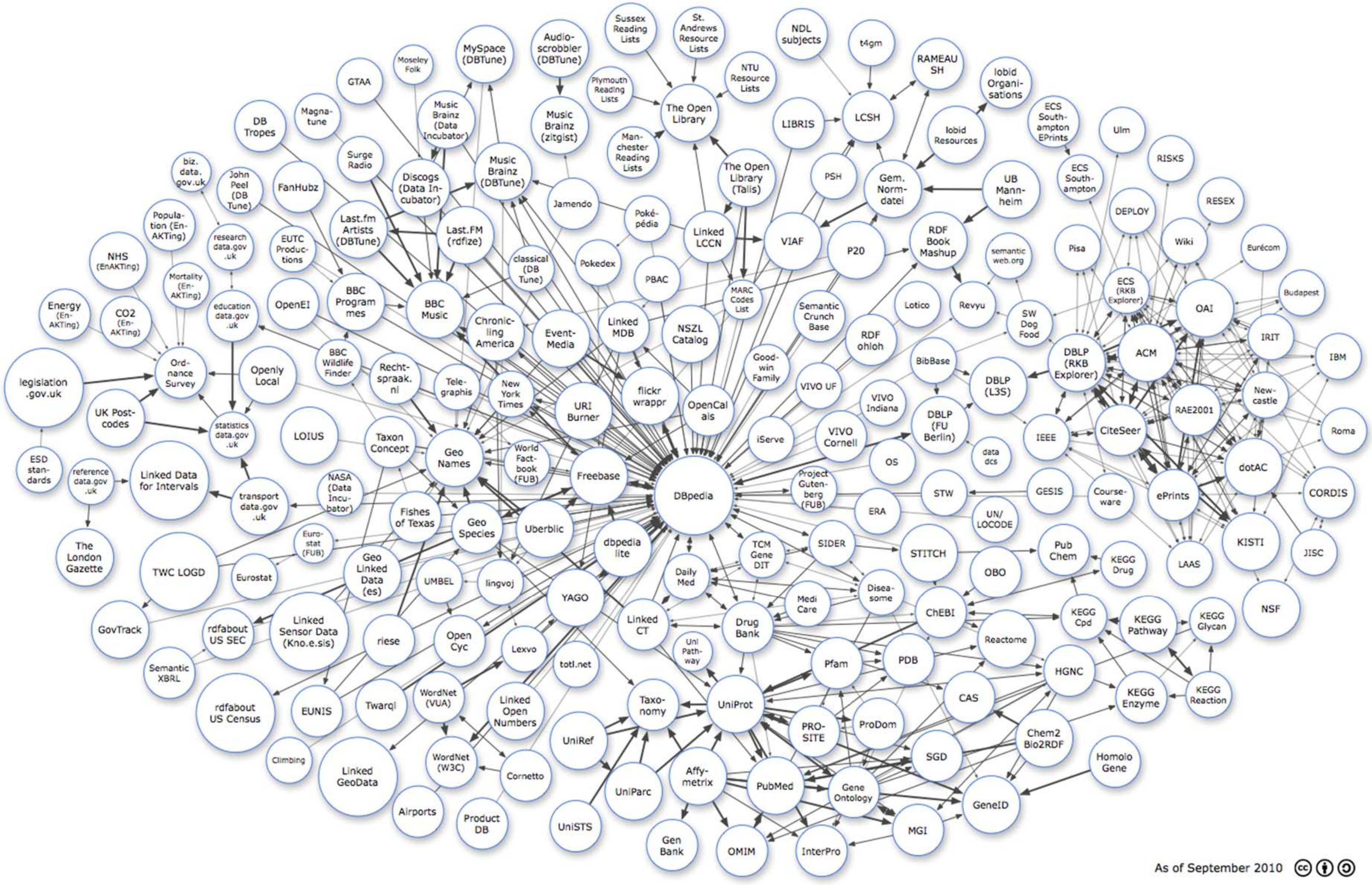
Re-using Cool URIs:
Entity Reconciliation Against LOD Hubs

Fadi Maali, Richard Cyganiak, Vassilios Peristeras

LDOW 2011

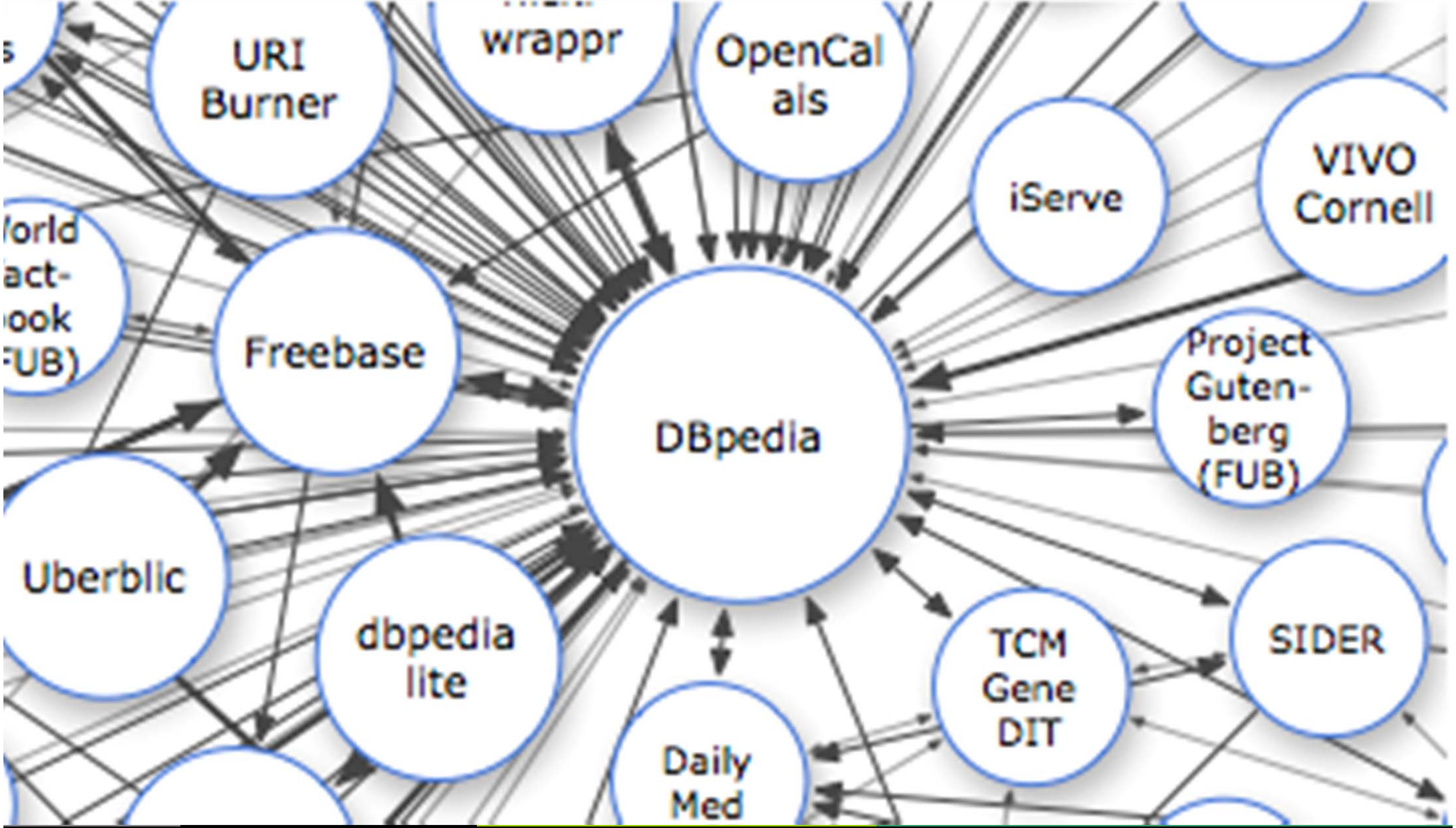
© Copyright 2009 Digital Enterprise Research Institute. All rights reserved.

