Linking up your data

Outline

- Why do you want links
  - Reuse
  - Reusability
- What types of Links
- How do you link
  - Linking techniques
- What data is out there to link to
  - The LOD cloud

Four rules of Linked Data

1. Use URIs to identify things (**Resources**).

2. Use HTTP URIs so that these things can be referred to and looked up ("**dereference**") by people and user agents

3. Provide useful information (i.e., a **structured description - metadata**) about the thing when its URI is dereferenced.

4. Include **links** to other, related URIs in the exposed data to improve discovery of other related information on the Web.

Reuse

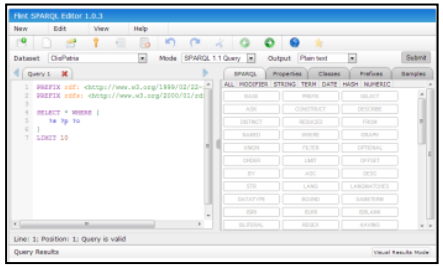
By establishing links from your data to external data, you can reuse the information in the external dataset

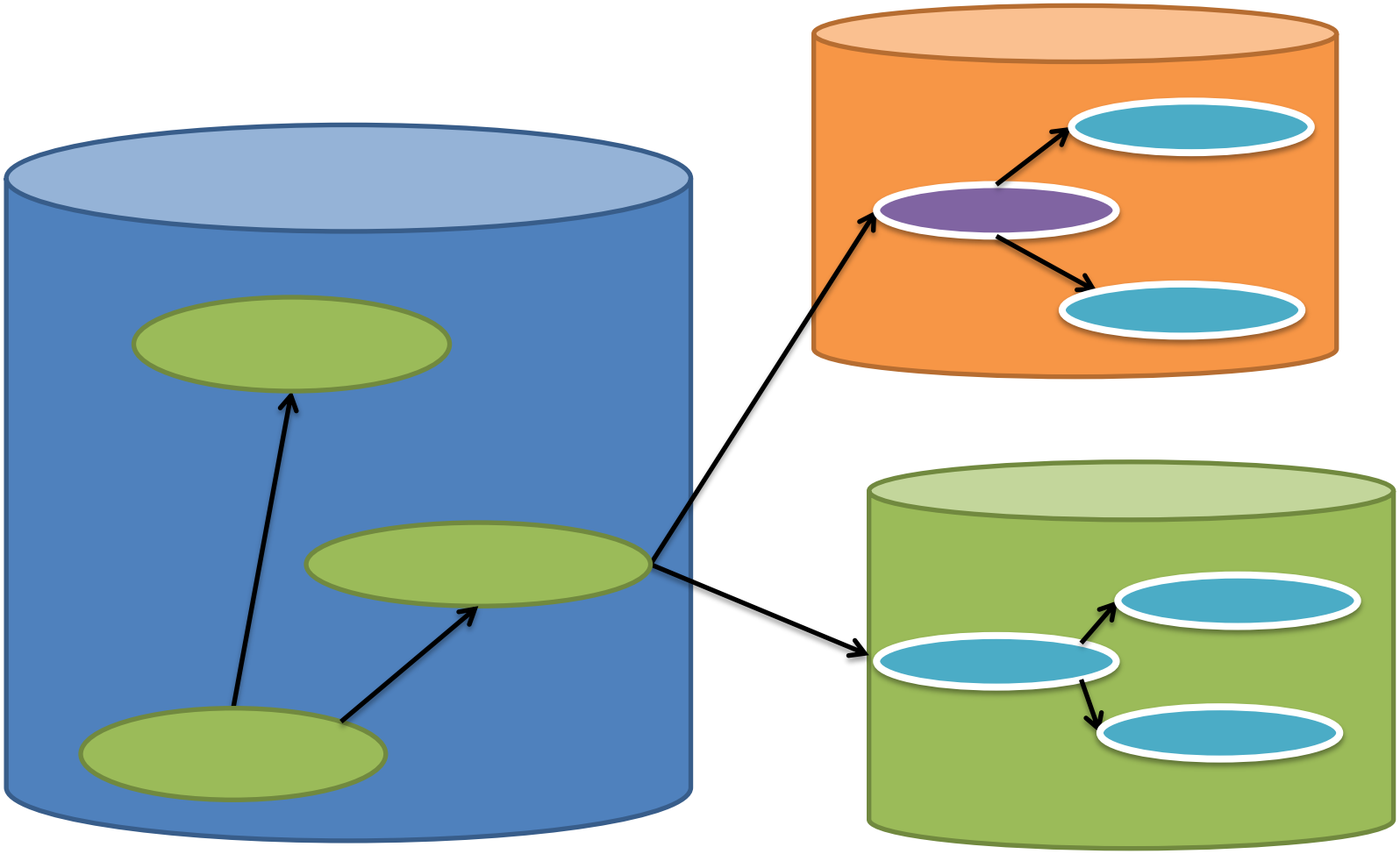
In other words: you get to focus on *your data* and have others *curate their data*

-> FREE LUNCH!

