Web of Data

Iztok Savnik, Famnit, UP

December, 2018.

Outline

1) Graph data model (RDF)

2) Popular graph databases on Web

3) Linked data and applications
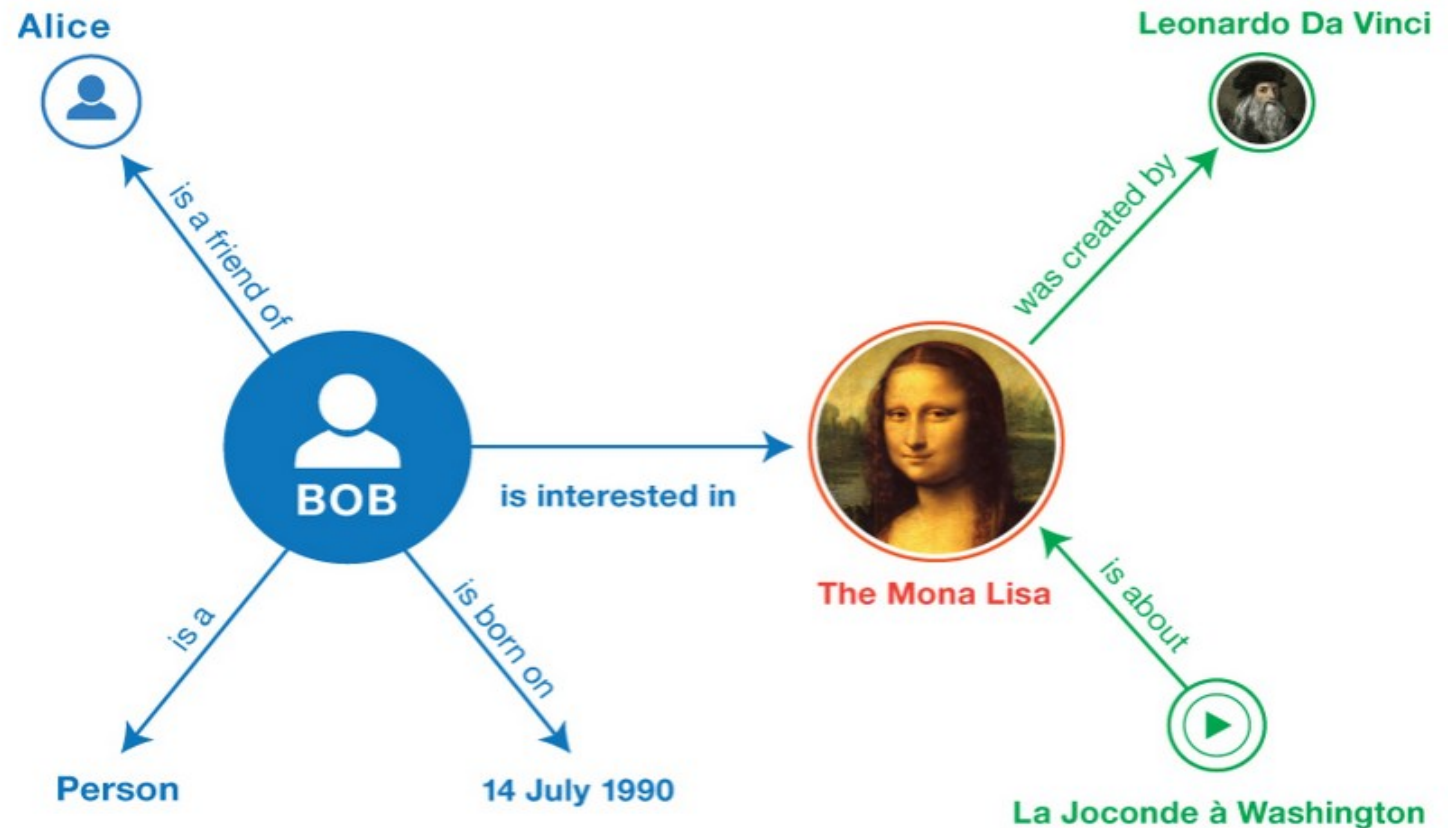
Graph data model (RDF + RDFS)

Graph data model

- Graph database
  - Database that uses graphs for the representation of data and queries

- Vertexes
  - Represent things, persons, concepts, classes, ...

- Arcs
  - Represent properties, relationships, associations, ...
  - Arcs have labels !

RDF

- Resource Description Framework
  - Tim Berners Lee, 1998-2009
  - This is movement !
- What is behind ?
  - Graphs
  - Triples (3)
  - Semantic data models
  - Human associative memory (psychology)
  - Associative neural networks
  - Hopfield Network

RDF

```
<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <the 4th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>
```

RDF syntax

- N3, TVS
- Turtle
- TriG
- N-Triples
- RDF/XML
- RDF/JSON

Name spaces

- Using short names for URL-s
  - Long names are tedious
- Simple but strong concept
- Defining name space:

prefix rdf:, namespace URI: http://www.w3.org/1999/02/22-rdf-syntax-ns#

prefix rdfs:, namespace URI: http://www.w3.org/2000/01/rdf-schema#

prefix dc:, namespace URI: http://purl.org/dc/elements/1.1/

prefix owl:, namespace URI: http://www.w3.org/2002/07/owl#

prefix ex:, namespace URI: http://www.example.org/ (or http://www.example.com/)

prefix xsd:, namespace URI: http://www.w3.org/2001/XMLSchema#

N-Triples

```
<http://example.org/bob#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://example.org/bob#me> <http://xmlns.com/foaf/0.1/knows> <http://example.org/alice#me> .
<http://example.org/bob#me> <http://schema.org/birthDate> "1990-07-04"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://example.org/bob#me> <http://xmlns.com/foaf/0.1/topic_interest> <http://www.wikidata.org/entity/Q12418> .
<http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/title> "Mona Lisa" .
<http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/creator> <http://dbpedia.org/resource/Leonardo_da_Vinci> .
<http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619> <http://purl.org/dc/terms/subject> <
```

Turtle

```
01   BASE   <http://example.org/>
02   PREFIX foaf: <http://xmlns.com/foaf/0.1/>
03   PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
04   PREFIX schema: <http://schema.org/>
05   PREFIX dcterms: <http://purl.org/dc/terms/>
06   PREFIX wd: <http://www.wikidata.org/entity/>
07
08   <bob#me>
09      a foaf:Person ;
10      foaf:knows <alice#me> ;
11      schema:birthDate "1990-07-04"^^xsd:date ;
12      foaf:topic_interest wd:Q12418 .
13
14   wd:Q12418
15      dcterms:title "Mona Lisa" ;
16      dcterms:creator <http://dbpedia.org/resource/Leonardo_da_Vinci> .
17
18   <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>
19      dcterms:subject wd:Q12418 .
```
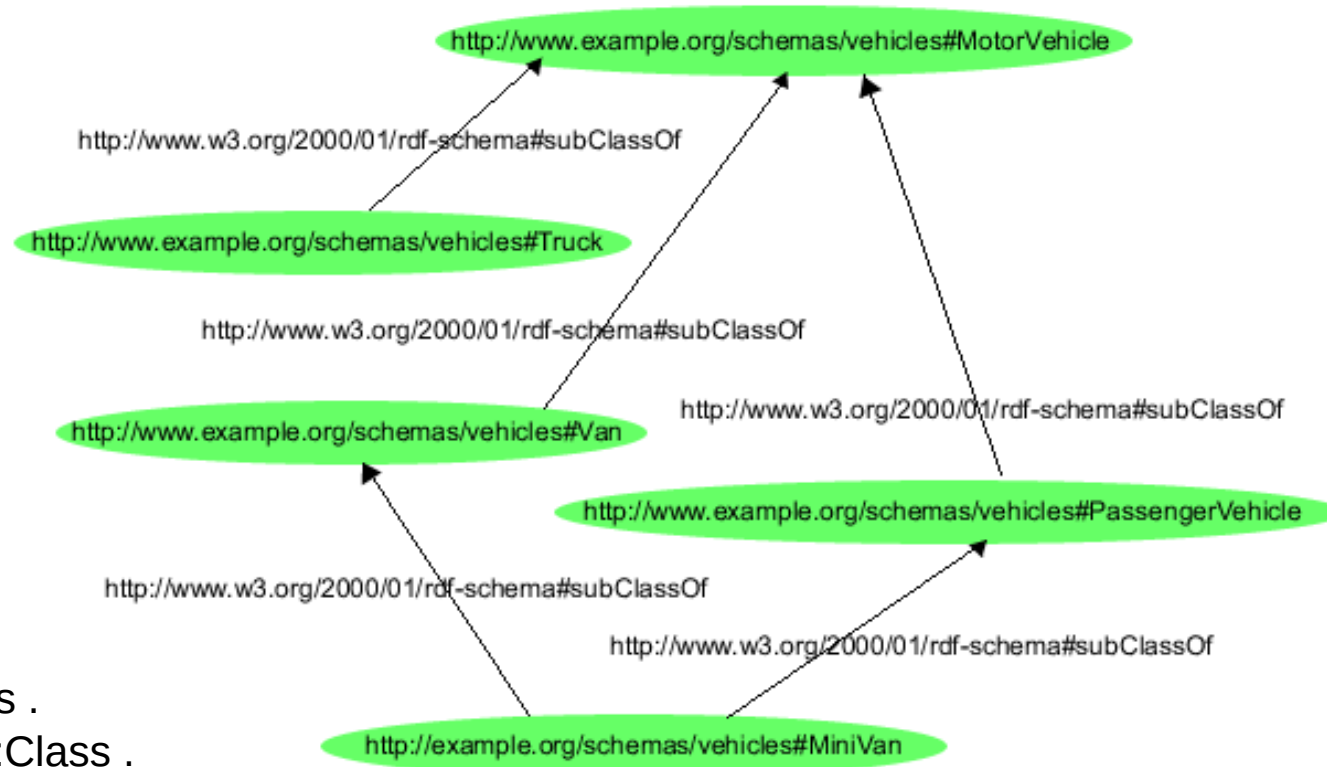
9