The Web of Data

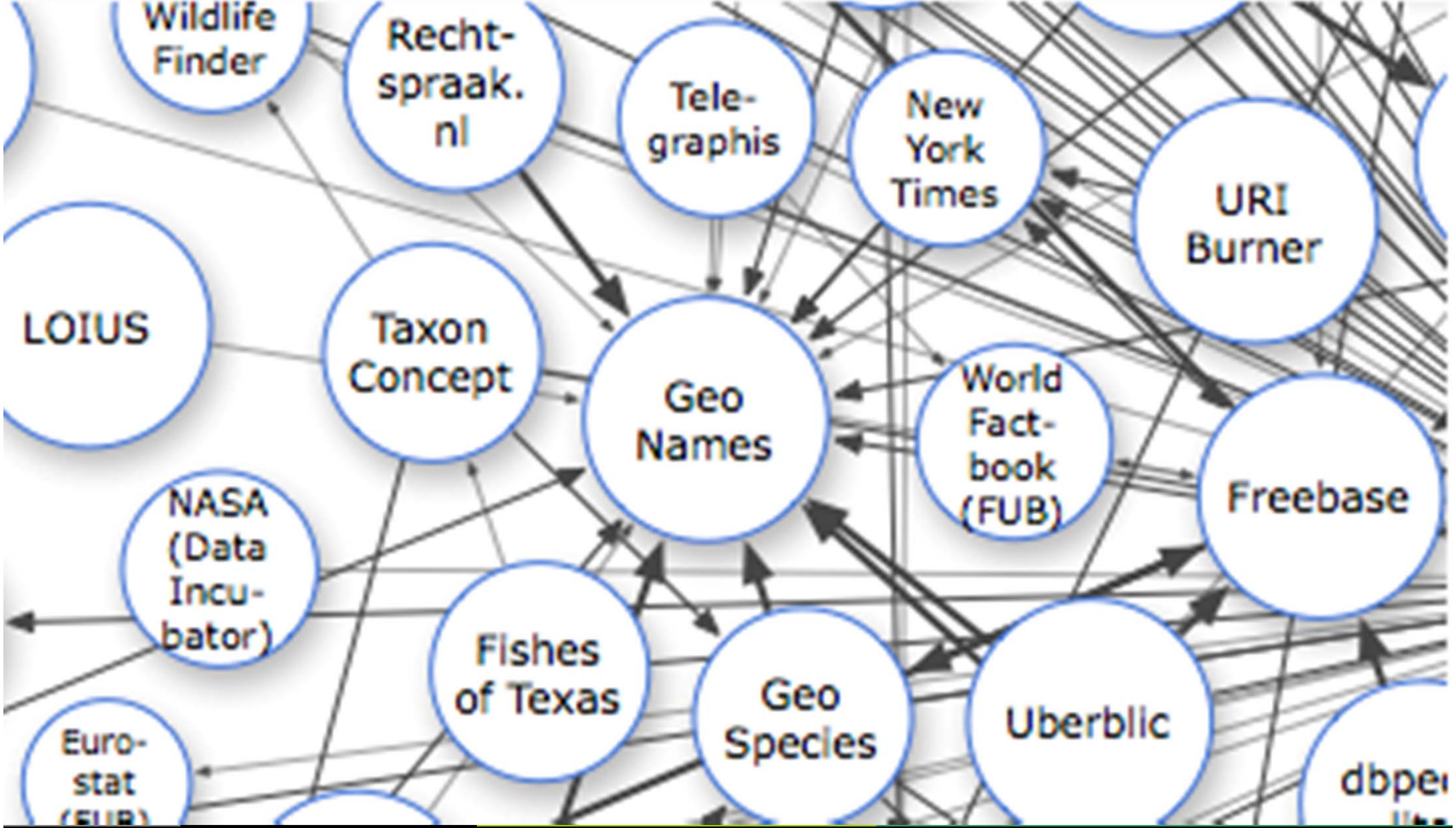


As of September 2010

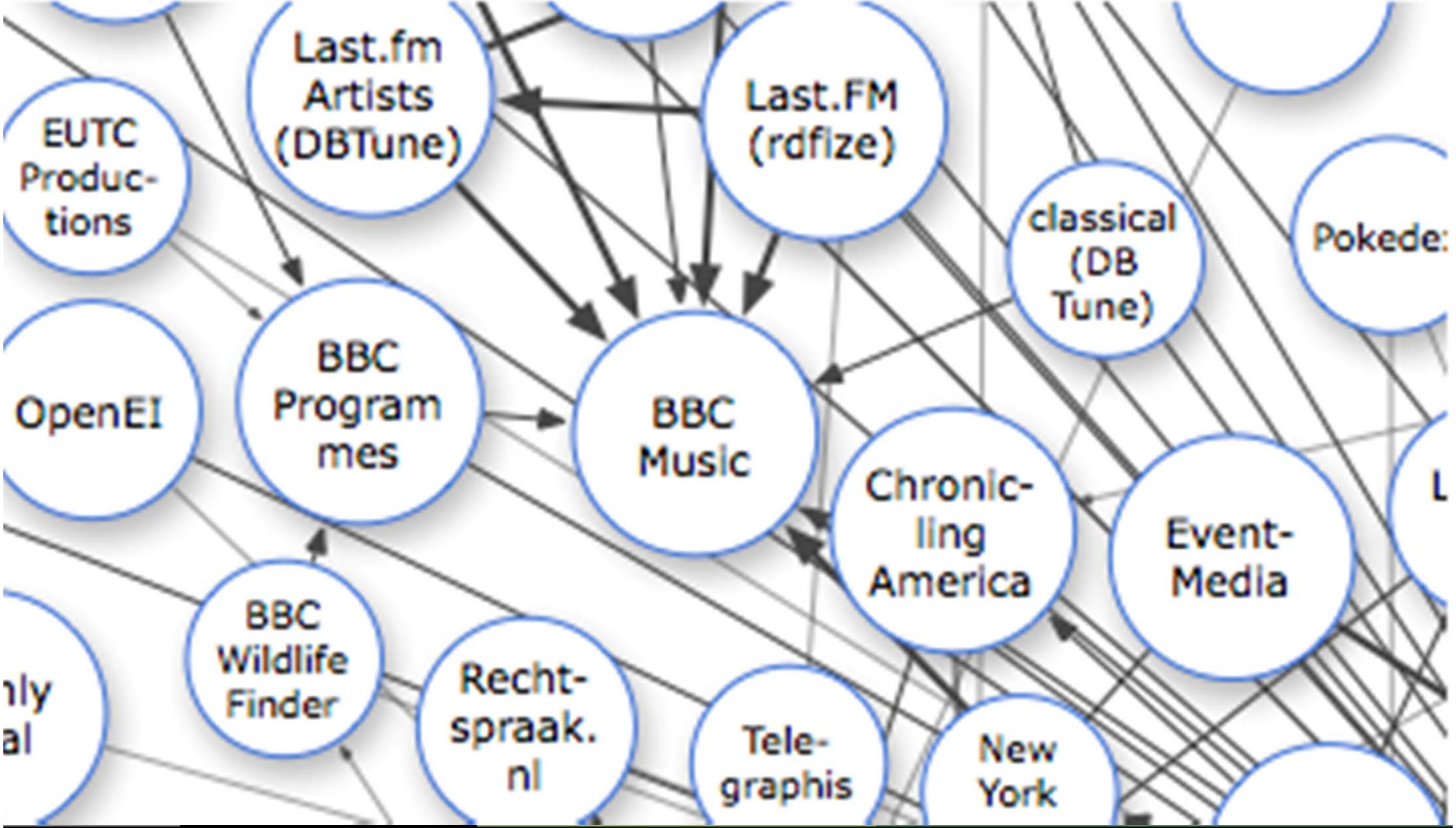
The Web of Data



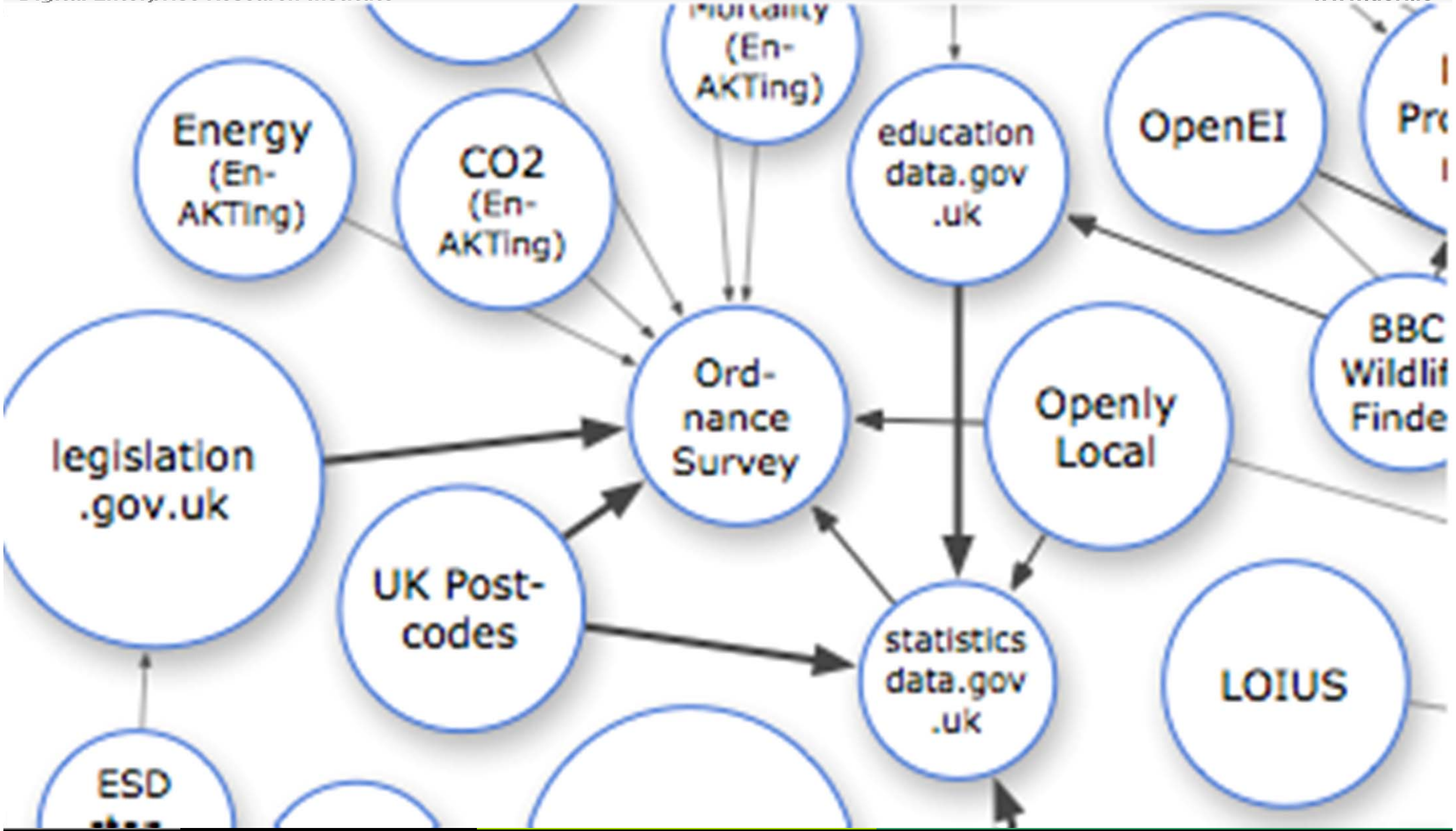
The Web of Data



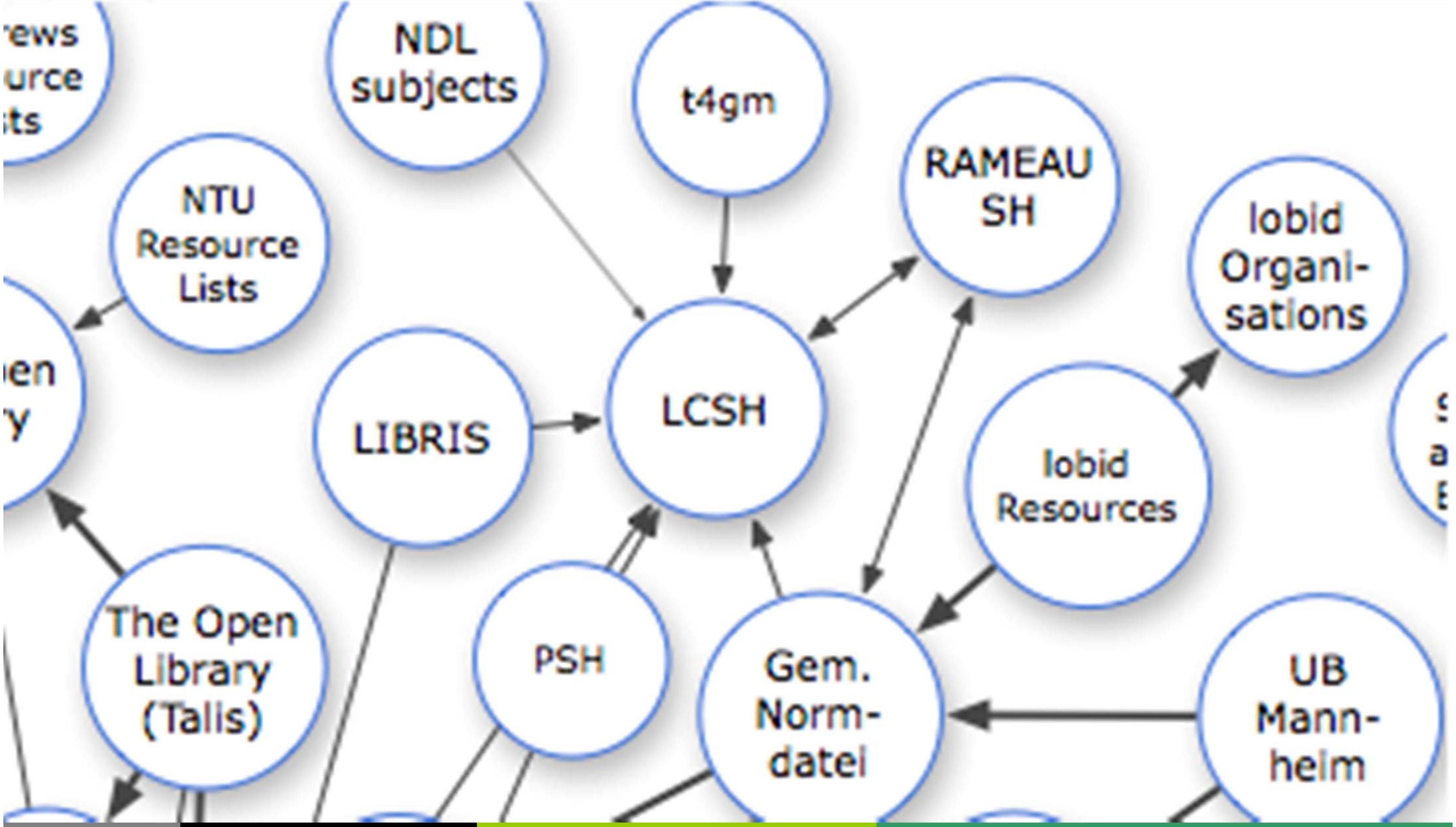
The Web of Data



The Web of Data



The Web of Data

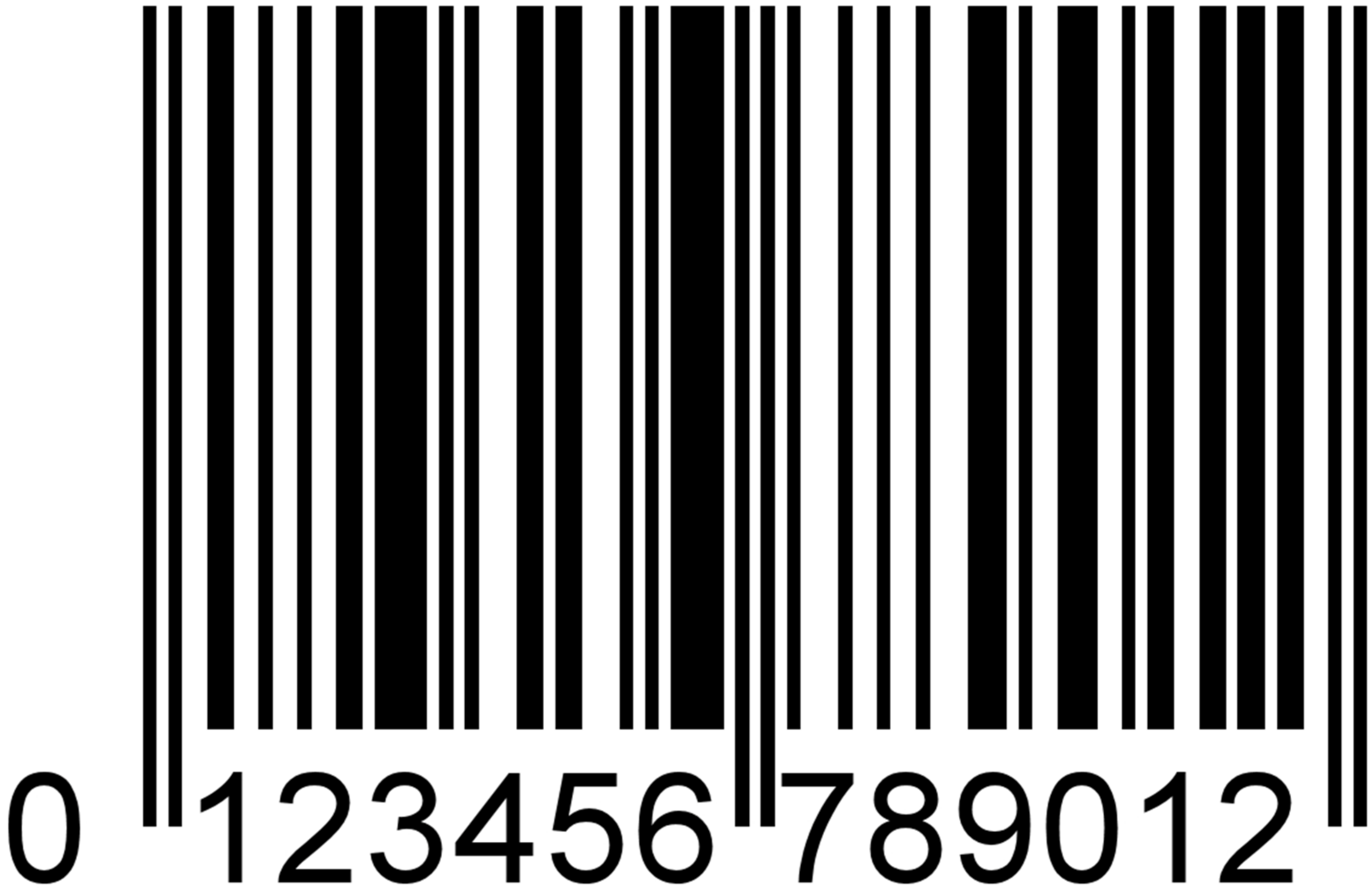


“LOD Hubs”

- LOD Hubs = datasets that attract many inlinks

- Music metadata community uses BBC Music identifiers
- UK government data community uses Ordnance Survey identifiers
- Library data community uses Library of Congress Subject Headings

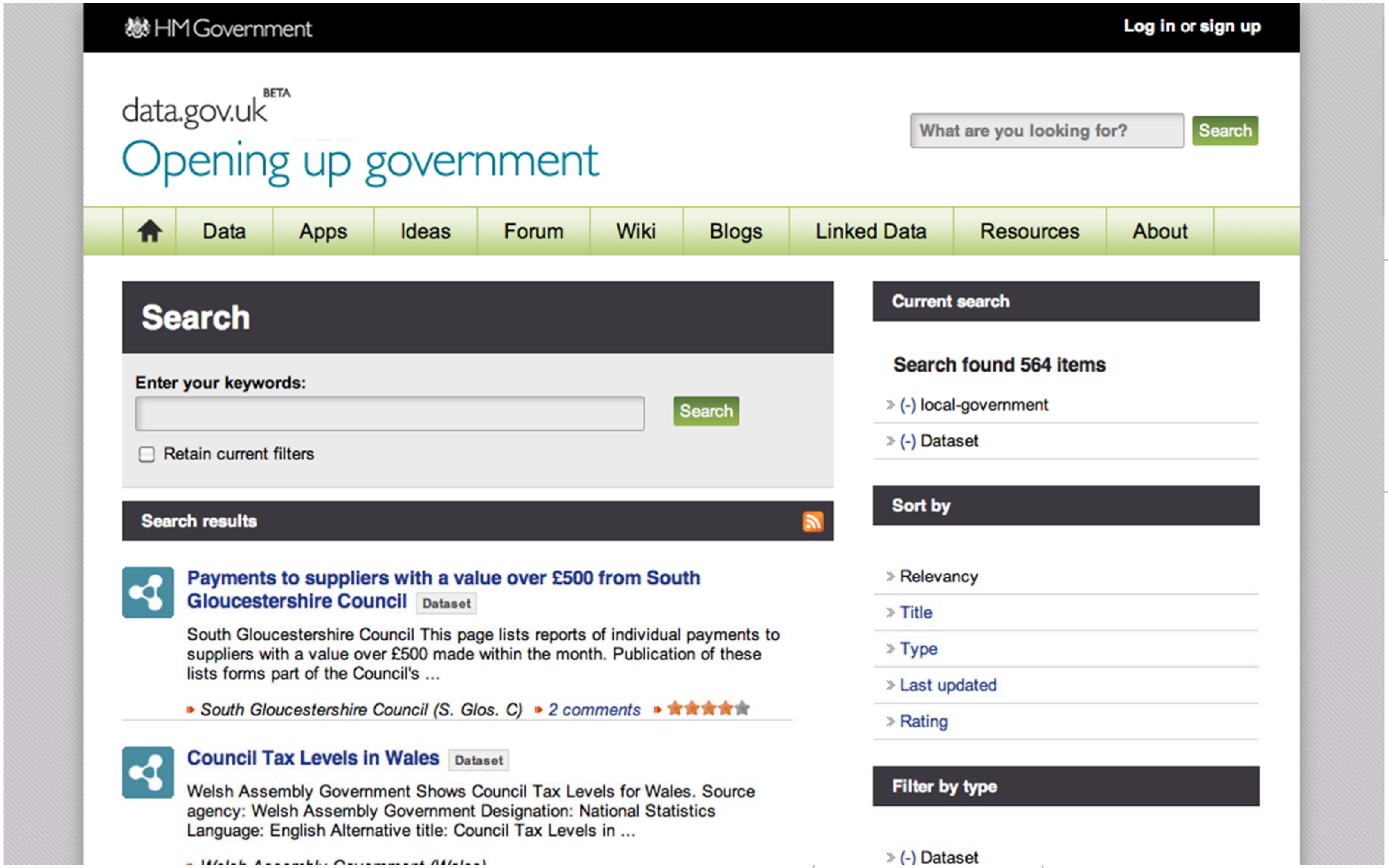
Standard identifiers



Standard identifiers

LOD Hubs are bar codes for a specific community.

For example, government data



data.gov.uk BETA

Opening up government

Search

Enter your keywords:

Search

□ Retain current filters

Search results

Payments to suppliers with a value over £500 from South Gloucestershire Council Dataset

South Gloucestershire Council This page lists reports of individual payments to suppliers with a value over £500 made within the month. Publication of these lists forms part of the Council's ...