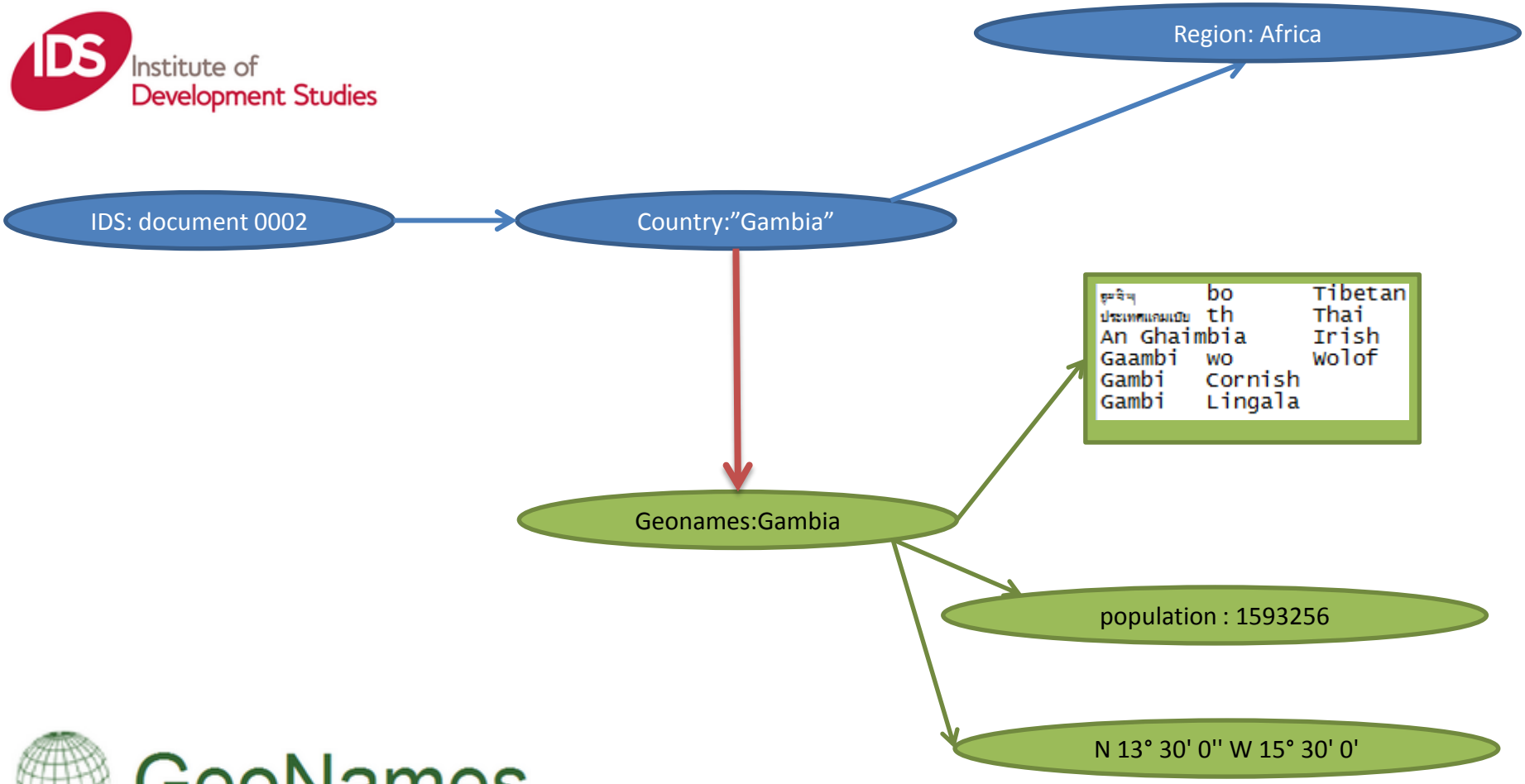
Web of Documents (WWW)
Linked Documents

Linked Data

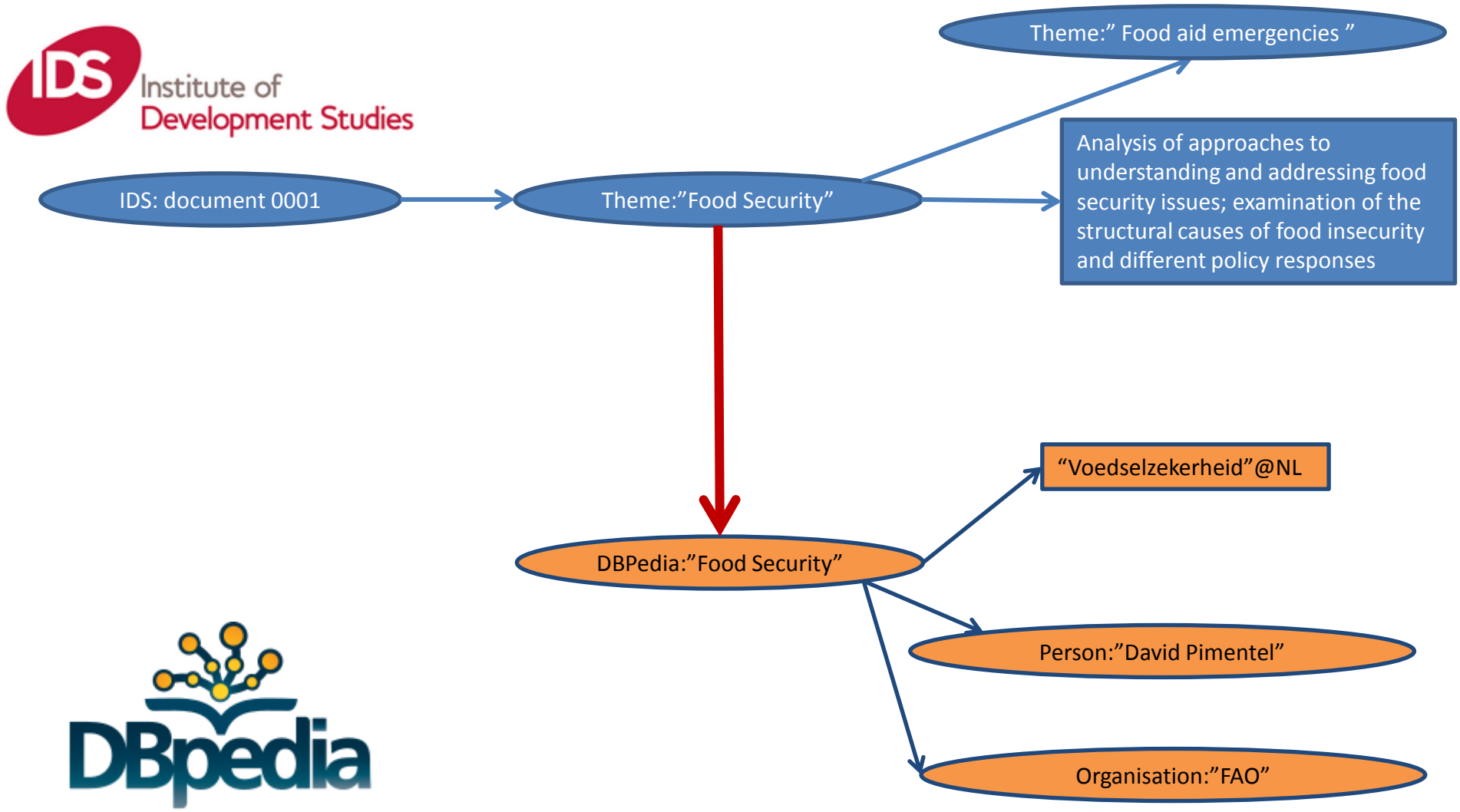




Example: Link to Geonames



Example: Links to DBPedia

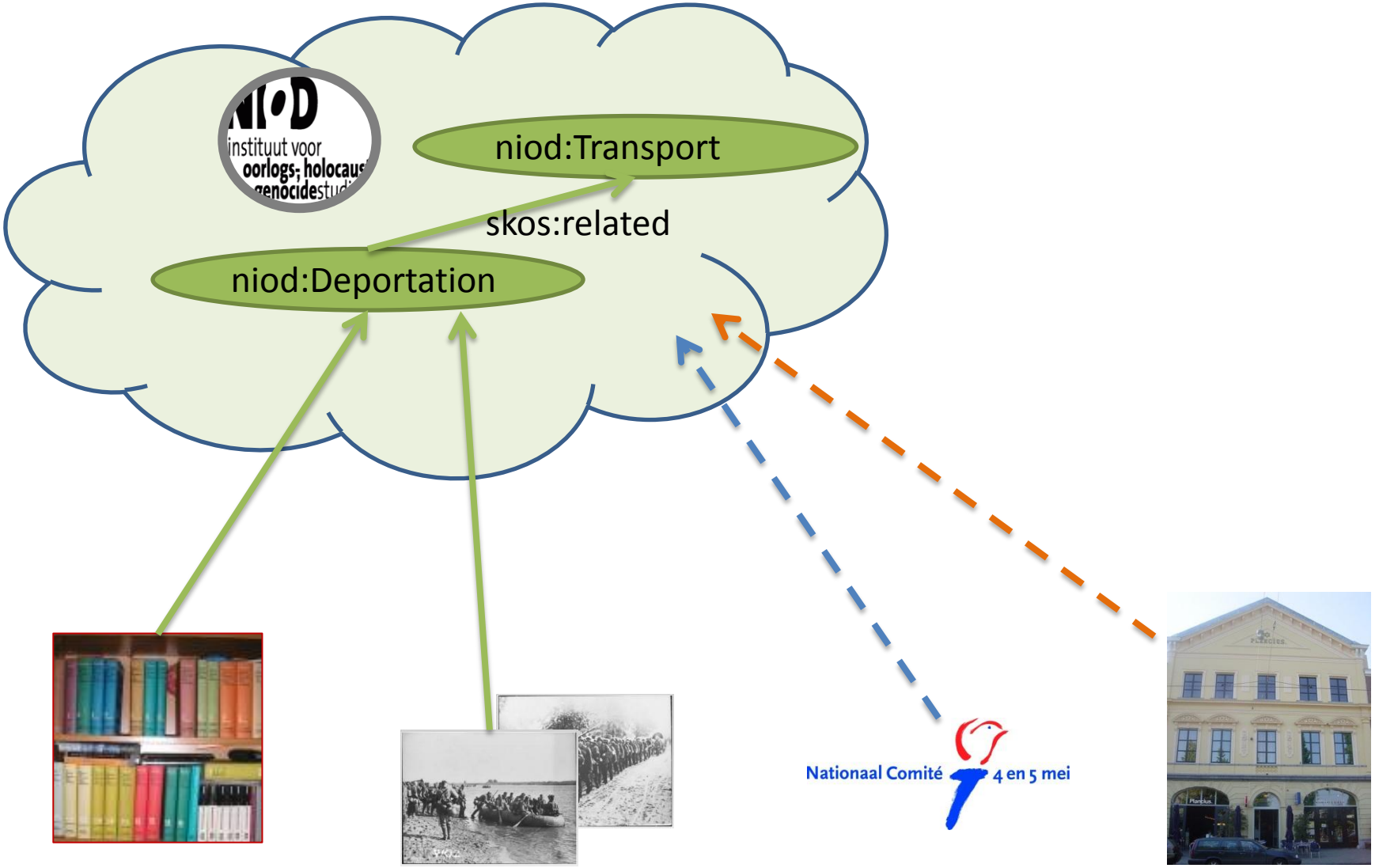


Reusability

- Works exactly the other way around. As a curator of information, I want to make my data as reusable as possible

- Make reverse In-links possible
- Include out-links to add meaning/context to my information

Example: Dutch National War Institute NIOD



Types of Links

- **Relationship Links**
  - point *at related things* in other data sources,
  - other people, places or genes…
- **Identity Links**
  - point at URI aliases used by other data sources to identify the same real-world object or abstract concept.
  - Get more information
  - Enable different views of the world
- **Vocabulary Links**
  - point from data to the definitions of the vocabulary terms that are used to represent the data
  - Interoperability

Relationship Links

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://biglynx.co.uk/people/dave-smith>
  rdf:type foaf:Person ;
  foaf:name "Dave Smith" ;
  foaf:based_near <http://sws.geonames.org/3333125/> ;
  foaf:based_near <http://dbpedia.org/resource/Birmingham> ;
  foaf:topic_interest <http://dbpedia.org/resource/Wildlife_photography> ;
  foaf:knows <http://dbpedia.org/resource/David_Attenborough> .
```

http://sws.geonames.org/3333125/ has all kinds of information about Birmingham

Identity Links

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix louvre: <http://www.louvre.fr/data/> .
@prefix dct: < http://purl.org/dc/terms/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .

louvre:mona_lisa
  rdf:type louvre:Painting ;
  rdfs:label "Mona Lisa" ;
  dct:date "1503-1505"
  owl:sameAs dbpedia:Mona_Lisa.
```

"1503-1505"

Louvre

"1503-1513"

Dbpedia

owl:sameAs , skos:exactMatch, …

Vocabulary Links

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix co: <http://biglynx.co.uk/vocab/sme#> .

