RKBExplorer: Repositories, Linked Data and Research Support

At
Eprints User Group, Open Repositories 2009

Hugh Glaser, Ian Millard & Les Carr
This is me
Background

• Rich resource of Repositories
• Other rich resources (NSF, CORDIS, ResearchIndex, Wikipedia)
• Where is the added value?
Linked Data

- Designed for just this
- “the Semantic Web done right, and the Web done right”
  - Tim Berners-Lee

- An Infrastructure for Naming things and Linking things
- OR can/should join
RKB and RKBExplorer

- RKB (ReSIST Knowledge Base) and RKBExplorer
- Knowledge-enabled infrastructure for cooperation in research into resilient systems
- Came out of CS AKTiveSpace
  - (Semantic Web Challenge winner 2003)
- Reasonably mature system and ongoing development
LOD Datasets on the Web: March 2009

4.5 billion triples
180 million data links
Harvest some ePrints repositories

This is one of several semantic repositories that contains and publishes RDF and co-reference data, forming the underlying distributed storage model behind the RKB Explorer initiative.

The data presented here comes from downloading and processing a number of eprints3 archives.

The following is a possibly incomplete list:

- Lancaster E-Prints URL: http://eprints.lancs.ac.uk/
- LSE Research Online URL: http://eprints.lse.ac.uk/
- Electronics & Computer Science EPrints Service - University of Southampton (ECS EPrints Service) URL: http://eprints.ecs.soton.ac.uk/
- Universiteit Twente Repository URL: http://doc.utwente.nl/
- UCL Eprints (University College London Eprints) URL: http://eprints.ucl.ac.uk/
- ePrints® OUDIR URL: http://eprints.lib.okayama-u.ac.jp/
- OPUS (Online Publications Store) URL: http://opus.bath.ac.uk/
- Caltech Authors URL: http://authors.library.caltech.edu/
- KFUPM ePrints URL: http://eprints.kfupm.edu.sa/
- Queensland University of Technology ePrints Archive (QUT ePrints Archive) URL: http://eprints.qut.edu.au/
- E-LIS URL: http://eprints.rclis.org/
- Open Research Online (ORO) URL: http://oro.open.ac.uk/
- KAR (Kent Academic Repository) URL: http://kar.kent.ac.uk/
- Цифровий репозиторій Харківської національної академії міського господарства (KNAME Digital Repository) (XHAMT) URL: http://eprints.kname.kharkov.ua/
- Munich RePEc Personal Archive URL: http://mpra.ub.uni-muenchen.de/
- UTas ePrints (University of Tasmania Eprints Repository) URL: http://eprints.utas.edu.au/
- Strathprints (University of Strathclyde Institutional Repository) URL: http://strathprints.strath.ac.uk/
- Bournemouth University EPrints URL: http://eprints.bournemouth.ac.uk/
One of Many KBs, including other ePrints

<table>
<thead>
<tr>
<th>URL</th>
<th>URL</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”
Now Look at an Author
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
Who is Carl Lagoze?
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
  – Not even (reliably) within most repositories
Co-Reference Service (CRS)

- CRS Subsystem
  - Find co-references
  - Store them
  - Publish them
    - Essentially:
      - URI\textsubscript{i} -> \{ URI\textsubscript{1}, ..., URI\textsubscript{i}, ..., URI\textsubscript{n} \}
    - Recommend a “Canon”
- Published by the Data Publisher
  - And possibly others
- Middleware aggregates co-references from recognised CRSeas
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/di/person/047157
4. http://dblp.rkbexplorer.com/id/people-1ed5a6029b22e0a6374e695e55214f05-90c423eb148125a8e6557345c1543c

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

Seungwoo Lee

Showing information queried from all repositories ...

Subject | Property | Object/Value | Source
---|---|---|---
Seungwoo Lee | akt:has-affiliation | Electrical and Computer Engineering Division, Pohang University of Science & Technology (POSTECH), Pohang, South Korea. gheelee@postech.ac.kr | acm-periodicals rdf >>
Seungwoo Lee | cts:isNameOf| POSTECH, Pohang, Korea | acm-proceedings rdf >>
Seungwoo Lee | kts:isLangOf| Person | datatypedproperties rdf >>
Seungwoo Lee | rdf:type | act:Affiliated-Person | acm-periodicals rdf >>
Seungwoo Lee | rdf:type | act:Generic-Agent | acm-periodicals rdf >>
Seungwoo Lee | rdf:type | act:Generic-Agent | dblp-publications-2001 rdf >>
Seungwoo Lee | rdf:type | act:Person | acm-periodicals rdf >>
Seungwoo Lee | rdf:type | act:Person | acm-proceedings rdf >>
Seungwoo Lee | rdf:type | act:Person | dblp-publications-2001 rdf >>
Seungwoo Lee | rdf:type | PER_char201MM | datatypedproperties rdf >>
Seungwoo Lee | rdf:type | PER_char201MM | objectsproperties rdf >>
Seungwoo Lee | rdf:type | PER_char201MM | resources rdf >>

Subject | Property | Object/Value | Source
---|---|---|---
Automatic acquisition of named entity tagged corpus from world wide web | akt:has-author | Seungwoo Lee | acm-proceedings rdf >>
A Corpus-Based Learning Method of Compound Noun Indexing Rules for Korean | akt:has-author | Seungwoo Lee | acm-periodicals rdf >>
SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP. | akt:has-author | Seungwoo Lee | dblp-publications-2001 rdf >>
A Corpus-Based Learning Method of Compound Noun | akt:has-author | Seungwoo Lee | dblp-publications-2001 rdf >>
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
Gadget – find out about people

Mark Borkum did this
Before and after inserting this paper in the Southampton ePrints repository and RKB has noticed
Note the position of Les Carr in Hugh’s related People
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?