South Gloucestershire Council (S. Glos. C) · *2 comments* · ★★★★☆

Council Tax Levels in Wales Dataset

Welsh Assembly Government Shows Council Tax Levels for Wales. Source agency: Welsh Assembly Government Designation: National Statistics Language: English Alternative title: Council Tax Levels in ...

Welsh Assembly Government (Wales)

Current search

Search found 564 items

- (-) local-government
- (-) Dataset

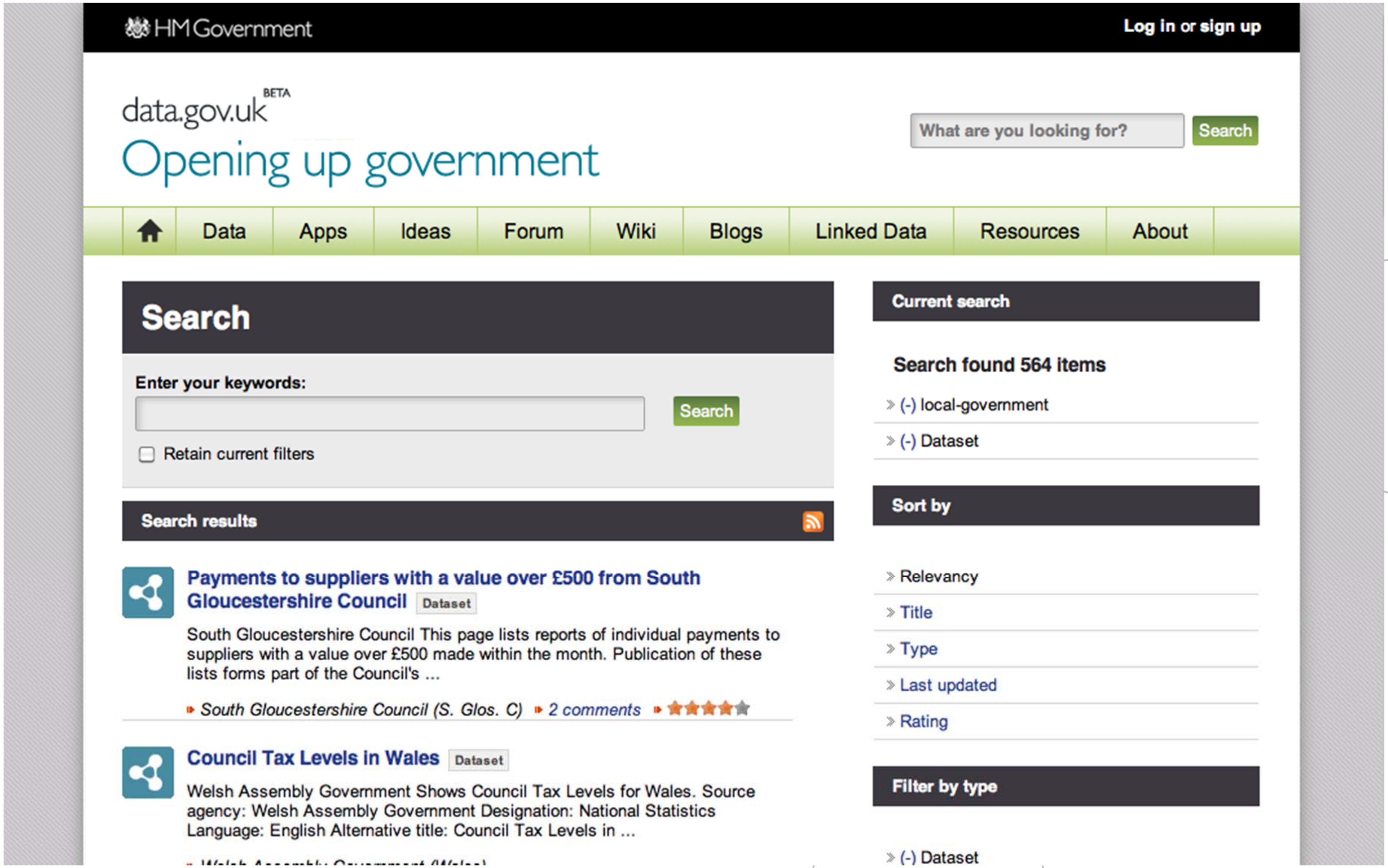
Sort by

- Relevancy
- Title
- Type
- Last updated
- Rating

Filter by type

- (-) Dataset

For example, government data



data.gov.uk BETA

Opening up government

Search

Enter your keywords:

Search

□ Retain current filters

Search results

Payments to suppliers with a value over £500 from South Gloucestershire Council Dataset

South Gloucestershire Council This page lists reports of individual payments to suppliers with a value over £500 made within the month. Publication of these lists forms part of the Council's ...