<http://biglynx.co.uk/vocab/sme#SmallMediumEnterprise>
rdf:type rdfs:Class ;
rdfs:label "Small or Medium-sized Enterprise" ;
rdfs:subClassOf <http://dbpedia.org/ontology/Company> .
rdfs:subClassOf <http://umbel.org/umbel/sc/Business> ;
rdfs:subClassOf <http://sw.opencyc.org/concept/Mx4rvVjQNpwpEbGdrcN5Y29ycA> ;
rdfs:subClassOf <http://rdf.freebase.com/ns/m/0qb7t> .
```

Using an external metadata schema makes it ***interpretable*** in that schema

You can link your *instance* data to external vocabulary

You can link your *vocabulary* to an external vocabulary

How to link

Choosing a target dataset

- What is the value of the data in the target data set?
- To what extent does this add value to my data set?
- Are there ongoing links to other data set so that applications can tap into a network of interconnected data sources?

- Is the target data set and its namespace under stable ownership and active maintenance?
- Are the URIs in the data set stable and unlikely to change?

Manual Links

- Just write RDF triples where the RDF Object is an external resource

  <mydata:M_Lisa owl:sameAs louvre:Mona_Lisa>

- Refrain from saying something about resources that are not in your namespace
  - ie. Don't add triples that have an external resource as RDF Subject

  ~~< louvre:Mona_Lisa owl:sameAs mydata:M_Lisa >~~

Generating Links

- Based on some key
  - Country codes
  - ISBN code
- Based on similarity
  - Person Names
  - Country names
- Complex combinations of methods and
  - Person Name + Birth date

Matching or Linking is a whole field on its own. Finding out which things are the same is hard!

It is core to the success of Linked Data

SILK

- Framework for finding links between datasets
- Silk Specification Language: XML language to define linking conditions (rules)
- Compare instances based on metrics such as string comparison, etc.
- Define rules manually or visually in SILK workbench
- Execute rules in SILK server

http://www4.wiwiss.fu-berlin.de/bizer/silk/