Additional RDF Constructs

- Complex values
  - Bags, lists, trees, graphs
- Empty nodes
- Types of atomic values
- Types of nodes
- Reification

RDF Schema

- RDFS

- Knowledge representation language

  - Not just graph any more !

  - AI Frames, Object Model

- Small dictionary for RDFS

  - rdfs:class, rdfs:subClassOf, rdfs:type

  - rdfs:property, rdfs:subPropertyOf

  - rdfs:domain, rdfs:range

Classes

ex:MotorVehicle rdf:type rdfs:Class .
ex:PassengerVehicle rdf:type rdfs:Class .
ex:Van rdf:type rdfs:Class .
ex:Truck rdf:type rdfs:Class .
ex:MiniVan rdf:type rdfs:Class .

ex:PassengerVehicle rdfs:subClassOf ex:MotorVehicle .
ex:Van rdfs:subClassOf ex:MotorVehicle .
ex:Truck rdfs:subClassOf ex:MotorVehicle .

ex:MiniVan rdfs:subClassOf ex:Van .
ex:MiniVan rdfs:subClassOf ex:PassengerVehicle .

12

SPARQL

- **S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage

- SPARQL query

  – Graph can include variables in place of constants

- Operations

  – JOIN (natural, left-join)

  – AND, FILTER, UNION, OPTIONAL

- Commercial DBMS-s

  – Implement RDF and SPARQL

Example SPARQL query

```
PREFIX
    abc: <http://mynamespace.com/exampleOntology#>
SELECT ?capital ?country
WHERE { ?x abc:cityname ?capital.
        ?y abc:countryname ?country.
        ?x abc:isCapitalOf ?y.
        ?y abc:isInContinent abc:africa. }
```


14
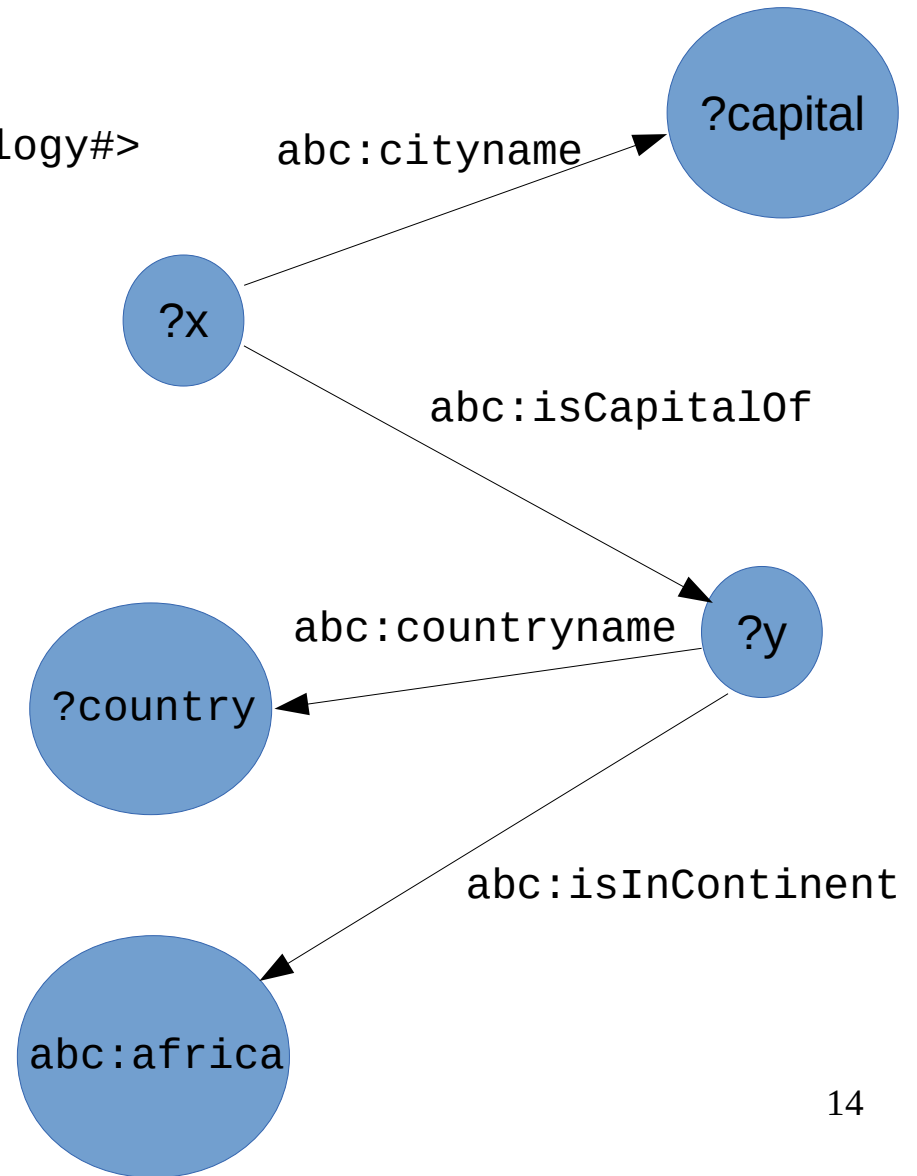
Logic - OWL

- **Ontology language**
  - Knowledge representation + Logic
- Based on **description logic**
  - Fragments of predicate calculus
  - Hierarchy of DL languages
- **OWL reasoners**
  - FaCT++, HermiT, RacerPro, Pellet, ...