South Gloucestershire Council (S. Glos. C) *2 comments* ★★★★☆

Council Tax Levels in Wales Dataset

Welsh Assembly Government Shows Council Tax Levels for Wales. Source agency: Welsh Assembly Government Designation: National Statistics Language: English Alternative title: Council Tax Levels in ...

Welsh Assembly Government (Wales)

Current search

Search found 564 items

- (-) local-government
- (-) Dataset

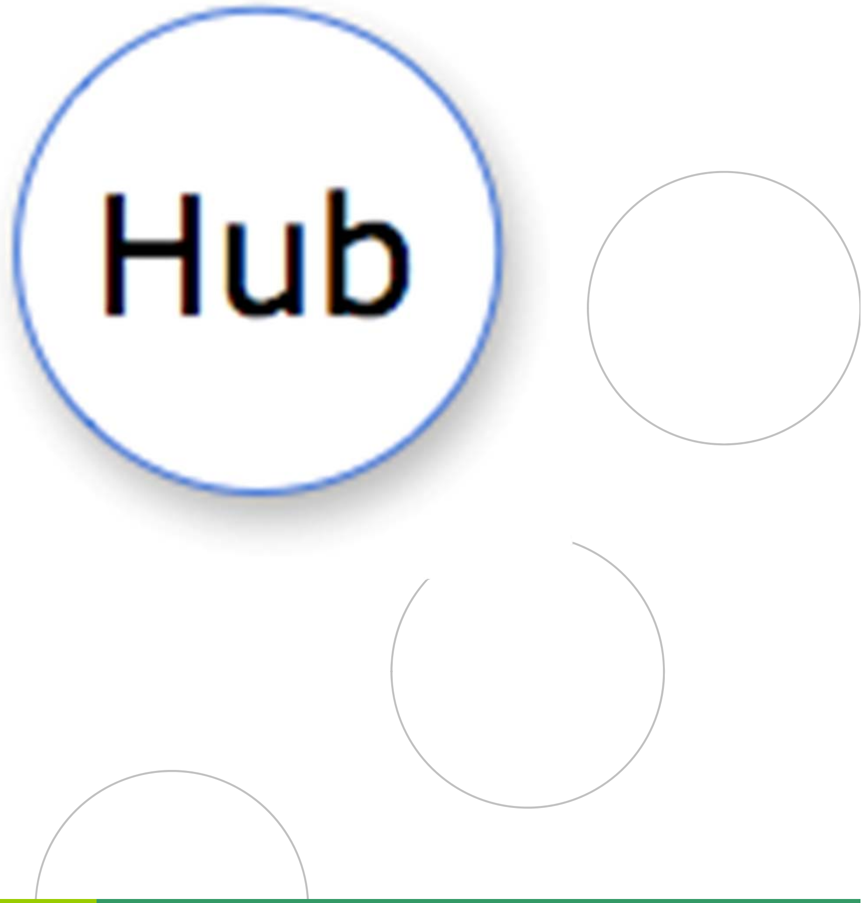
Sort by

- Relevancy
- Title
- Type
- Last updated
- Rating

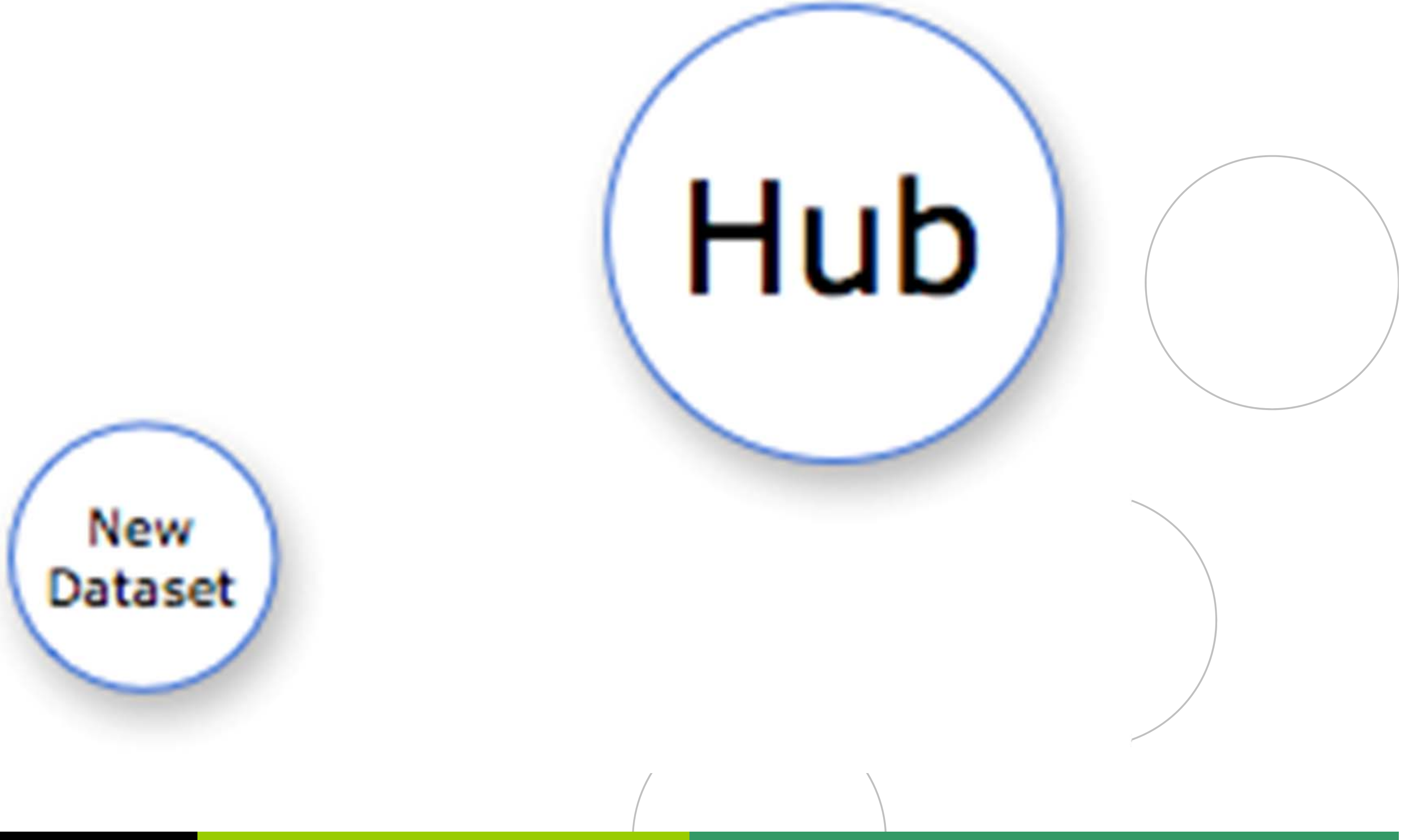
Filter by type

- (-) Dataset

How to Attract Links



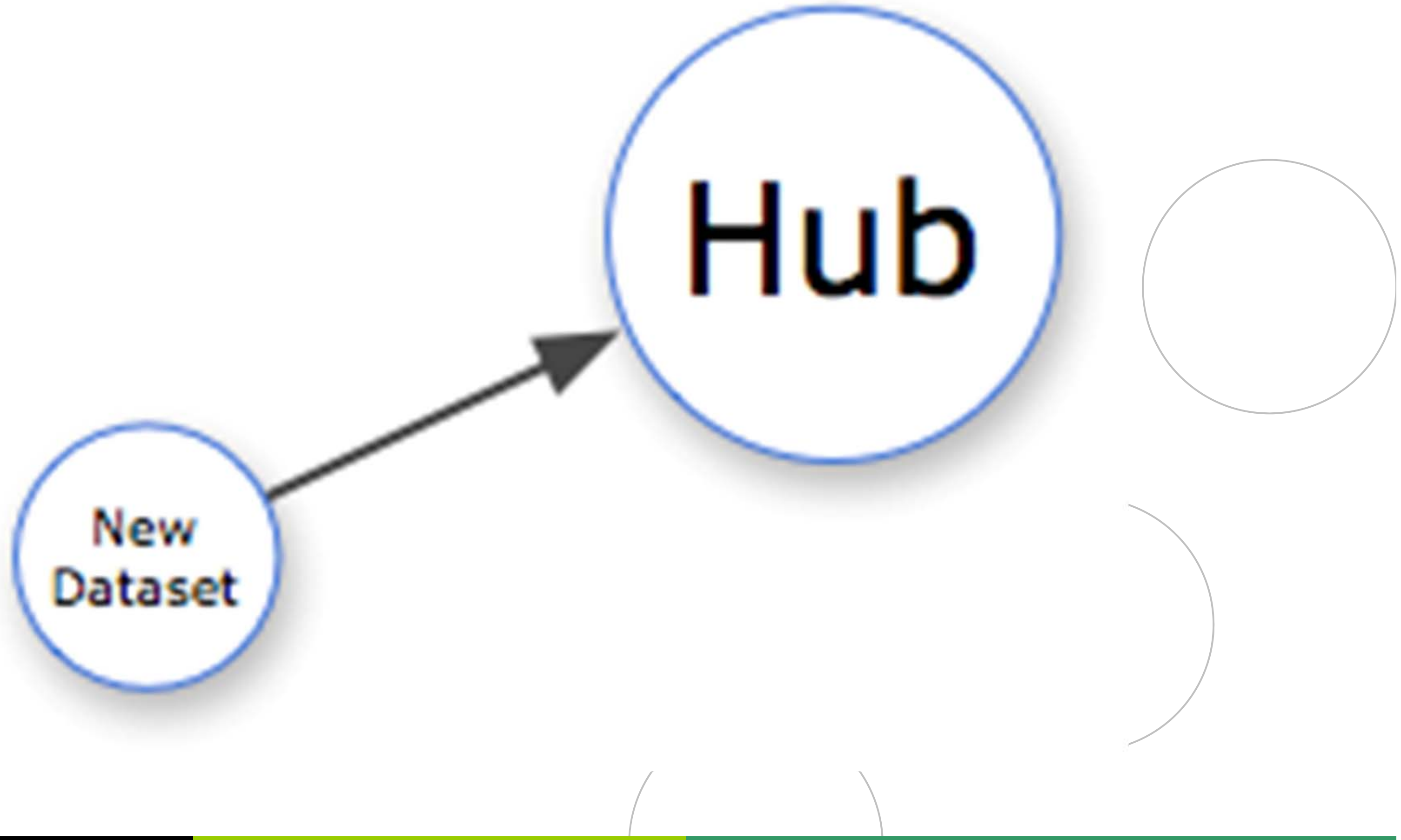
How to Attract Links



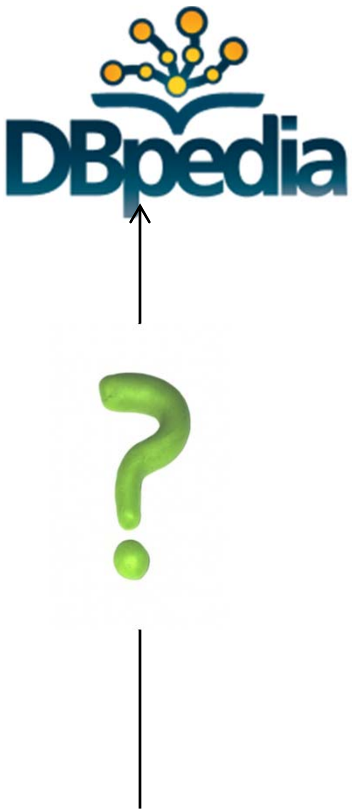
Hub

New Dataset

How to Attract Links



Reconciliation



City	State	Country
Cambridge	Massachusetts	United States

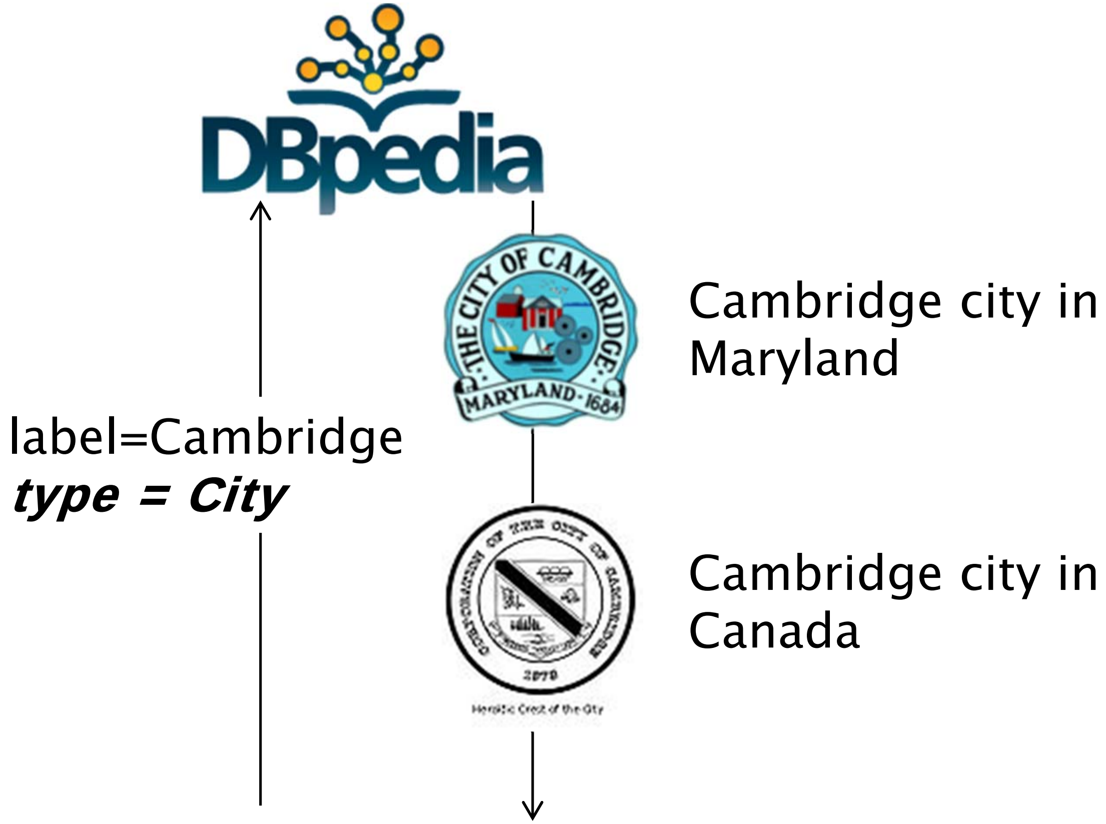
Reconciliation

label=Cambridge

Cambridge Bay in Canada

City	State	Country
Cambridge	Massachusetts	United States

Reconciliation


DBpedia
THE CITY OF CAMBRIDGE MARYLAND 1684
Heraldic Crest of the City

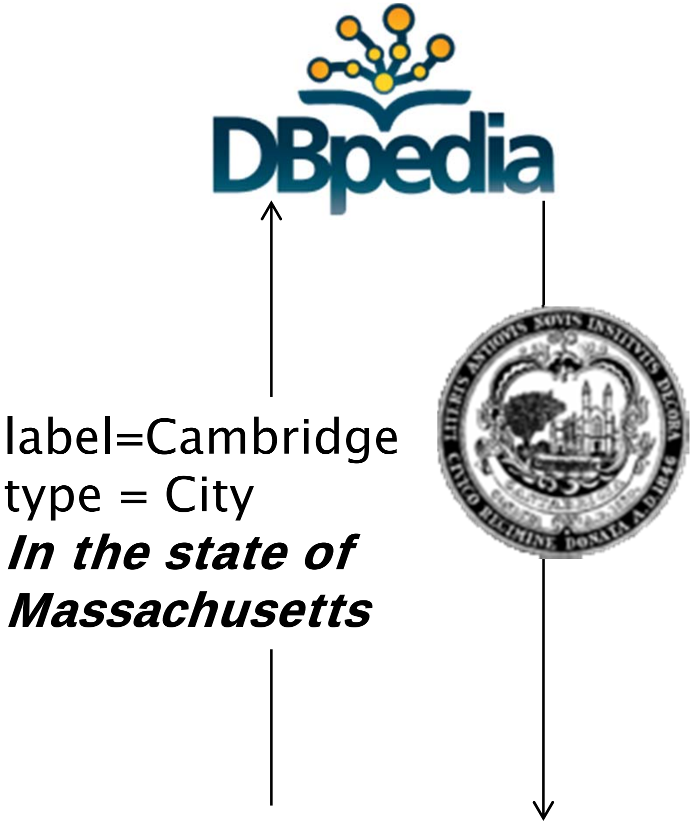
label=Cambridge
type = City

Cambridge city in Maryland

Cambridge city in Canada

City	State	Country
Cambridge	Massachusetts	United States

Reconciliation



label=Cambridge
type = City
In the state of Massachusetts

Cambridge city in Massachusetts ✓

City	State	Country