```
<Interlink>
  <LinkType>owl:sameAs</LinkType>
  <SourceDataset dataSource="sider" var="a">
    <RestrictTo>?a rdf:type sider:drugs</RestrictTo>
  </SourceDataset>
  <TargetDataset dataSource="drugbank" var="b">
    <RestrictTo>?b rdf:type drugbank:drugs</RestrictTo>
  </TargetDataset>
  <LinkCondition>
    <Aggregate type="max">
      <Compare metric="levenshteinDistance" threshold="1">
        <TransformInput function="lowerCase">
          <Input path="?a/rdfs:label"/>
        </TransformInput>
        <TransformInput function="lowerCase">
          <Input path="?b/rdfs:label"/>
        </TransformInput>
      </Compare>
    </Aggregate>
  </LinkCondition>
  <Filter limit="1"/>
</Interlink>
```

ClioPatria CPACK: Amalgame alignment platform

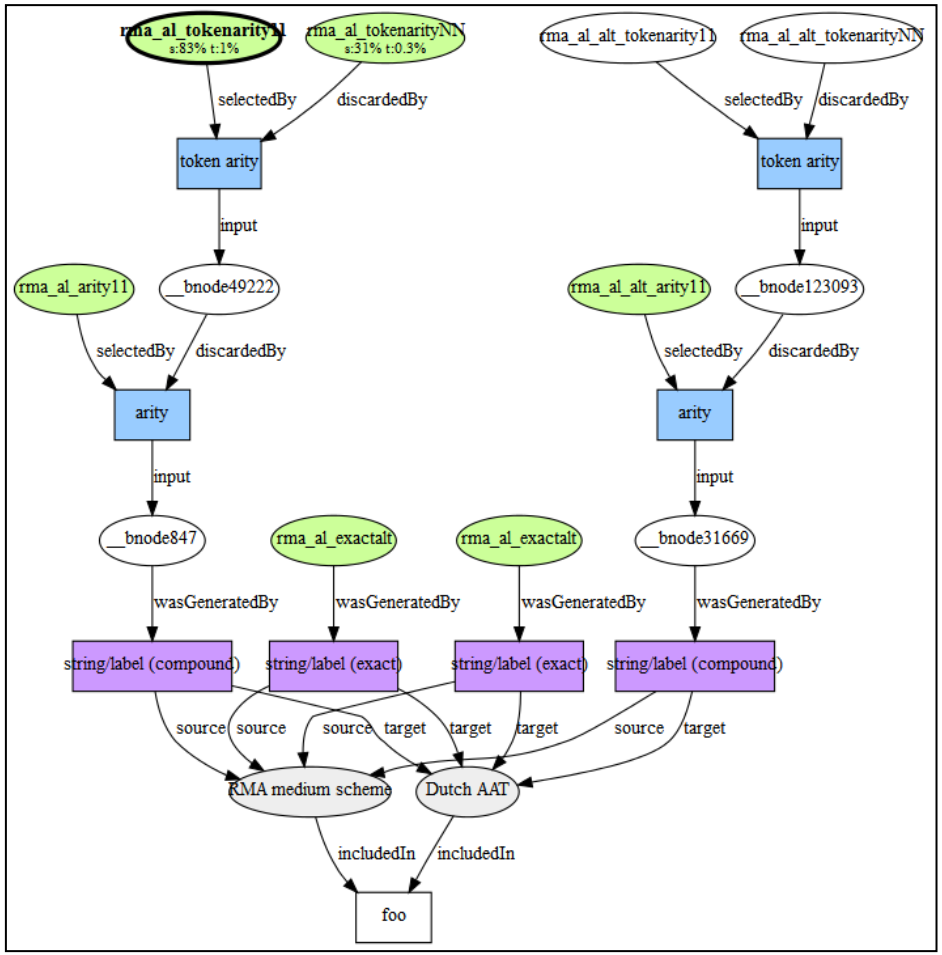
Semi-automatic matching

Simple automatic techniques

String match (exact/sim)

Selection, set splitting

chained together by hand



What can I link to?

Flickr/ horrigans

datahub.io/group/lodcloud

datahub
The easy way to get, use and share data

Datasets

/ Groups / **Linking Open Data Cloud**