Protégé
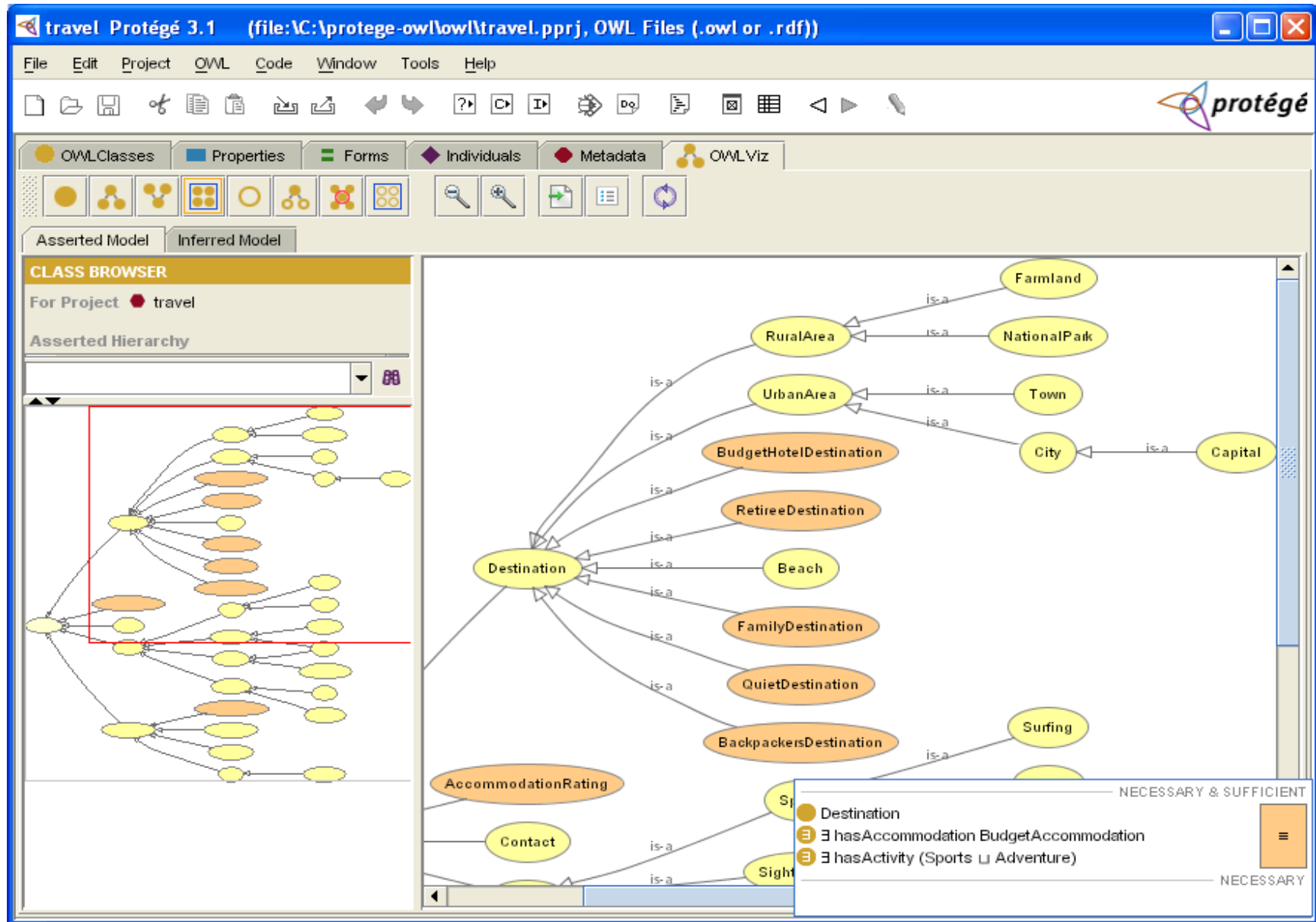
Popular graph databases on Web

Terminology

- Linked data
  - Linked Open Data
- Open data
- Graph databases
- Knowledge bases
- Knowledge graphs

Wordnet

- Princeton's large lexical database of English.
  - Cognitve synonims: synsets
    - 117,000 synsets
  - Synsets are linked by:
    - conceptual-semantic relationships, and
    - lexical relationships.
    - Include definitions of synsets.
  - Main relationships:
    - Synonymy, hyponymy (ISA), meronymy (part-whole), antonymy

Linked Open Data

- Datasets are represented in RDF
  - Wikipedia, Wikibooks, Geonames, MusicBrainz, WordNet, DBLP bibliography
- Number of triples: 33 Giga (10^9) (2011)
- Governments:
  - USA, UK, Japan, Austria, Belgium, France, Germany, ...
- Active community
  - http://en.wikipedia.org/wiki/Open_Data
  - http://www.w3.org/LOD

LOD zbirke na spletu: Maj 2007
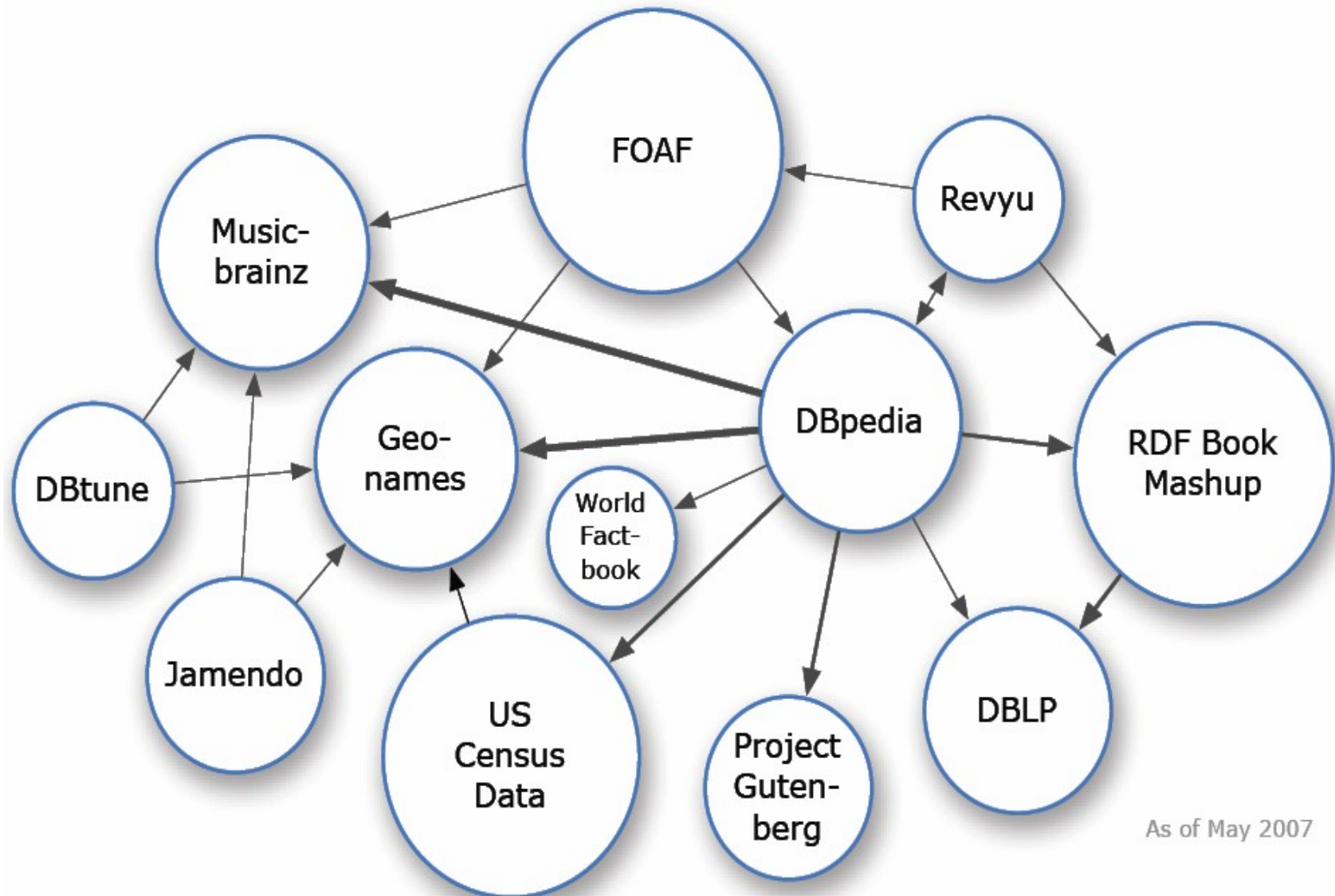
- Nad 500 miljonov RDF trojic
- Okoli 120,000 RDF povezav med pod.viri