Cambridge	Massachusetts	United States

Approaches

- SPARQL
- SPARQL + full-text search
- Silk Server
- Semantic Web search engines

SPARQL

- Based on regular expressions
- Pros
  - □ Standardised
  - □ Zero-effort approach
- Cons
  - □ Slow
  - □ Not good at text search
  - □ No ranked results

SPARQL + full-text search

- Based on full-text extension for SPARQL
- Pros
  - More forgiving string matching
  - Ranking
  - Zero-effort (depending on your SPARQL store)
- Cons
  - Proprietary syntax

Silk Server

- Pros
  - Powerful declarative link specification
  - Variety of similarity functions
- Cons
  - Configuration needs to prepared
  - Silk Server needs to be deployed
  - Silk Server tightly couples its input and reference data

Semantic Web Search Engine

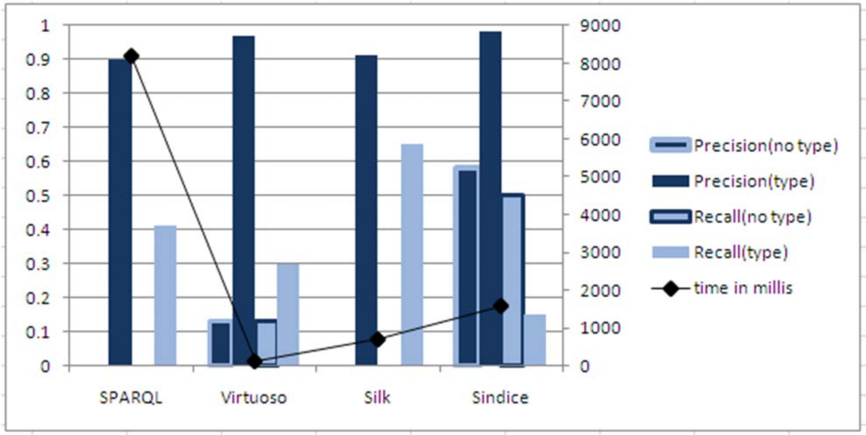
- Based on Sindice API
- Pros
  - Zero-effort approach (if your dataset is indexed in Sindice)
  - Search distributed RDF datasets (e.g. FOAF profiles)
- Cons
  - Noisy

Benchmark

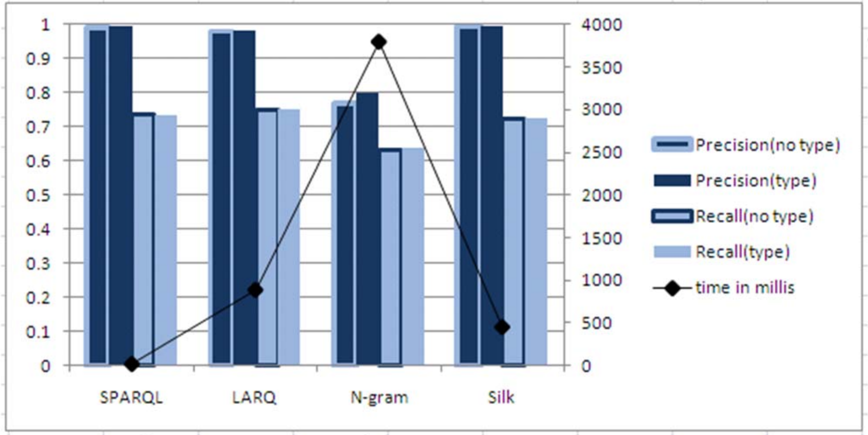
- Data Interlinking benchmark (part of IM@OAEI2010)
- We reconciled DailyMed against:
  - □ DBpedia SPARQL endpoint (http://dbpedia.org/sparql)
  - □ Sider dump file (part of the benchmark)

Results

- SPARQL with REGEX is unsuitable (performance)
- Except if labels are very consistent
- Type restrictions are very effective
- Silk has best recall (but requires custom link spec)