Linking Open Data Cloud

This group catalogs data sets that are available on the Web as Linked Data and contain data links pointing at other Linked Data sets. The descriptions of the data sets in this... read more

Followers **3**

Datasets **337**

Groups Clear All

Linking Open Data C... (337)

Library Linked Data (47)

Datasets | Activity Stream | About

Search...

337 datasets found

theses.fr

Source : theses.fr theses.fr is the french dissertations yet, but will be as open as possible NB : The RDFa of

example/rdf+xml meta/sitemap

Calames

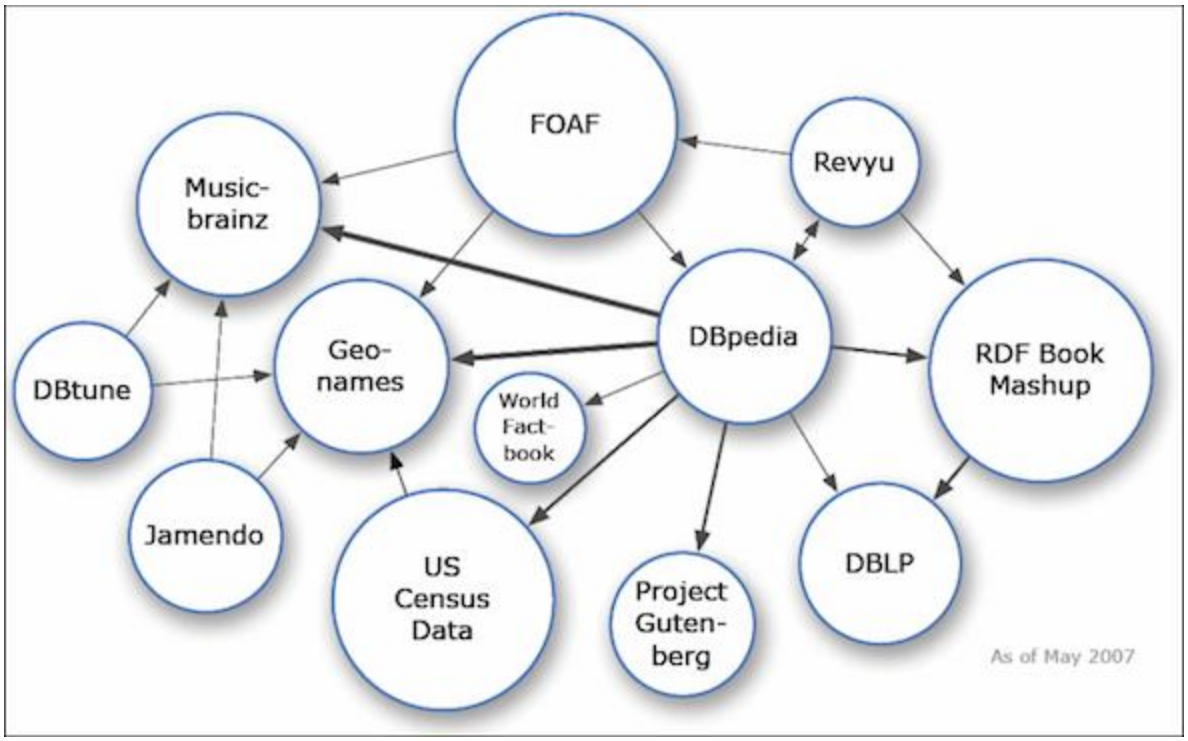
Source : Calames Calames is the French academic u maintained by ABES Licensing : not yet, but will be as

example/rdfa meta/sitemap

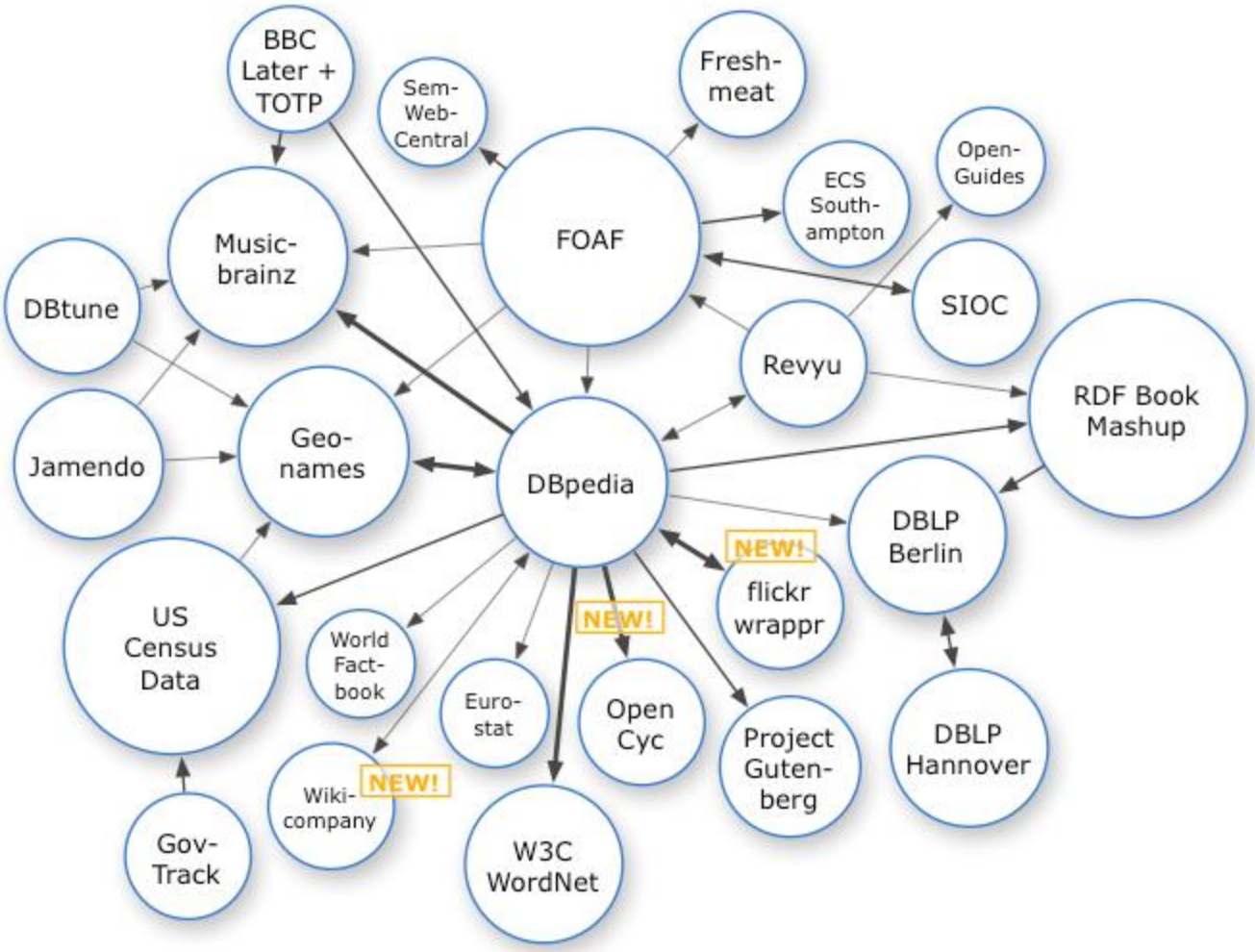
IdRef: Sudoc authority data

Source : IdRef IdRef is a Web application dedicated t academic union catalogue, maintained by ABES. IdRe

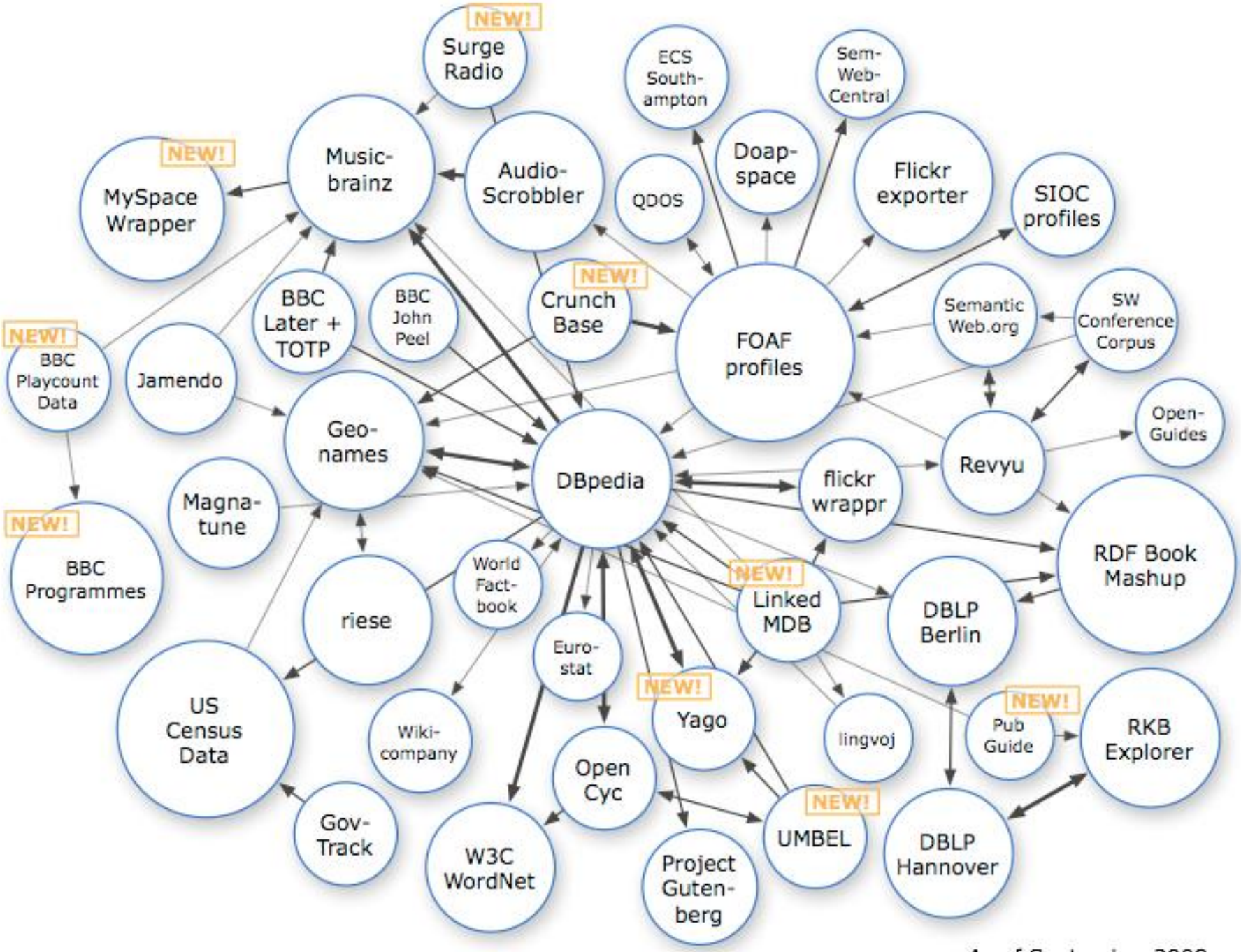
example/rdf+xml uri

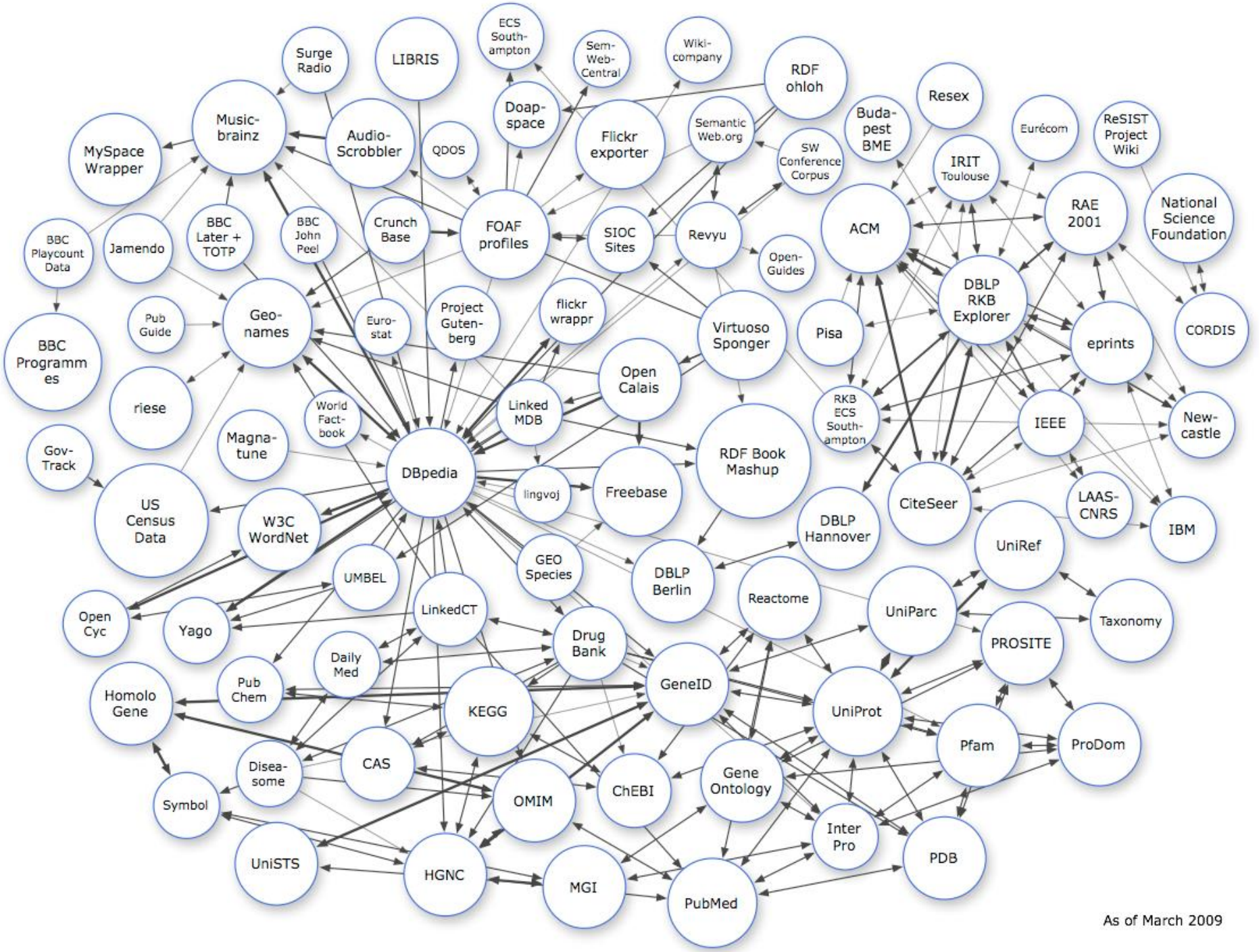

FOAF
Music-
brainz
Revyu
DBpedia
Geo-
names
DBtune
World
Fact-
book
RDF Book
Mashup
Jamendo
US
Census
Data