LOD zbirke na spletu: September 2008

As of September 2008

LOD zbirke na spletu: Marec 2009

As of March 2009
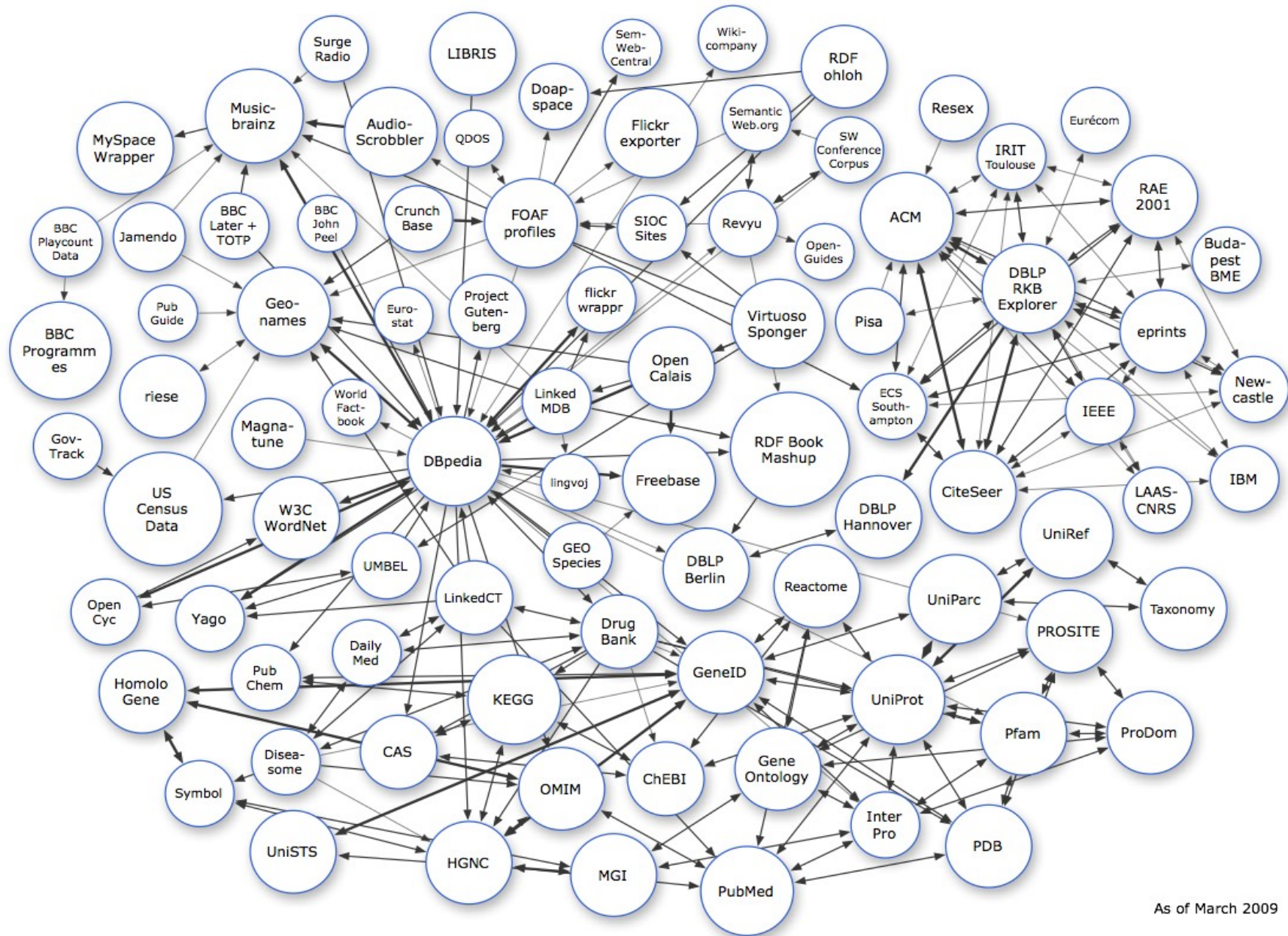
LOD zbirke na spletu: Julij 2009
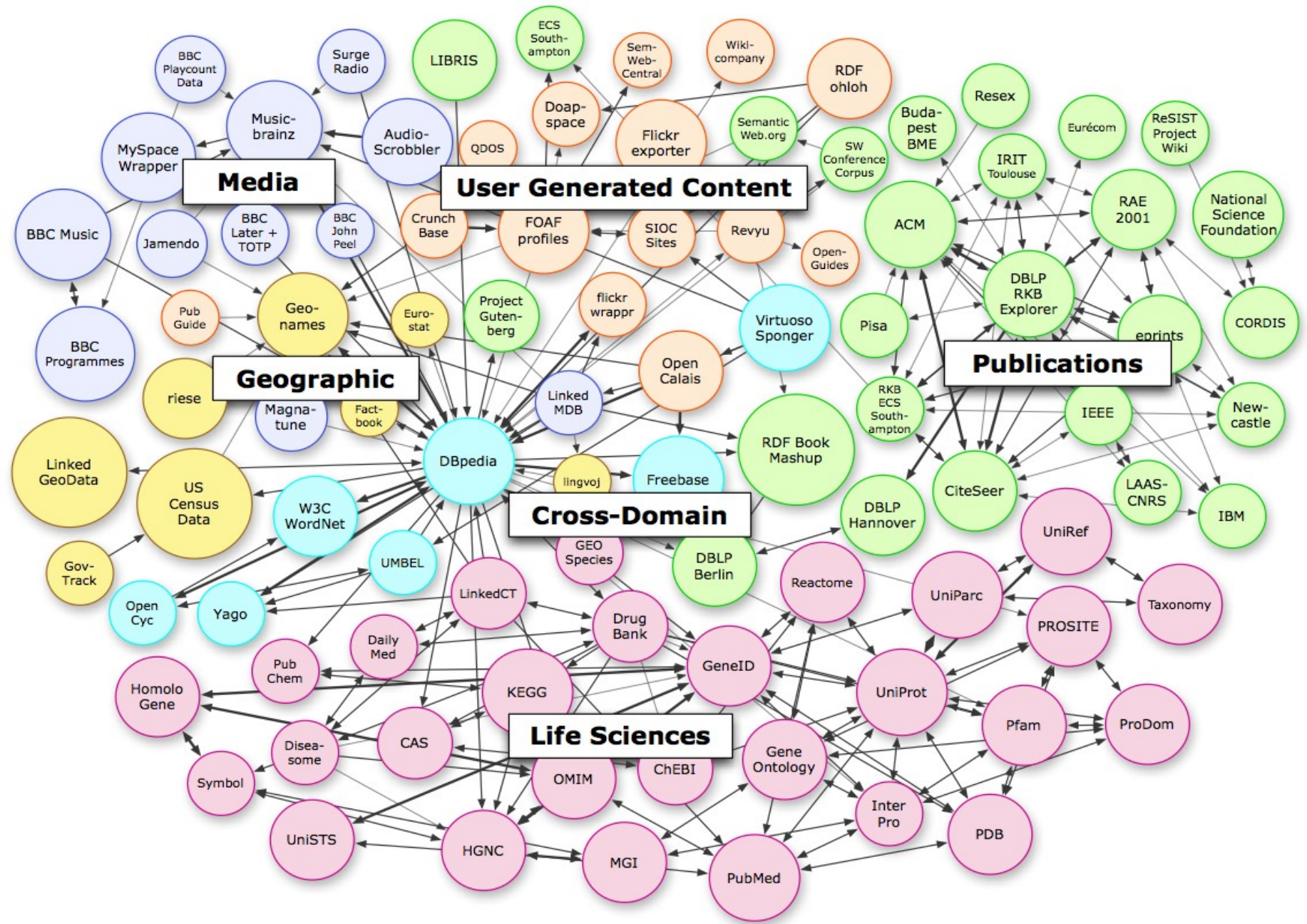
LOD zbirke na spletu: Sept 2010

LOD zbirke na spletu: Sept 2011

As of September 2011

Legend:
- Media
- Geographic
- Publications
- User-generated content
- Government
- Cross-domain
- Life sciences

