Open System

- RKBExplorer is only one interface
  - And not a required part
- Services:
  - Details for a paper (the right hand pane in RKBExplorer):
  - Network of people for a publication (lower pane):
    - ...
- Other Interfaces (using the services)
  - Personal Web pages
  - iPhone
  - iGoogle Gadget
Or a Paper
Or a Couple of People

And how they are linked
And Why they are Linked?

Carl Lagoze is connected to Dean Krafft

They are linked by 9 relations.

**Publications**

They have co-authored 6 papers:

- Core services in the architecture of the national science digital library (NSDL)
- An information network overlay architecture for the NSDL
- Metadata aggregation and automated digital libraries*: A retrospective on the NSDL experience

(3 more)

**Affiliations**

They are both affiliated to Cornell University.

**Projects**

They are both members of 2 projects:

- NSDL Technical Network Services: A Cyberinfrastructure Platform for STEM Education
- Collaborative Project: Core Integration - Leading NSDL toward Long-Term Success
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.

Equivalent URIs...
1. (Canon) http://acm.rkbexplorer.com/id/person-407157
4. http://dblp.rkbexplorer.com/id/people-1ec5a600299222dd6374695e5f21405-906423eb148125a6e65c573dc5a15e430c

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

Seungwoo Lee

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>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>dblp-publications-2001 rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>acm-periodicals rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>Science &amp; Technology (POSTECH), Pohang, South Korea</td>
<td>acm-periodicals rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>Electrical and Computer Engineering Division, Pohang University of Science &amp; Technology (POSTECH), Pohang, South Korea</td>
<td>acm-periodicals rdf &gt;&gt;</td>
</tr>
</tbody>
</table>

<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>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>Seungwoo Lee</td>
<td>rdfs:label</td>
<td>POSTECH, Pohang, Korea</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Property</th>
<th>Object</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic acquisition of named entity tagged corpus from world wide web</td>
<td>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>A Corpus-Based Learning Method of Compound Noun Indexing Rules for Korean</td>
<td>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>A Corpus-Based Learning Method of Compound Noun Indexing Rules for Korean</td>
<td>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>acm-proceedings rdf &gt;&gt;</td>
</tr>
<tr>
<td>A Corpus-Based Learning Method of Compound Noun</td>
<td>rdfs:label</td>
<td>Seungwoo Lee</td>
<td>dblp-publications-2001 rdf &gt;&gt;</td>
</tr>
</tbody>
</table>
Finding Co-reference

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 Sindico to search first.

Search results from Sindico, with co-references applied...

- **Les Carr**
  2. http://southampton.rkbexplorer.com/id/person-9be028dd84b079a54318...fa4814122
  3. http://southampton.rkbexplorer.com/id/person-a63585b18c81bfdf7dab...84425e0250
  
  Show 118 more

- **Les Carr**
    - rdf+xml - rdf - json - text

- **Les Carr**
  1. http://dblp.l3s.de/db2/resource/authors/Les_Carr
    - rdf+xml - rdfs - json - text

Currently serving 23268924 URIs in 8454882 bundles!
Where is it Taught?

Budapest University of Technology and Economics

Courses taught at Budapest University of Technology and Economics, Budapest:
- Software Verification and Validation
- Management of Computing Infrastructure

Istvan Majzik
Gadget – find out about people

Mark Borkum did this
This is a page that gives a simple demonstration showing papers which have been deemed related through textual analysis by IAI Saarbrucken. Up to the top 20 are listed for each paper, when they meet a simple thresholding:

1 – very strong – 0.9 – strongly – 0.7 – related – 0.6 – ignored – 0

The 1980 paper **Exception Handling and Software-Fault Tolerance** [browse]

is very strongly related to

- [browse] 2003 "Automatic detection and masking of non-atomic exception handling" [PDF]
- [browse] 1989 "Formal Verification of Programs with Exceptions"
- [browse] 1983 "Programming Reliable and Robust Software in ADA"

is strongly related to

- [browse] 1998 "Improving software robustness with dependability cases" [PDF]
- [browse] 1999 "Wrapping windows NT software for robustness" [PDF]
- [browse] 1981 "Exception Handling and Error Recovery Techniques in Modular Systems - An Application to the Isaure System"
- [browse] 2003 "Deadlock resolution via exceptions for dependable Java applications" [PDF]
- [browse] 2002 "Robust software - no more excuses" [PDF]

is related to

- [browse] 1995 "Fault tolerance in concurrent object-oriented software through coordinated error recovery" [PDF]
- [browse] 2004 "Implementing simple replication protocols using CORBA portable interceptors and Java serialization" [PDF]
- [browse] 1984 "Fault Tolerance Using Communicating Sequential Processes"
- [browse] 2001 "Middleware support for voting and data fusion" [PDF]
Concluding Remarks

- ePrints today, other systems tomorrow
  - Other related technologies (such as OAI-ORE)
  - Are they right for this?
- Please don’t stop at the repository
- Go on and get the added value of Linked Data
- ePrints has plans to publish RDF
  - Will the schema (ontology) by expressive enough
- Worry about your co-reference
  - Do you have IDs in your repository?
  - Can you reliably identify all the papers of a single person?
And so ...

dotAC: Exploring the UK Research Landscape

contact@dotac.info