Project
Guten-
berg
DBLP
As of May 2007

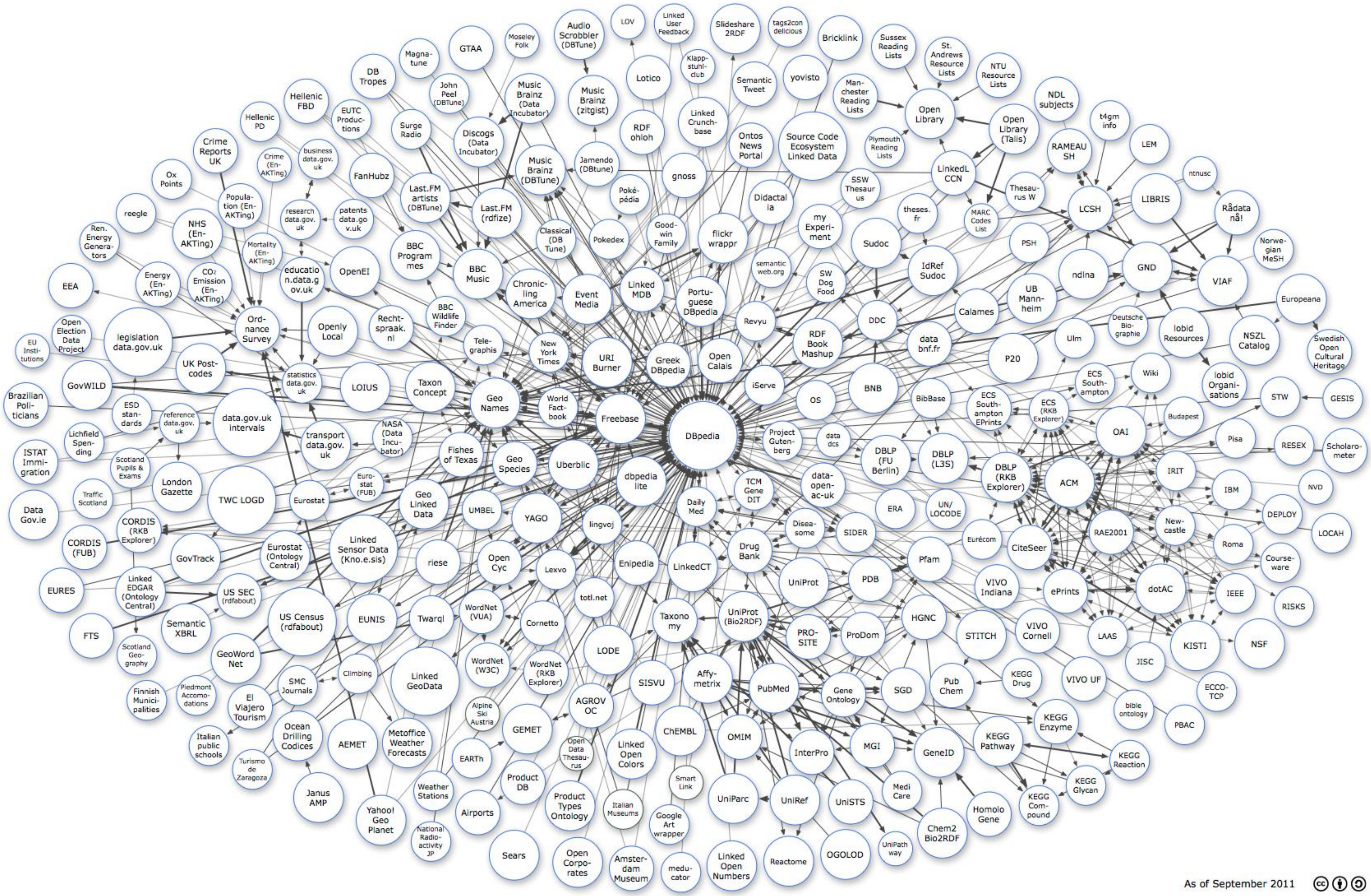
May 2007


BBC Later + TOTP
Sem-Web-Central
Fresh-meat
Open-Guides
ECS South-ampton
FOAF
Music-brainz
DBtune
SIOC
Revyu
RDF Book Mashup
Jamendo
Geo-names
DBpedia
DBLP Berlin
NEW!
flickr wrappr
US Census Data
World Fact-book
Euro-stat
NEW!
Open Cyc
Project Guten-berg
DBLP Hannover
NEW!
Wiki-company
Gov-Track
W3C WordNet

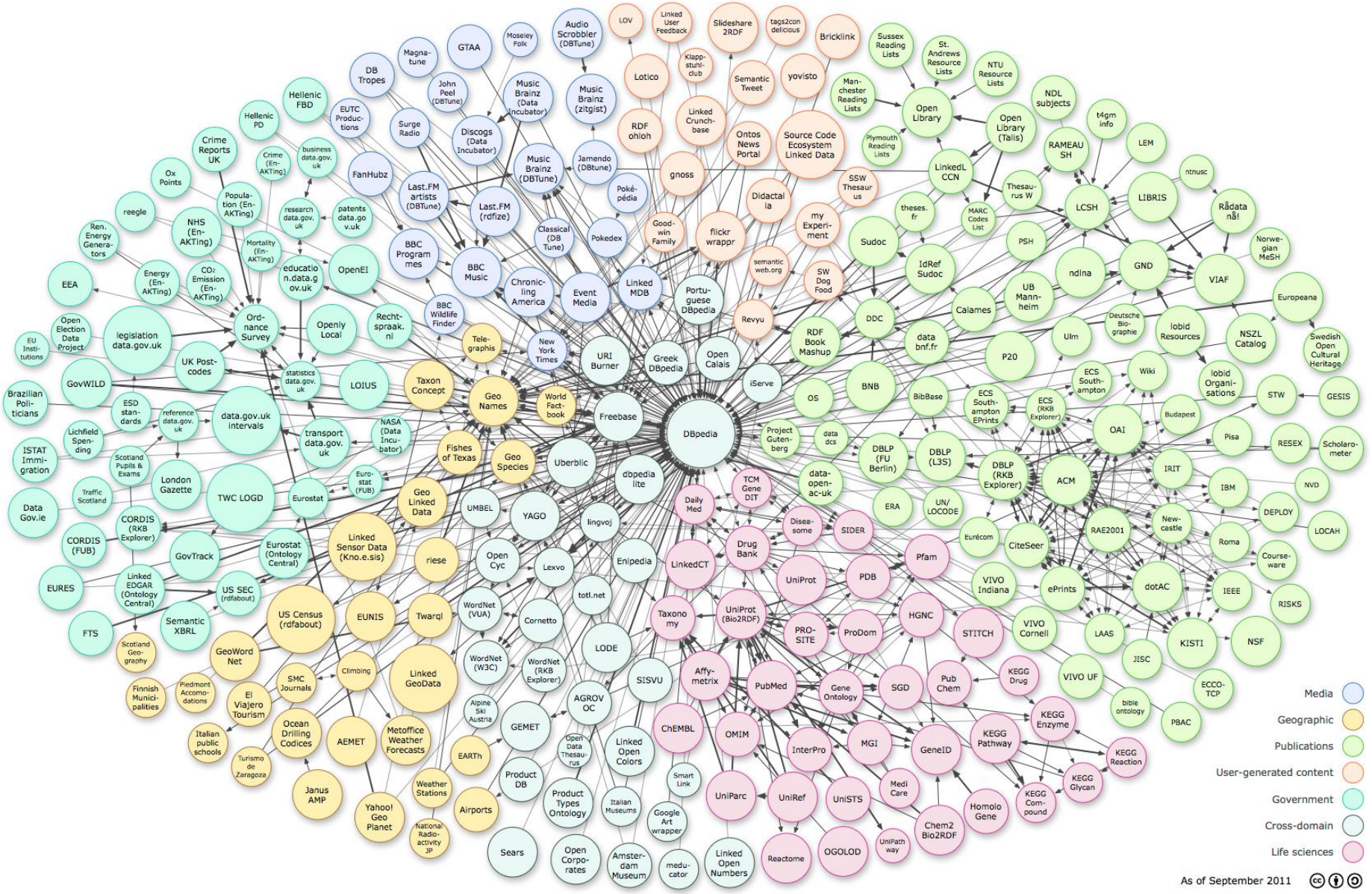
Oct 2007


NEW!
Surge Radio
ECS South-ampton
Sem-Web-Central
NEW!
MySpace Wrapper
Music-brainz
Audio-Scrobbler
QDOS
Doap-space
Flickr exporter
SIOC profiles
BBC Later + TOTP
BBC John Peel
NEW!
Crunch Base
FOAF profiles
Semantic Web.org
SW Conference Corpus
NEW!
BBC Playcount Data
Jamendo
Geo-names
DBpedia
Open-Guides
Revyu
Magna-tune
flickr wrappr
RDF Book Mashup
NEW!
BBC Programmes
World Fact-book
NEW!
Linked MDB
DBLP Berlin
riese
Euro-stat
US Census Data
NEW!
Yago
NEW!
Pub Guide
RKB Explorer
Wiki-company
lingvoj
Open Cyc
NEW!
UMBEL
Gov-Track
W3C WordNet
Project Guten-berg
DBLP Hannover
As of September 2008


Surge Radio
LIBRIS
ECS South-ampton