LOD Cloud, 2014

Legend:
- Media
- Geographic
- Publications
- User-Generated Content
- Government
- Cross-Domain
- Life Sciences
- Social Networking
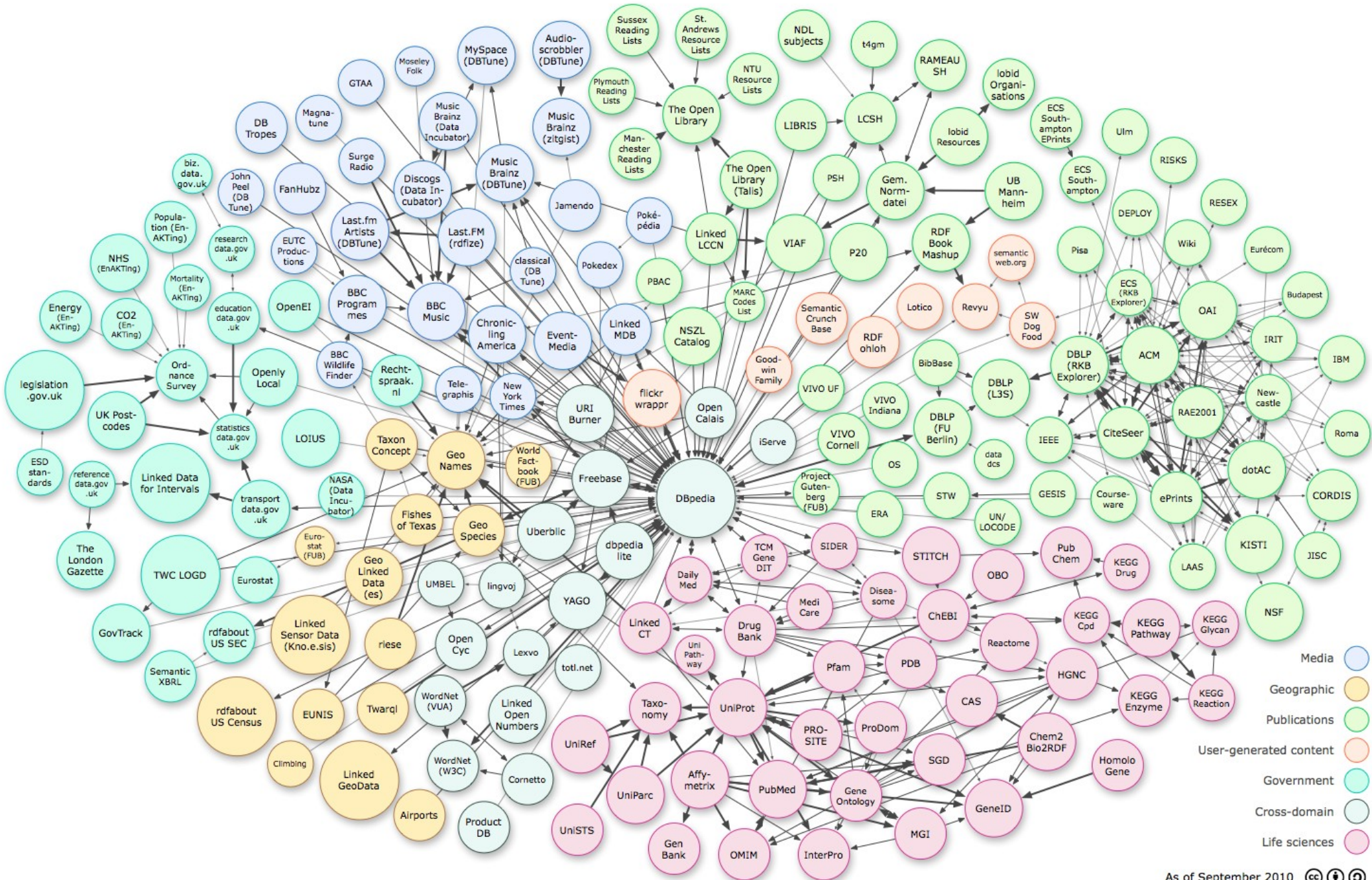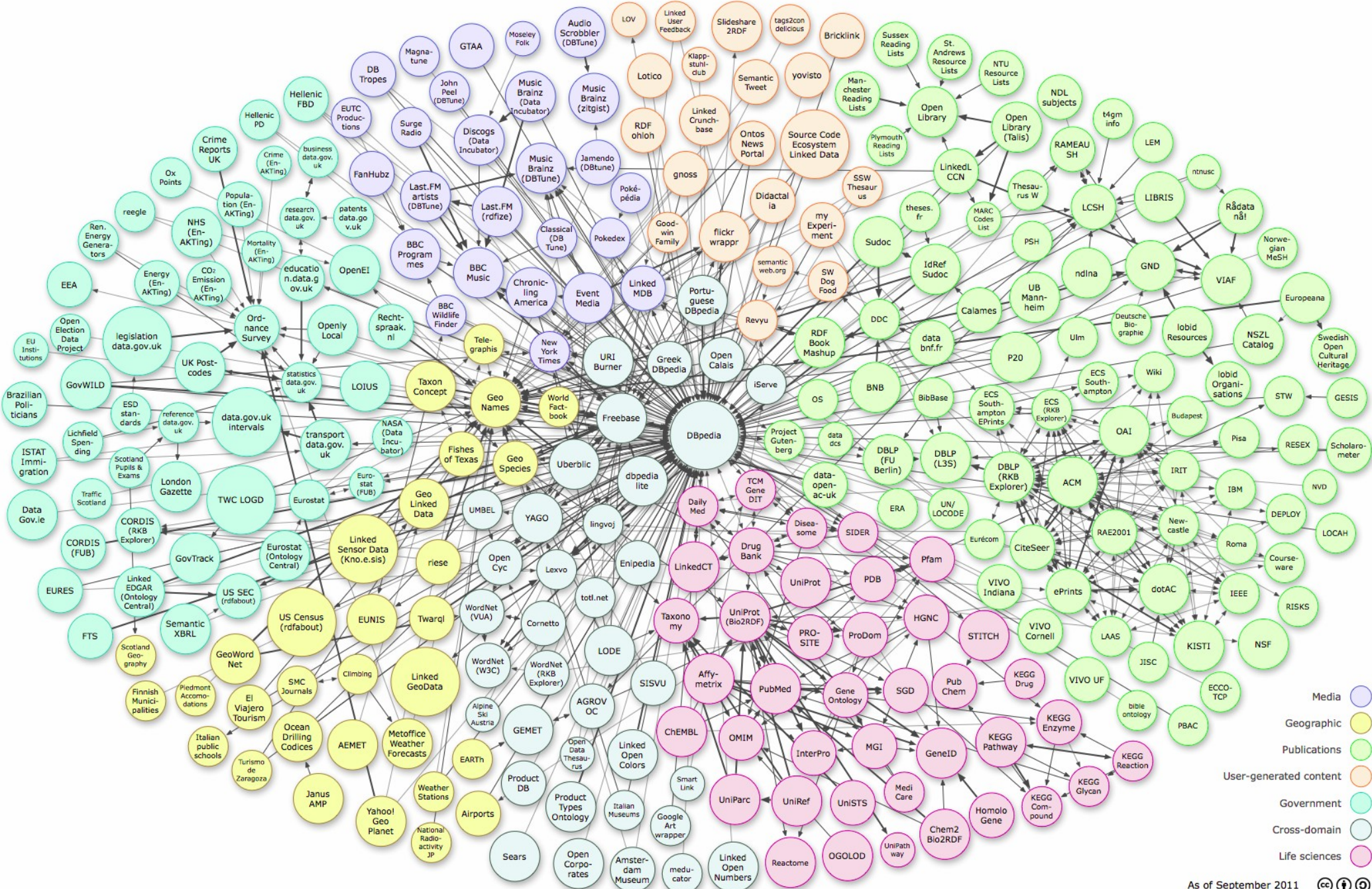
Crawlable Linked Datasets as of April 2014
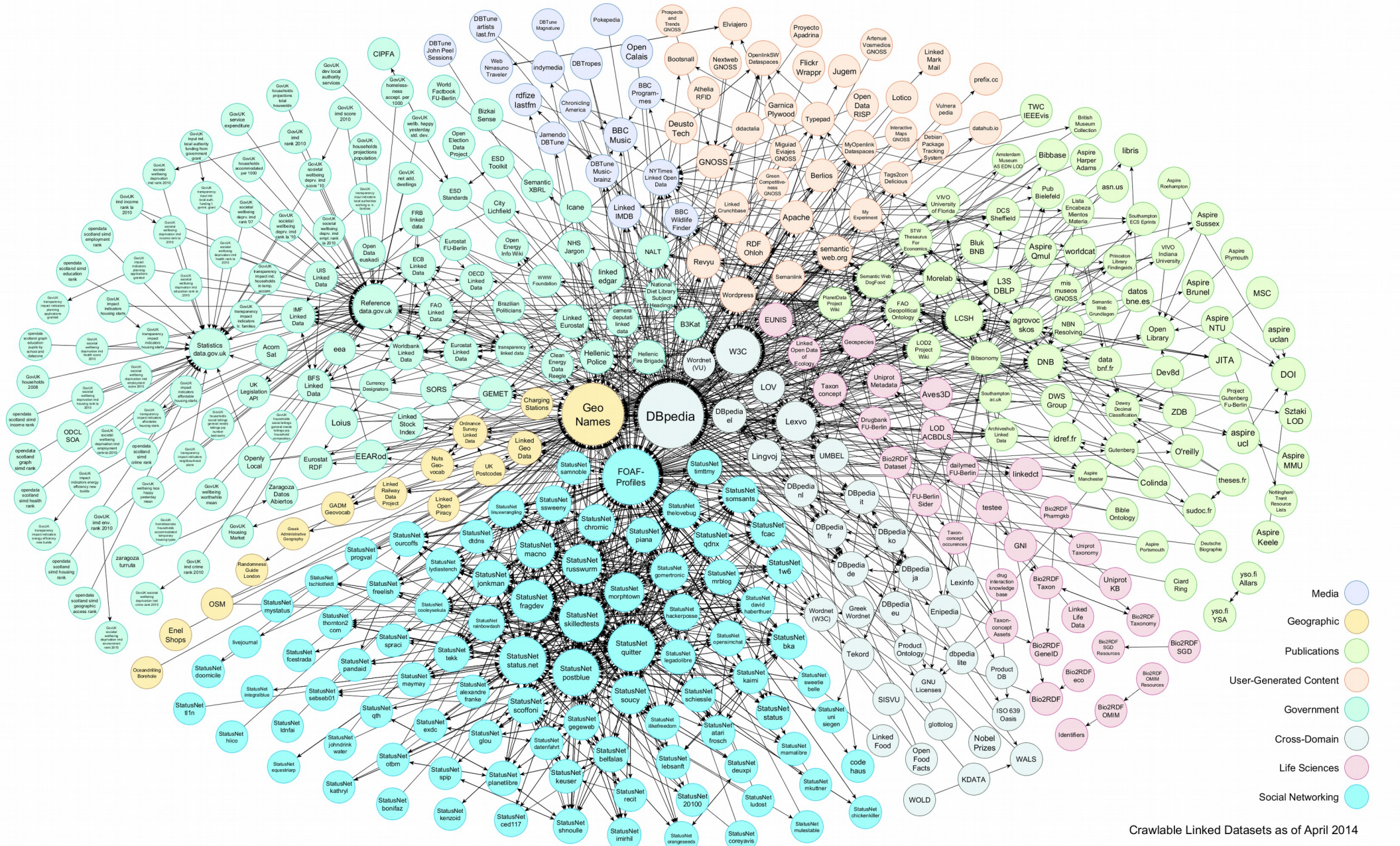
LOD Cloud, 2018