Services performance against DBpedia

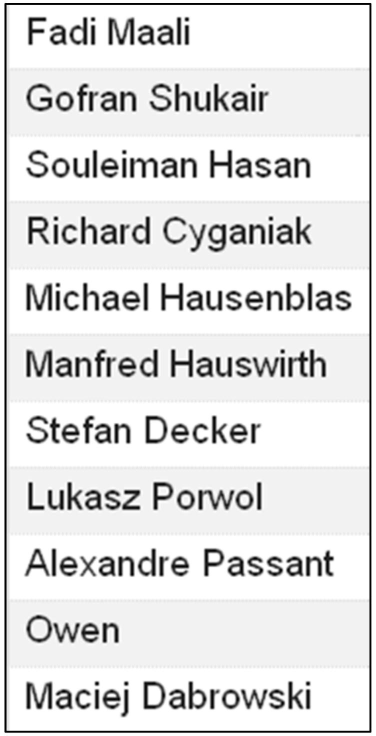


Services performance against Sider RDF dump file

Google Refine + RDF

+ RDF

Example



Fadi Maali

Gofran Shukair

Souleiman Hasan

Richard Cyganiak

Michael Hausenblas

Manfred Hauswirth

Stefan Decker

Lukasz Porwol

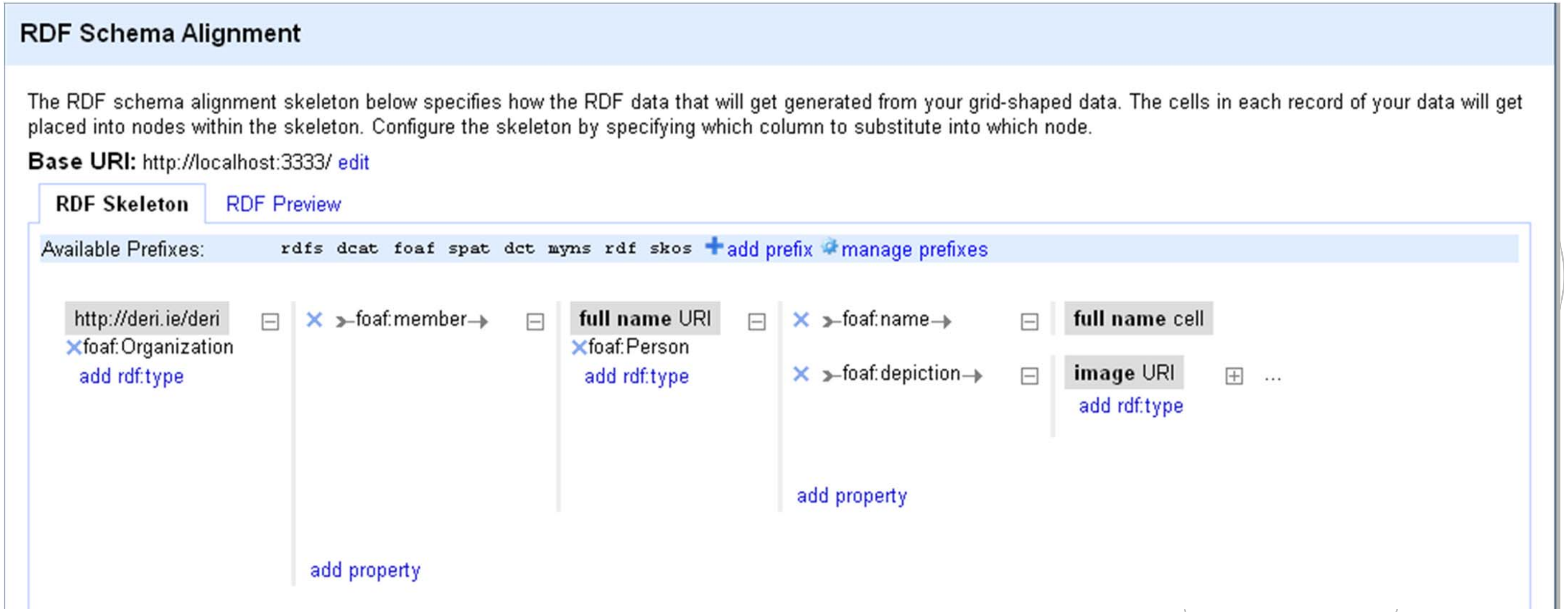
Alexandre Passant

Owen

Maciej Dabrowski

List of people from DERI

Example

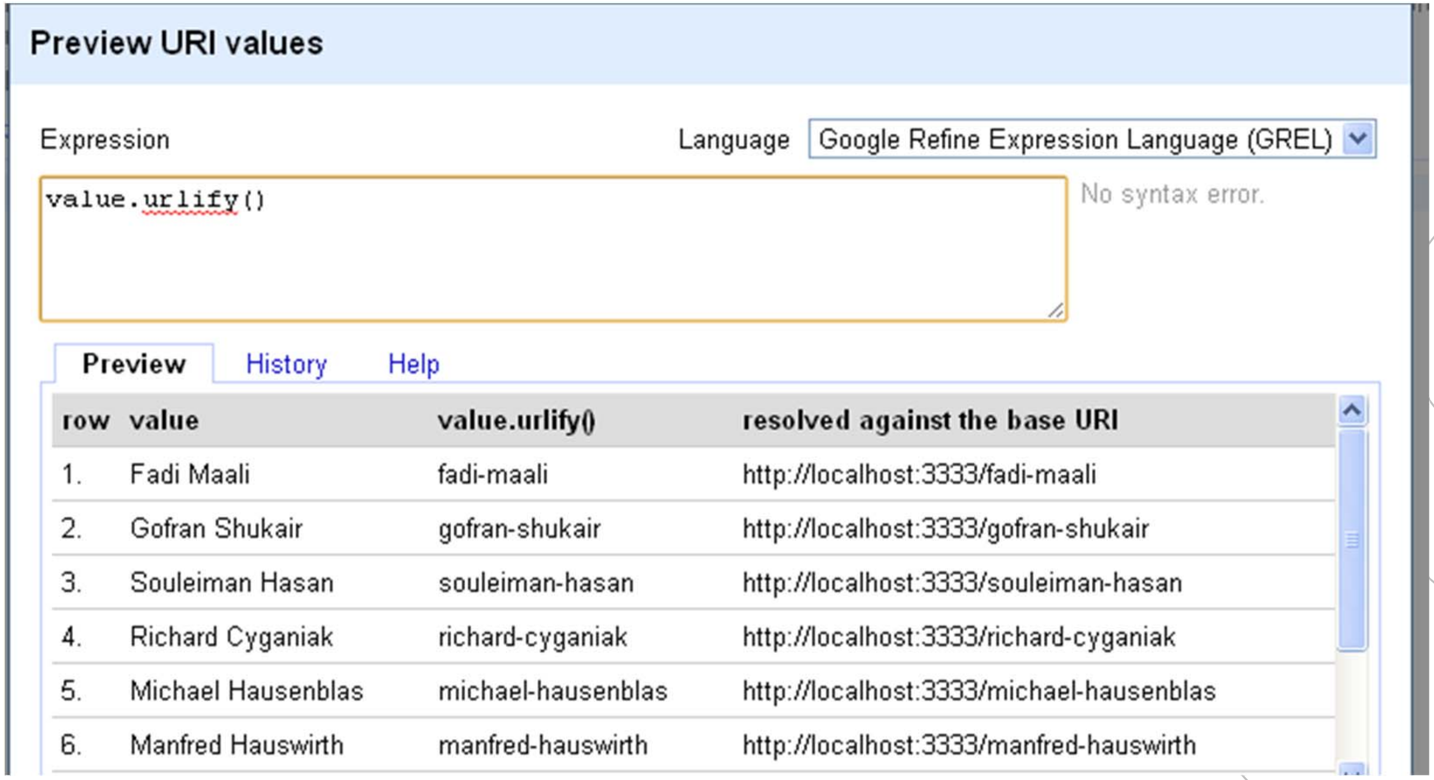


RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://localhost:3333/ edit

Example



Preview URI values

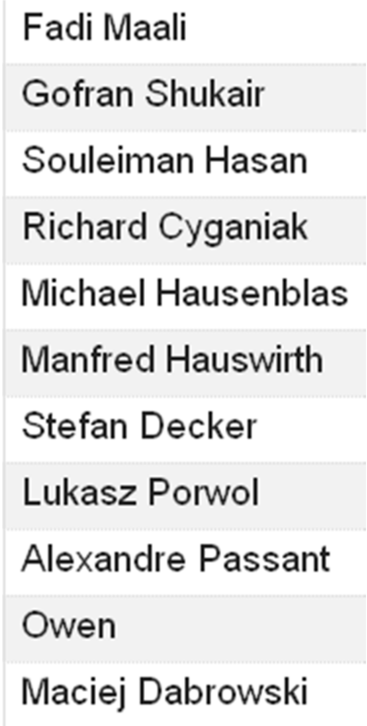
Expression: value.urlify()

Language: Google Refine Expression Language (GREL)

No syntax error.

row	value	value.urlify()	resolved against the base URI
1.	Fadi Maali	fadi-maali	http://localhost:3333/fadi-maali
2.	Gofran Shukair	gofran-shukair	http://localhost:3333/gofran-shukair
3.	Souleiman Hasan	souleiman-hasan	http://localhost:3333/souleiman-hasan
4.	Richard Cyganiak	richard-cyganiak	http://localhost:3333/richard-cyganiak
5.	Michael Hausenblas	michael-hausenblas	http://localhost:3333/michael-hausenblas
6.	Manfred Hauswirth	manfred-hauswirth	http://localhost:3333/manfred-hauswirth

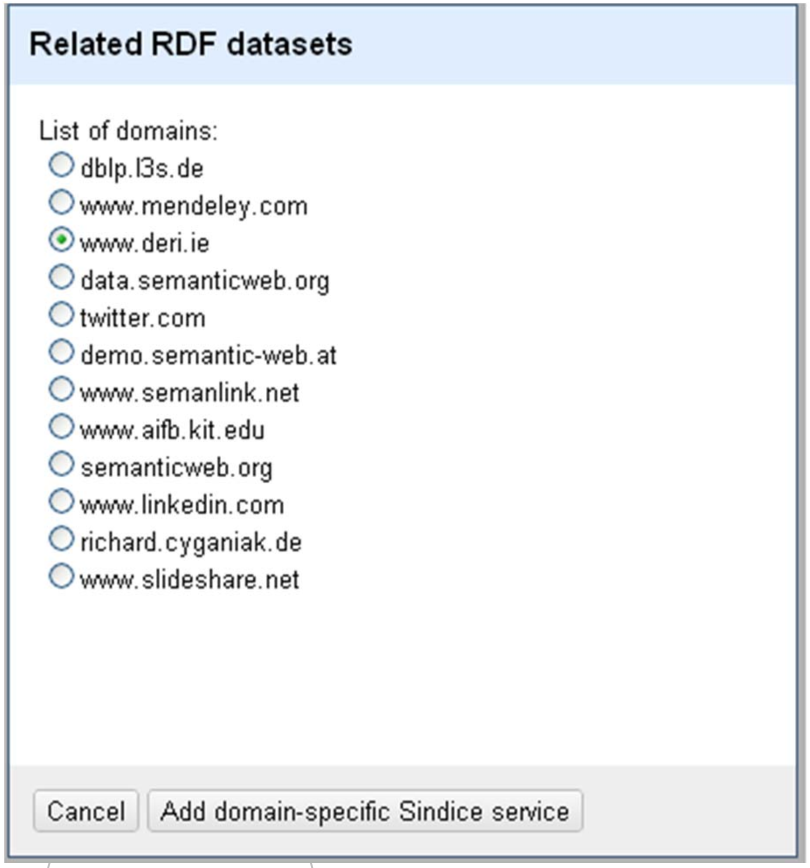
Example



Fadi Maali
Gofran Shukair
Souleiman Hasan
Richard Cyganiak
Michael Hausenblas
Manfred Hauswirth
Stefan Decker
Lukasz Porwol
Alexandre Passant
Owen
Maciej Dabrowski