Sem-Web-Central
Wiki-company
RDF ohloh
Resex
Buda-pest BME
Eurécom
ReSIST Project Wiki
Music-brainz
Audio-Scrobbler
QDOS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
IRIT Toulouse
MySpace Wrapper
BBC Playcount Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
National Science Foundation
BBC Programmes
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
eprints
CORDIS
riese
World Fact-book
Linked MDB
Open Calais
RKB ECS South-ampton
IEEE
New-castle
Gov-Track
Magna-tune
DBpedia
RDF Book Mashup
US Census Data
W3C WordNet
lingvoj
Freebase
DBLP Hannover
CiteSeer
LAAS-CNRS
IBM
UniRef
UMBEL
GEO Species
DBLP Berlin
Reactome
Open Cyc
Yago
LinkedCT
Drug Bank
UniParc
Taxonomy
PROSITE
Daily Med
Pub Chem
Homolo Gene
KEGG
GeneID
UniProt
Pfam
ProDom
Disea-some
CAS
OMIM
ChEBI
Gene Ontology
Symbol
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB
As of March 2009



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Domain	Number of datasets	Triples	%	(Out-)Links	%
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,328	7.11 %
Government	49	13,315,009,400	42.09 %	19,343,519	3.84 %
Publications	87	2,950,720,693	9.33 %	139,925,218	27.76 %
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	38.06 %
User-generated content	20	134,127,413	0.42 %	3,449,143	0.68 %
	295	31,634,213,770		503,998,829	

Table 3.1: Number of data sets, amount of triples, and amount of RDF links per topical domain (http://lod-cloud.net/state/)

Most used vocabularies

Vocabulary prefix	Vocabulary link	Number of usages in data sets
dc	http://purl.org/dc/elements/1.1/	92 (31.19 %)
foaf	http://xmlns.com/foaf/0.1/	81 (27.46 %)
skos	http://www.w3.org/2004/02/skos/core#	58 (19.66 %)
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	25 (8.47 %)
xhtml	http://www.w3.org/1999/xhtml/vocab#	19 (6.44 %)
akt	http://www.aktors.org/ontology/portal#	17 (5.76 %)
bibo	http://purl.org/ontology/bibo/	14 (4.75 %)
mo	http://purl.org/ontology/mo/	13 (4.41 %)
vcard	http://www.w3.org/2006/vcard/ns#	10 (3.39 %)
sioc	http://rdfs.org/sioc/ns#	10 (3.39 %)
cc	http://creativecommons.org/ns#	8 (2.71 %)
geonames	http://www.geonames.org/ontology#	6 (2.03 %)
frbr	http://purl.org/vocab/frbr/core#	6 (2.03 %)

http://lod-cloud.net/state/