Basic Statistics

Criterion	Average	Min	Max	Median	Total
Triples	67,544.15	0	47,054,407	337.0	192,230,648
Entities	18,105.28	0	9,319,918	80.0	54,225,309
Literals	30,137.45	0	31,476,008	166.0	90,261,655
Blanks	3,554.83	0	3,565,513	0.0	10,646,711
Blanks as subject	1,742.85	0	1,910,532	0.0	5,219,831
Blanks as object	1,812.01	0	3,564,789	0.0	5,426,969
Subclasses	1.6	0	2,000	0.0	4,779
Typed subjects	7,387.12	0	6,990,722	39.0	22,124,421
Labeled subjects	1,219.97	0	1,440,595	0.0	3,653,811
Average properties per entity	4.98	0.0	91.16	3.71	
Average string length typed	13.28	0.0	436.0	0.0	
Average string length untyped	391.77	0.0	181,576.0	10.0	
Average class hierarchy depth	3.24	1	9	None	
Links	15,379.59	0	13,252,430	57.0	46,061,873
Average property hierarchy depth	1.5	1	3	None	
Vocabularies	4.27	1	18	3.0	12,110
Classes	4.36	1	330	3.0	10,384
Properties	17.58	1	254	16.0	49,916

Cumulative numbers

9960 datasets

149,423,660,620 triples from **2973 datasets** (192,230,648 triples from **2838 dumps**, 149,231,429,972 from **151 datasets via SPARQL**)

Problems with **6971 datasets** (70.1%): **6578 dumps having errors**, **393 SPARQL endpoints with errors**

Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated
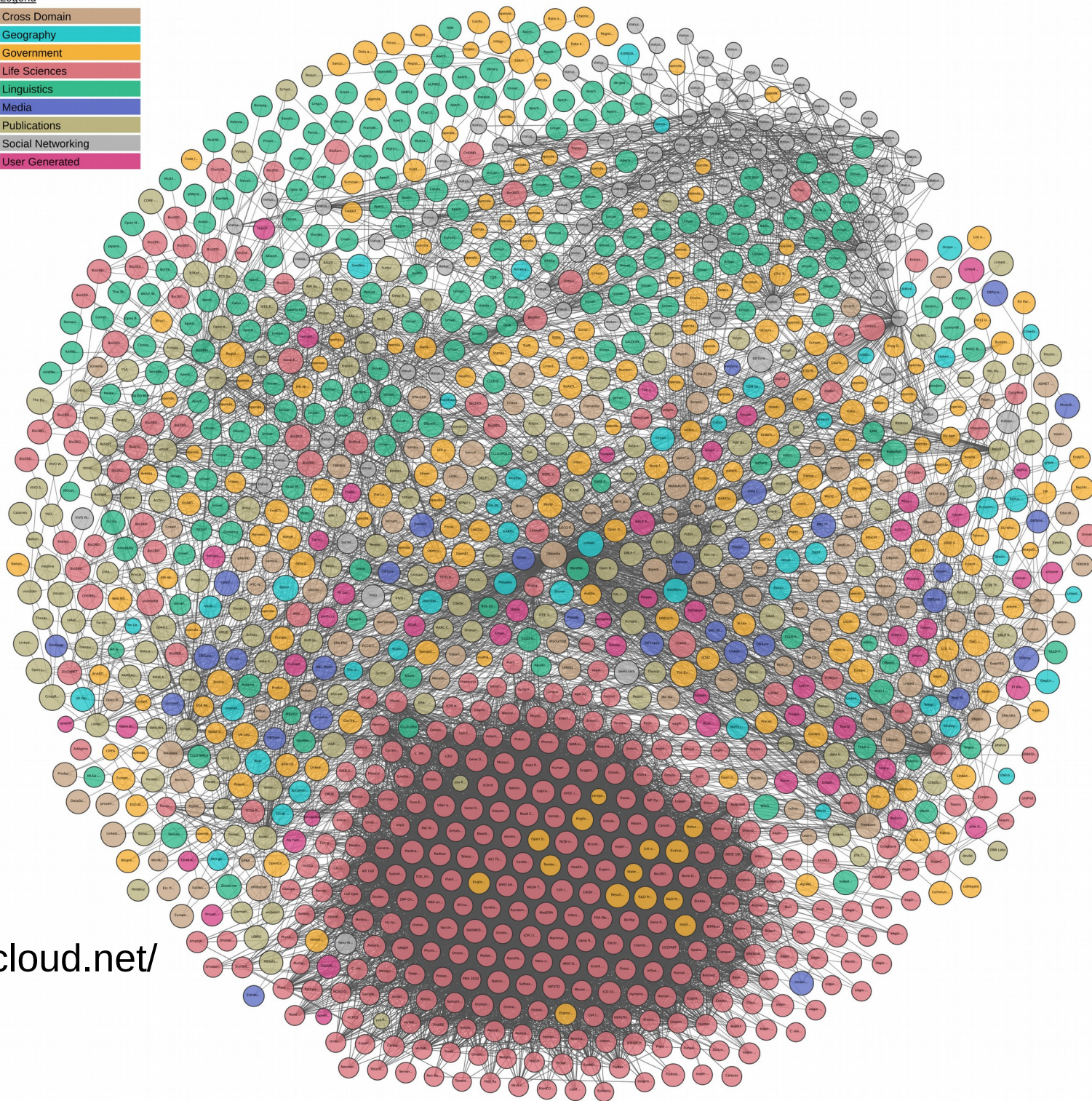
http://lod-cloud.net/

Open Data

🏠 / Organizations / Europeana Foundation / Europeana Linked Open Data

Europeana Linked Op en Data

Followers

3

🏛 **Organization**

Europeana Foundatio n

Europeana is Europe's platform for digital cultural heritage. We create new ways for people to engage with their cultural history, whether for work, learning or

⊹ Dataset   👥 Groups   🕐 Activity Stream

Europeana Linked Open Data

All Europeana datasets can be explored, accessed and downloaded through the Europeana API at http://labs.europeana.eu. The data is represented in the Europeana Data Model (EDM). The described resources are each addressable and dereferenceable by their URIs; for instance, http://data.europeana.eu/item/92056/BD9D5C6C6B02248F187238E9D7CC09EAF17BEA59 leads either to an HTML page on the Europeana portal for the object it identifies or to raw, machine-processable data on this object. The data can be downloaded in JSON or RDF.

⊕ **Download Data Package**

Data and Resources

EDM schema
The data structured using the Europeana Data Model (E

ⓘ **More information**   ☐ **Go to resource**

VOID description for data.europeana.eu
No description for this resource

ⓘ **More information**   ☐ **Go to resource**

Europe   Europeana   culturalheritage   format-edm   format-json   format-rdf   licence-metadata   lld   lod   lodcloud-diagram-20...   lodcloud-diagram-20...   lodlam

Additional Info

Field	Value
Source	http://data.europeana.eu/.well-known/void.ttl
Author	Valentine Charles

Freebase

- Free, knowledge graph:
  - people, places and things,
  - 3,041,722,635 facts, 49,947,845 topics
- Semantic search engines are here !

Freebase

- Based on graphs:
  - nodes, links, types, properties, namespaces
- Google use of Freebase
  - Knowledge graph
  - Words become concepts
  - Semantic questions
  - Semantic associations
  - Browsing knowledge
  - Knowledge engine
- Available in RDF

Knowledge graph

- Google's Knowledge Graph
  - 70 billion facts, oct 2016
  - Box to the right of search results, since 2012
  - Google Assistant and Google Home voice queries
- Knowledge Vault, Google, 2014
  - Initiative to succeed the capabilities of the Knowledge Graph
    - … to deal with facts, automatically gathering and merging information from across the Internet into a knowledge base capable of answering direct questions, such as "Where was Madonna born"
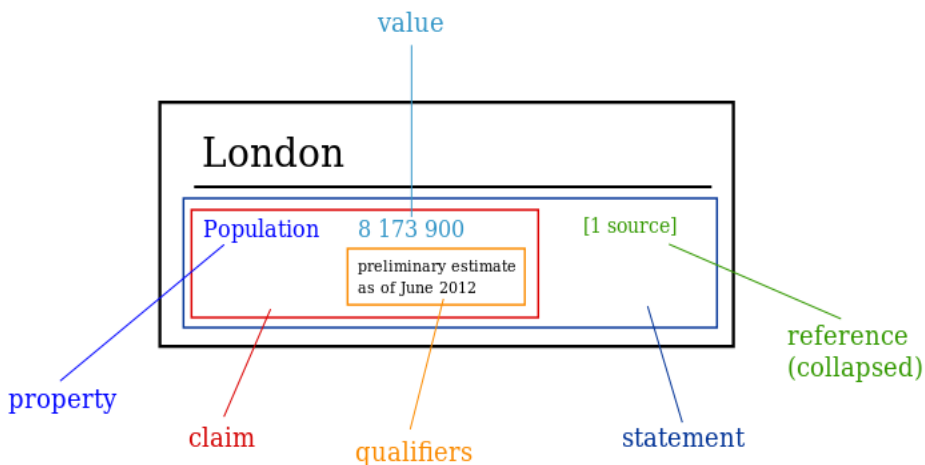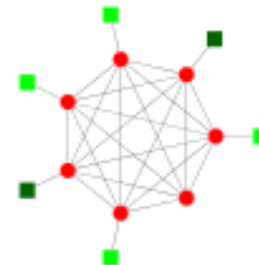
YAGO