List of people from DERI

Find related RDF datasets

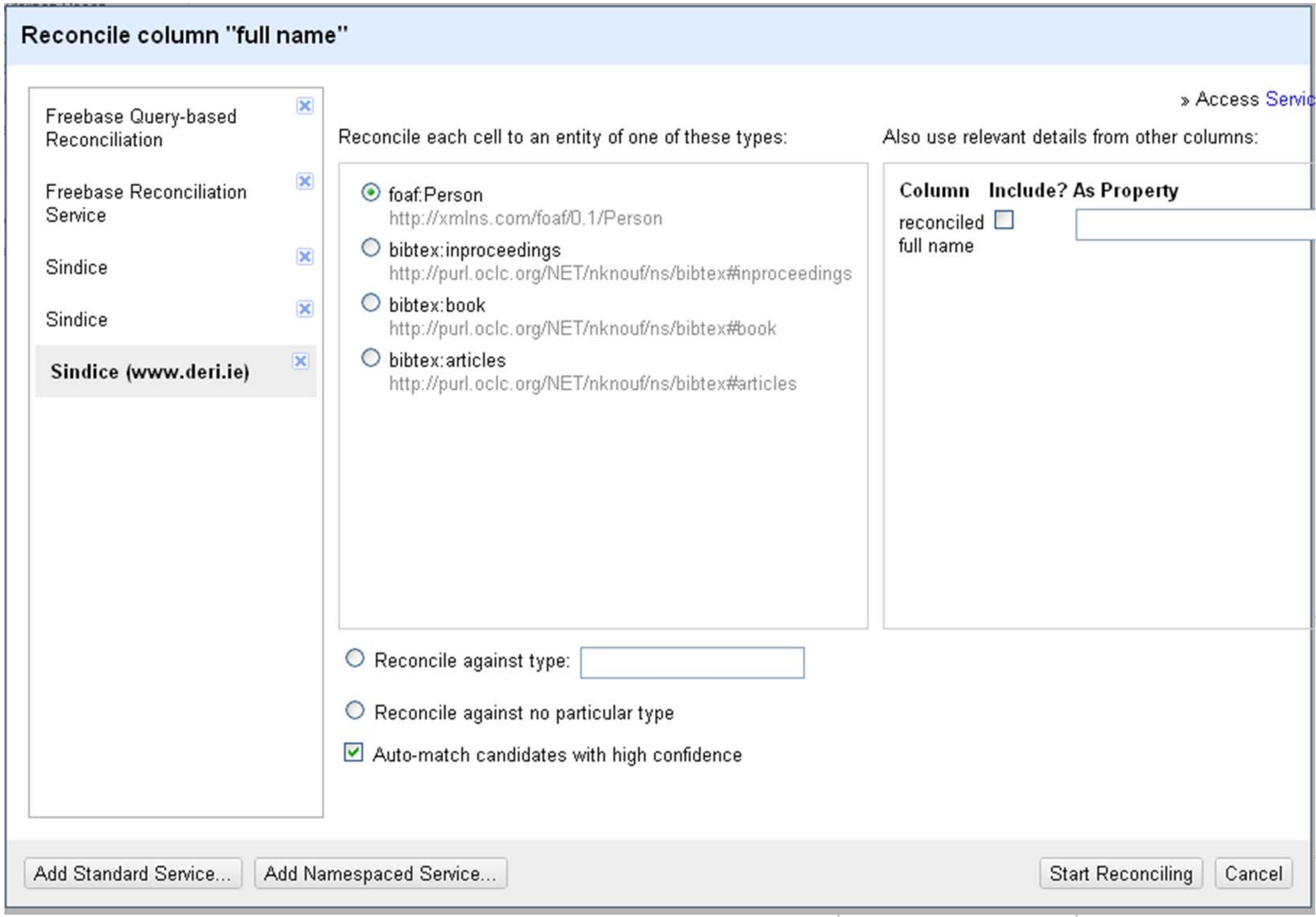


Related RDF datasets

List of domains:

- ○ dblp.l3s.de
- ○ www.mendeley.com
- ● www.deri.ie
- ○ data.semanticweb.org
- ○ twitter.com
- ○ demo.semantic-web.at
- ○ www.semanlink.net
- ○ www.aifb.kit.edu
- ○ semanticweb.org
- ○ www.linkedin.com
- ○ richard.cyganiak.de
- ○ www.slideshare.net

Example



Reconcile column "full name"

Freebase Query-based Reconciliation
Freebase Reconciliation Service
Sindice
Sindice
Sindice (www.deri.ie)
» Access Servic

Reconcile each cell to an entity of one of these types:

- ● foaf:Person http://xmlns.com/foaf/0.1/Person
- ○ bibtex:inproceedings http://purl.oclc.org/NET/nknouf/ns/bibtex#inproceedings
- ○ bibtex:book http://purl.oclc.org/NET/nknouf/ns/bibtex#book
- ○ bibtex:articles http://purl.oclc.org/NET/nknouf/ns/bibtex#articles

Also use relevant details from other columns:

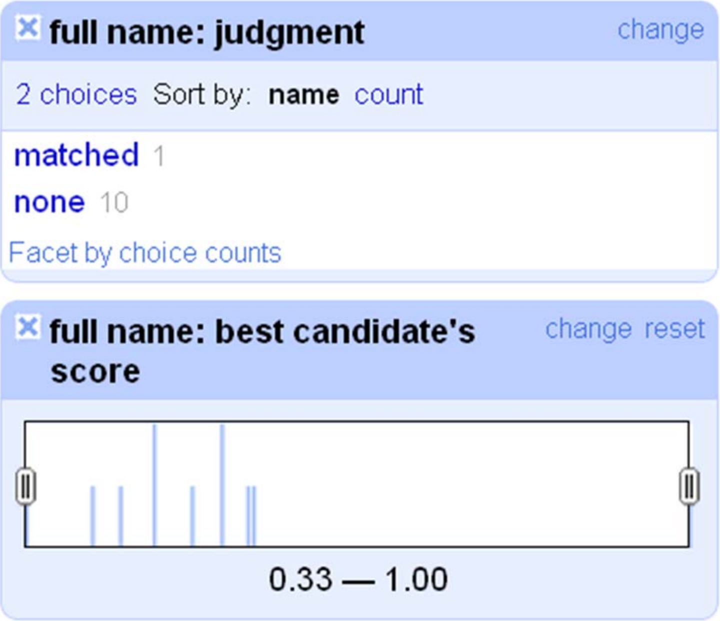
Column	Include?	As Property
reconciled full name	□	

○ Reconcile against type:

