Searching on the Open Semantic Web using a URI Identity Management Approach

Afraz Jaffri {a.o.jaffri@ecs.soton.ac.uk}
Dependable Systems and Software Engineering group
Electronics and Computer Science
University of Southampton

Aim of Research

In order for Semantic Web search engines and other applications to work with the increasing amount of RDF data that is being made available on the Web, there needs to be a URI management system that will track URI usage and *coreference* between URI's. The system will:

- Detect and group together URI's referring to the same resource
- Integrate with Linked Data
- Provide a query mechanism so groups of URI's can be quickly discovered by Semantic Web agents
- Track the Provenance of URI's
- Provide search functionality through the use of enhanced keywords mapped to URI's

Why do we need URI Identity Management?

URI's for 'Hugh Glaser':
http://acm.rkbexplorer.com/rdf/resource-P112732
http://citeseer.rkbexplorer.com/rdf/resource-CSP109020
http://citeseer.rkbexplorer.com/rdf/resource-CSP109011
http://dblp.rkbexplorer.com/rdf/resource-27de9959
http://www.ecs.soton.ac.uk/info/#person-00021

URI's for 'Spain':
http://dbpedia.org/resource/Spain
http://www4.wiwiss.fu-berlin.de/factbook/resource/Spain
http://sws.geonames.org/2510769
http://www.4.wiwiss.fu-berlin.de/eurostat/resource/countries/Espa%C3%Blla

What is the problem with owl:sameAs?

```
<rdf:Description  rdf:about="<URI-1>">
<vcard:FN>Hugh Glaser</vcard:FN>
<vcard:EMAIL>hg@ecs.soton.ac.uk</vcard:EMAIL>
<vcard:ROLE>Reader</vcard:ROLE></rdf>
```

```
<rdf:Description  rdf:about="<URI-2>">
<vcard:FN>Hugh Glaser</vcard:FN>
<vcard:EMAIL>hg1@soton.ac.uk</vcard:EMAIL>
<vcard:ROLE>Lecturer</vcard:ROLE></rdf>
```

Assert <URI-1> <owl:sameAs> <URI-2>

SELECT ?x WHERE {<URI-1> <vcard:EMAIL> ?x} Returns hg1@soton.ac.uk hg@ecs.soton.ac.uk

Which email belongs to which role?

Using owl:sameAs means that both URI's become indistinguishable even though they may refer to different entities according to the context in which they are used.

Solution

Consistent Reference Services

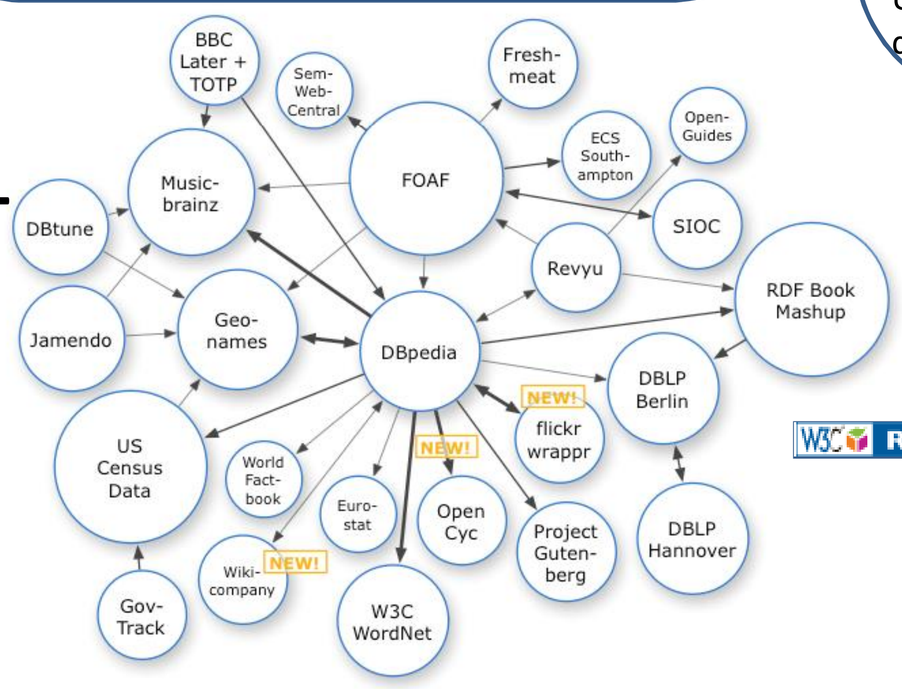
URI'S referring to the same resource are grouped together in '*Bundles*'
A bundle has properties:

Coref:hasCanonicalReference – One URI in a bundle can be made to be the canonical representation i.e. The preferred URI

Coref:hasEquivalentReference – The URI's in a bundle are grouped together using this predicate

Coref:updatedOn – The date of the last update to the bundle

2 Billion Triples

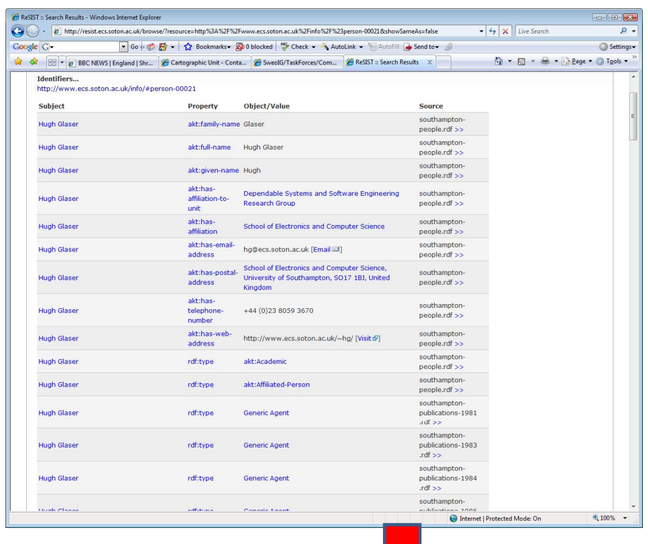


RDF

Knowledge Mediator

Finds all possible equivalences for a URI using the algorithm:

```
findEquivalence(URI u) {
    Dereference u;
    while (u coref:hasEquivalentReference a) {
        add a to equivalences;
        findEquivalence(a);
        }
}
```



Results (Knowledge)

User

Keywords

Knowledge Manager

Matches keywords to URI's using rdfs:label and similar properties

RDF

acm
CORDIS
Proceedings of the IEEE
CiteSeer.IST Scientific Literature Digital Library
NSF DBLP
THE RISKS DIGEST

60 Million Triples

Evaluation Method

- Use precision and recall metrics on URI's of known resources in order to determine coverage of URI coreference
- Perform usability testing on input and output interfaces
- Compare results between linkage with owl:sameAs and linkage with CRS's
- Formalise a theory of URI linkage

This work is supported under the ReSIST Network of Excellence (NoE) which is sponsored by the Information Society Technology (IST) priority of the EU Sixth Framework programme (FP6) under contract number IST-4-026764-NOE