- 10 Mega (10^6) concepts
  - 120M facts about these entities
  - Max Planc Institute, Informatik
  - Accuracy of 95%
- Includes:
  - Wikipedia, WordNet, GeoNames
  - Links Wordnet to Wikipedia taxonomy (350K concepts)
  - Anchored in time and space

Wikidata

- Free knowledge base with 46,769,977 items
  - 14,913,910 - 2015
- Collecting structured data
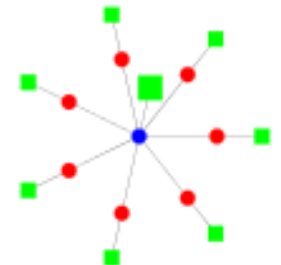- Properties of
  - person, organization,

London
Population 8 173 900
preliminary estimate as of June 2012
[1 source]

value
property
claim
qualifiers
statement
reference (collapsed)

Former system:
interwiki links
between all languages

Phase 1 of Wikidata:
links of all languages
to one central point

Former system:
Independent information about
infoboxes in all languages

Phase 2 of Wikidata:
Information for infoboxes
of all languages
on one central point

Wikidata

label ——— Douglas Adams (Q42) ————— item identifier

English writer and humorist
description ——— Douglas Noël Adams | Douglas Noel Adams ————— aliases
► In more languages

Statements

property ——— educated at   St John's College ————— value

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

——— qualifiers

rank ———

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

——— opened references

statement group ———

+ add reference

Brentwood School

end time	1970
start time	1959

► 0 references ——— collapsed reference

+ add (statement)

This diagram shows the most important terms used in Wikidata

Wikidata

- 2015 - Free knowledge base with 14,550,852 items
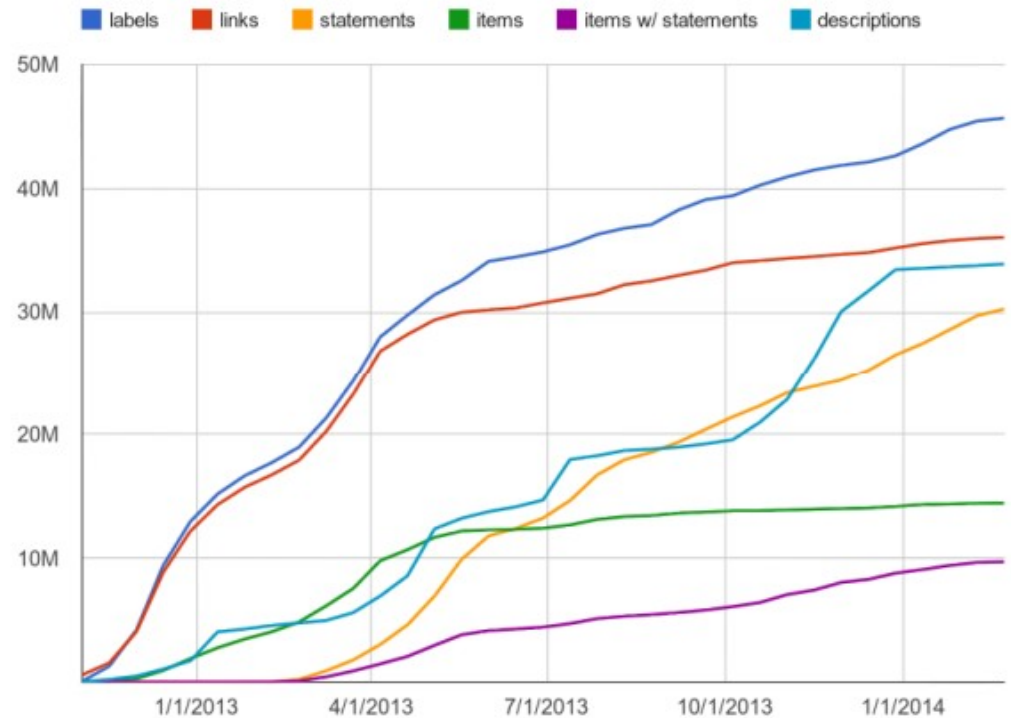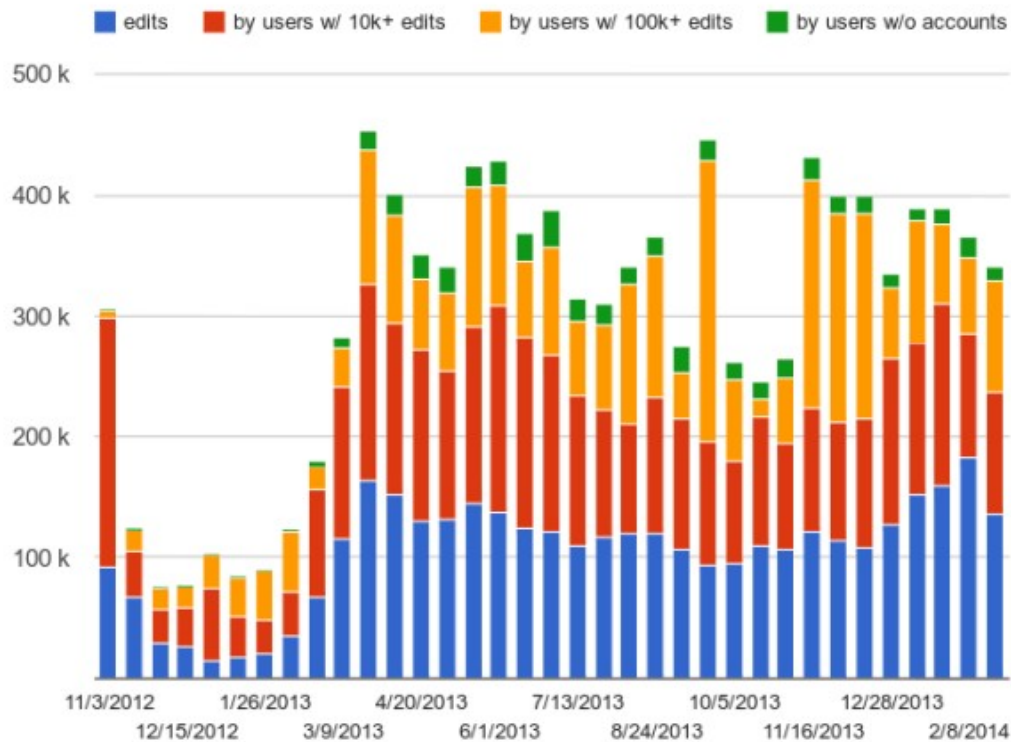
Wikidata

- As of November 2018, Wikidata information is used in 58.4% of all English Wikipedia articles
  - Wikidata is shown in
    - 64% of all Wikipedias' pages,
    - 93% of all Wikivoyages,
    - 34% of all Wikiquotes,
    - 32% of all Wikisources, and
    - 27% of Wikimedia Commons

Wikidata

- Currently contains 53,147,530 items.

- 811,351,583 edits since the project launch.