○ Reconcile against no particular type

☑ Auto-match candidates with high confidence

Example



full name: judgment change

2 choices Sort by: name count

matched 1

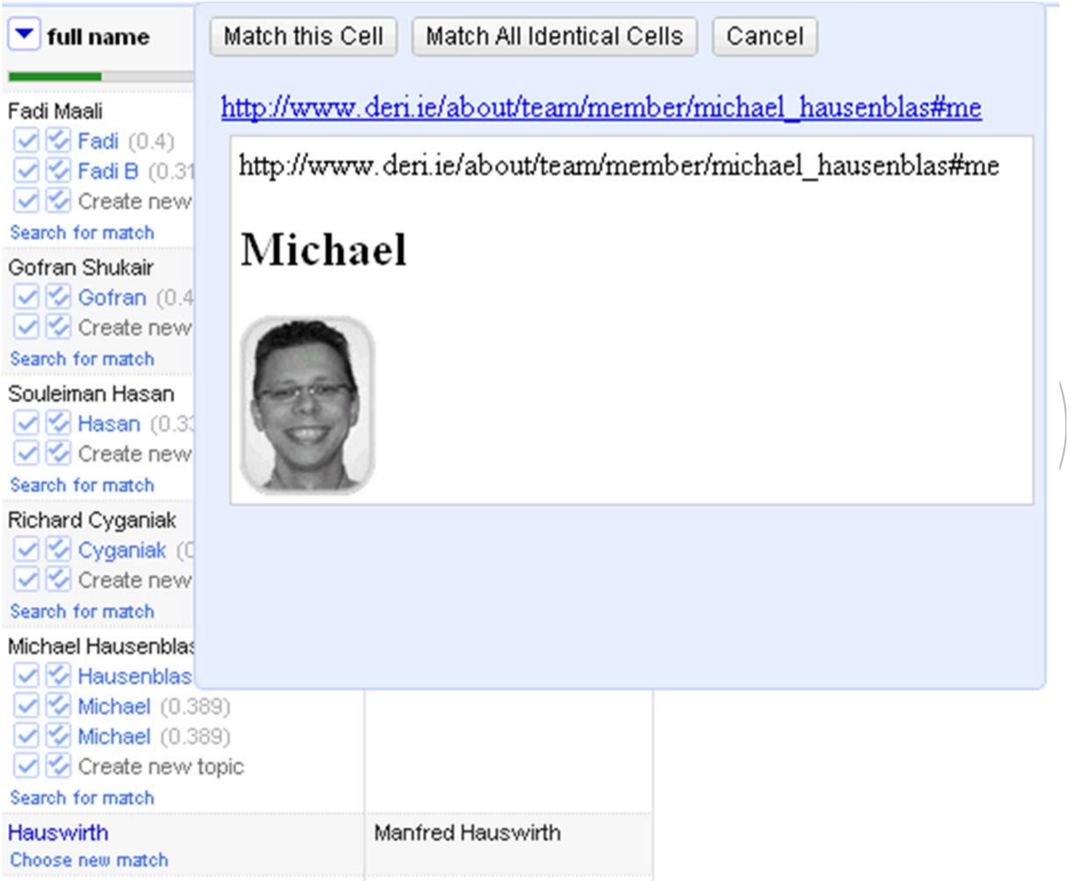
none 10

Facet by choice counts

full name: best candidate's score change reset

0.33 — 1.00

Reconciliation result facets



full name

Match this Cell | Match All Identical Cells | Cancel

http://www.deri.ie/about/team/member/michael_hausenblas#me

http://www.deri.ie/about/team/member/michael_hausenblas#me

Michael

Fadi Maali
- Fadi (0.4)
- Fadi B (0.31
- Create new
- Search for match

Gofran Shukair
- Gofran (0.4
- Create new
- Search for match

Souleiman Hasan
- Hasan (0.33
- Create new
- Search for match

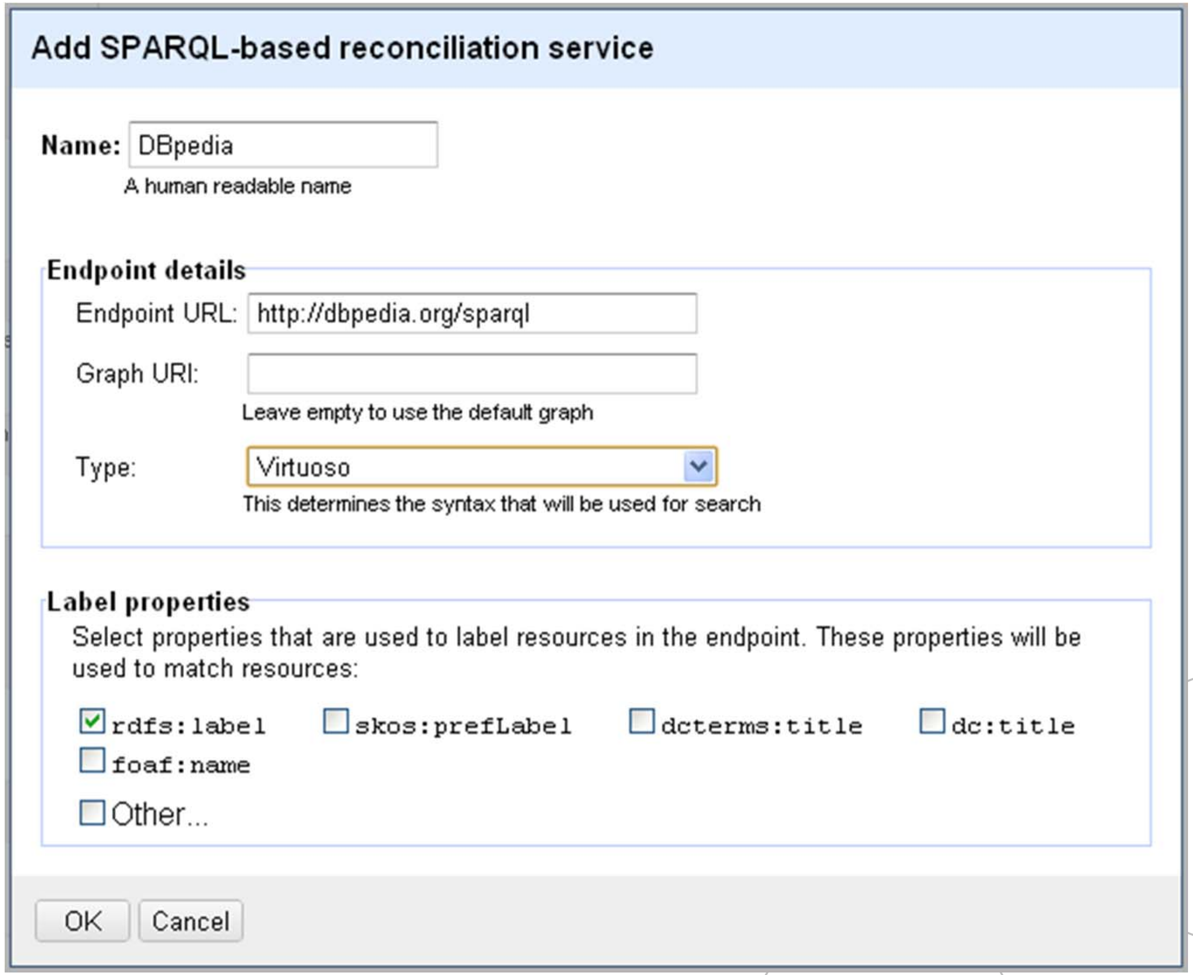
Richard Cyganiak
- Cyganiak (0
- Create new
- Search for match

Michael Hausenblas
- Hausenblas
- Michael (0.389)
- Michael (0.389)
- Create new topic
- Search for match

Hauswirth Manfred Hauswirth
- Choose new match

Resource Preview

Reconcile against a SPARQL endpoint



Add SPARQL-based reconciliation service

Name: DBpedia
A human readable name

Endpoint details

Endpoint URL: http://dbpedia.org/sparql

Graph URI:
Leave empty to use the default graph

Type: Virtuoso
This determines the syntax that will be used for search

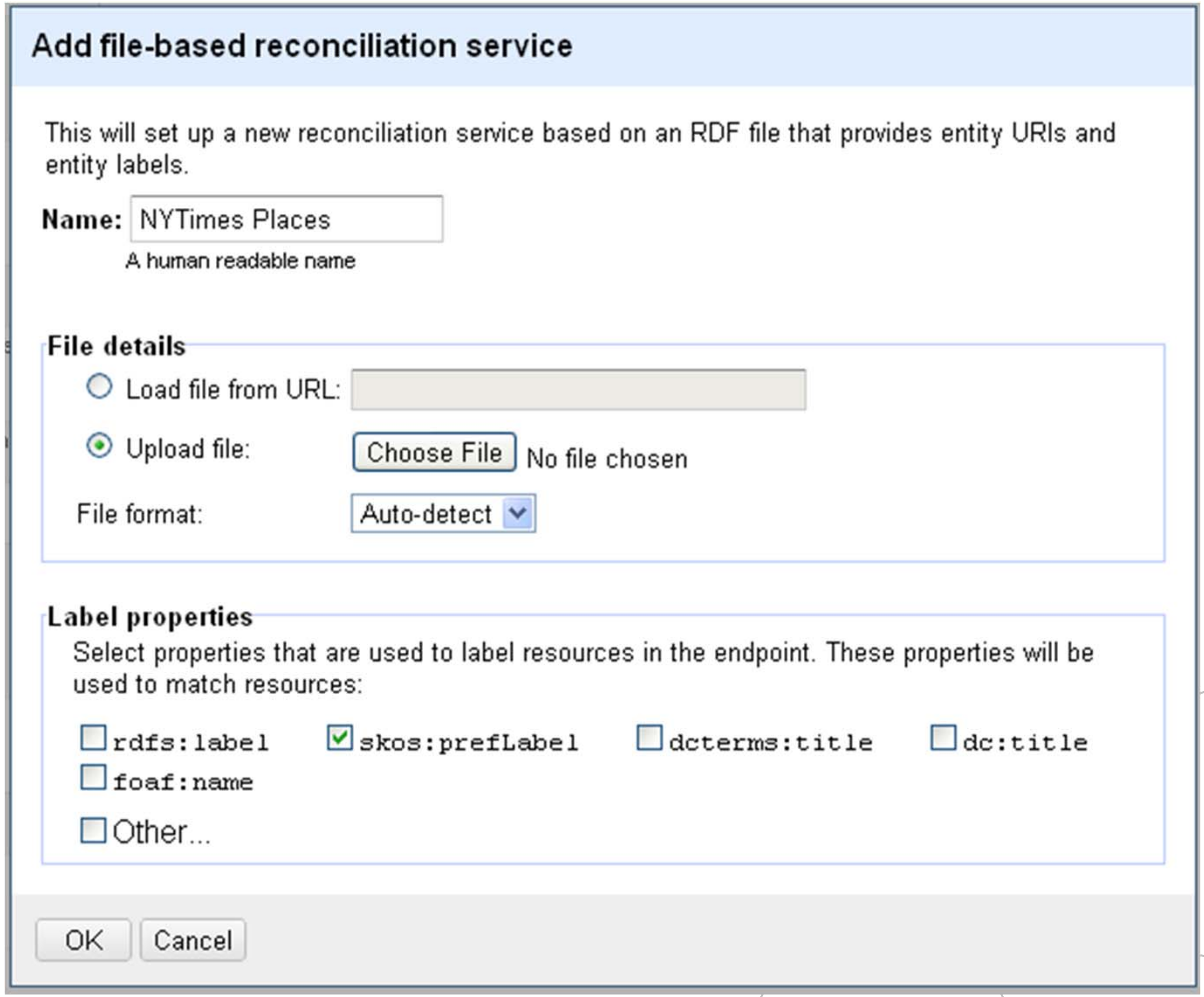
Label properties

Select properties that are used to label resources in the endpoint. These properties will be used to match resources:

☑ rdfs:label □ skos:prefLabel □ dcterms:title □ dc:title
□ foaf:name

□ Other...

Reconcile against an RDF dump



Add file-based reconciliation service

This will set up a new reconciliation service based on an RDF file that provides entity URIs and entity labels.

Name: NYTimes Places

A human readable name

File details

○ Load file from URL:

● Upload file: Choose File No file chosen

File format: Auto-detect

Label properties

Select properties that are used to label resources in the endpoint. These properties will be used to match resources:

□ `rdfs:label` ☑ `skos:prefLabel` □ `dcterms:title` □ `dc:title`
□ `foaf:name`

□ Other...

5-star plan for open data

★ Make your stuff available on the Web

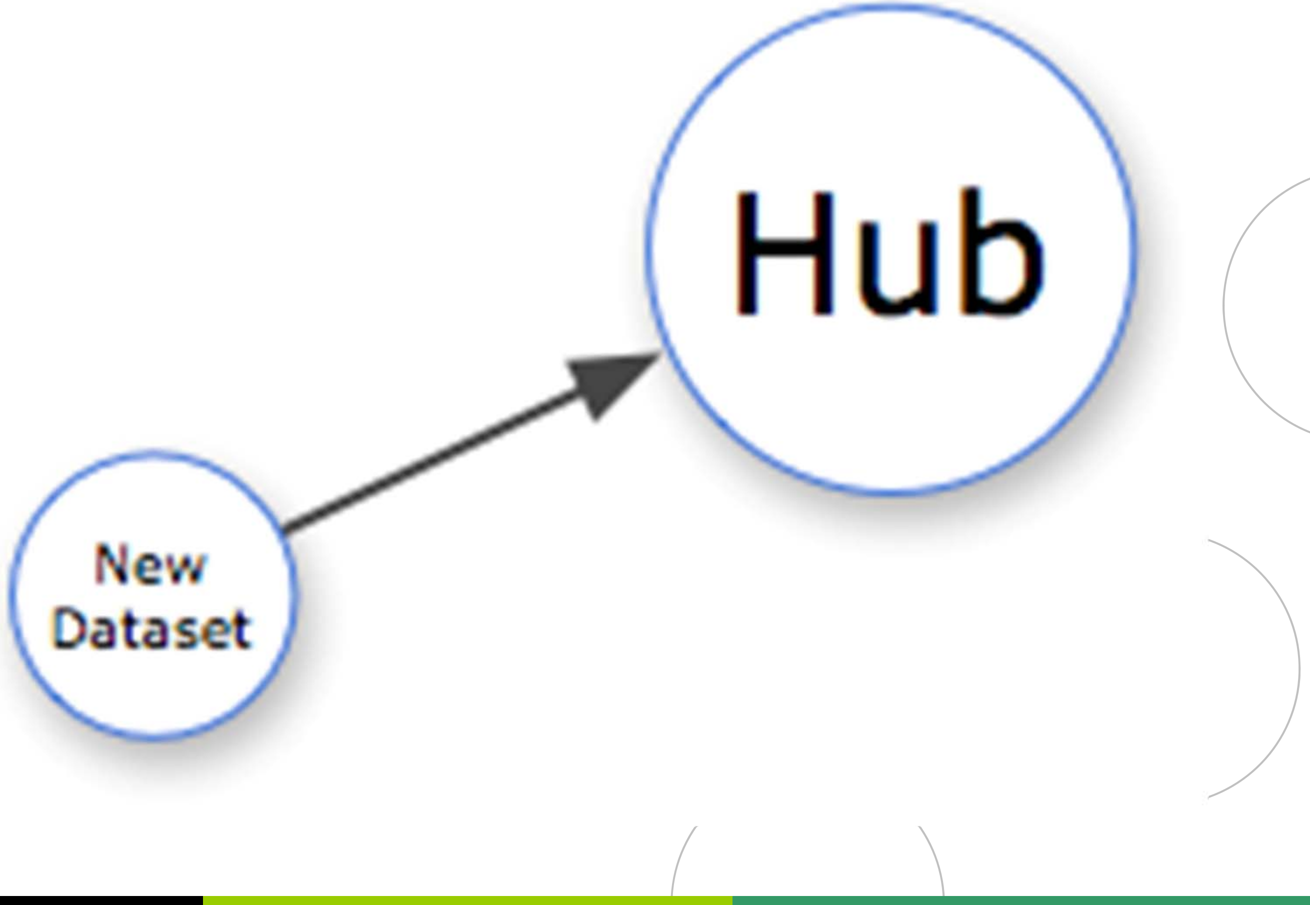
★★ Make it available as structured data
(e.g., an Excel sheet instead of image scan of a table)

★★★ Use a non-proprietary format
(e.g., a CSV file instead of an Excel sheet)

★★★★ Use linked data format
(i.e., URIs to identify things, and RDF to represent data)

★★★★★ Link your data to other people's data to provide context

Making this easier!



- RDF Extension for Google Refine
  http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/
- Reconciliation will be in the upcoming next version

Thanks!