- There are currently 20,106 active users

- July, 2018

Wikidata (44,316,713)

- ■ human: 4,378,742 (9.9%)
- ■ taxon: 2,509,963 (5.7%)
- ■ administrative territorial entity: 1,812,457 (4.1%)
- ■ architectural structure: 2,119,774 (4.8%)
- ■ occurrence: 1,651,237 (3.7%)
- ■ chemical compound: 759,000 (1.7%)
- ■ film: 268,609 (0.6%)
- ■ thoroughfare: 522,260 (1.2%)
- ■ astronomical body: 155,105 (0.3%)
- ■ Wikimedia list article: 299,810 (0.7%)
- ■ Wikimedia disambiguation page: 1,280,367 (2.9%)
- ■ Wikinews article: 211,930 (0.5%)
- ■ scholarly article: 18,771,018 (42.4%)
- ■ other P31/P279: 5,602,972 (12.6%)
- ■ no P31/P279: 3,973,469 (9.0%)

Module:Statistical data/by project/classes, 2018-07-16

Webdata Query Service

Webdata Query Service

Cyc - knowledge base

- Knowledge base

  - Doug Lenat

  - Conceptual networks (ontologies)

  - Higher ontology, basic theories, specific theories

  - Predefined semantic relationships

  - 500.000 terms, including about 17.000 types of relations, and about 7.000.000 assertions relating these terms

- Common sense reasoner

  - Based on predicate calculus

  - Rule-based reasoning

Cyc
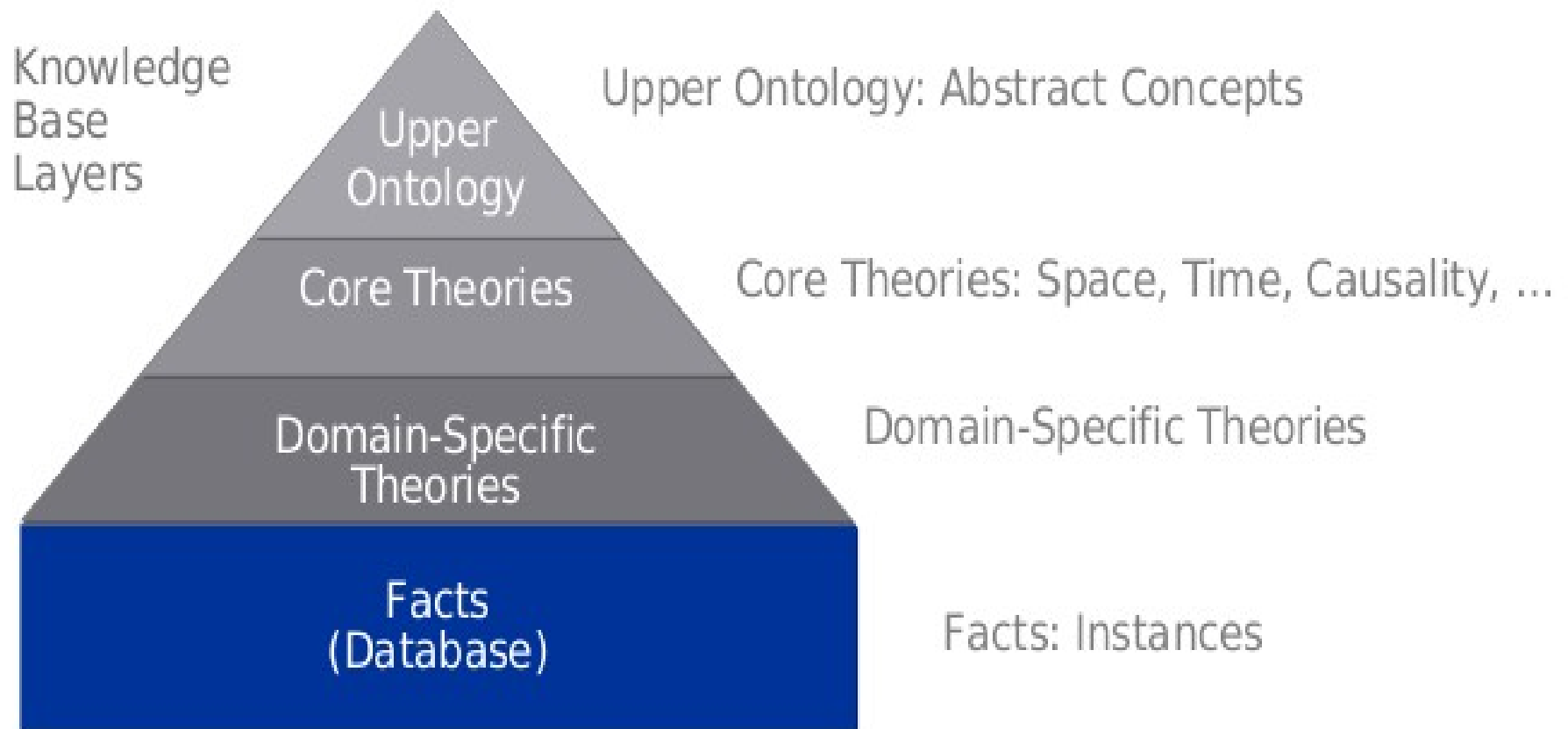
Stolp semantičnega spleta
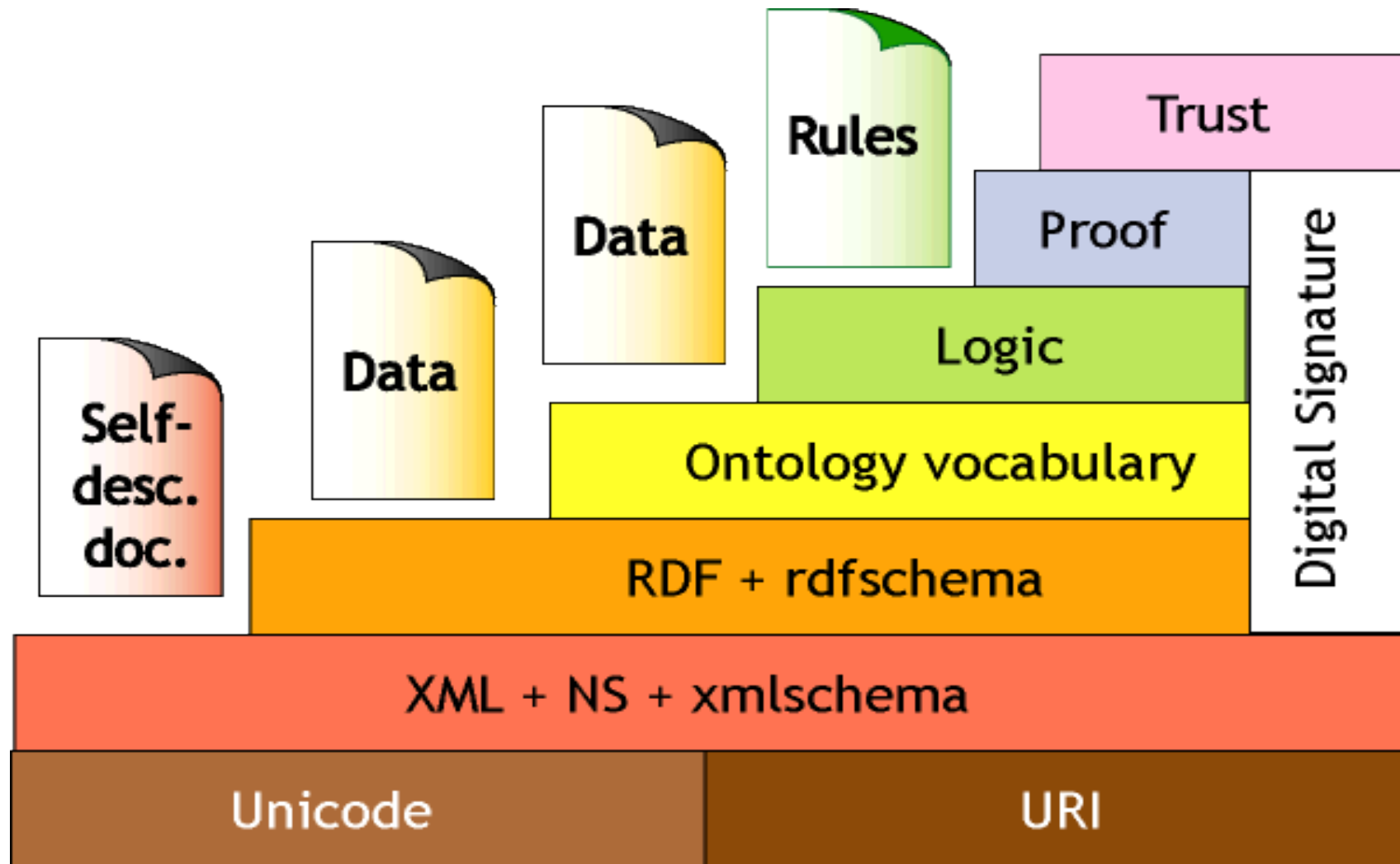
Linked data and applications

Linked data

1. Od spleta dokumentov do spleta podatkov
   - Spletni API-ji in Linked Data
2. Linked Data implementacija na spletu
   - Kateri podatki obstajajo?
3. Aplikacije
   - Kaj se dogaja s podatki?
4. Naslednji koraki
   - Kaj manjka?

Klasični splet

En sam globalen inform. prostor

1. URL za:
   - Globalni unikatni IDs
   - Poizvedovalni mehanizmi
2. HTML kot skupna oblika vsebine
3. Hyper-povezave

Web Browsers

Search Engines

HTML   HTML   HTML

hyper-links

A   B   C
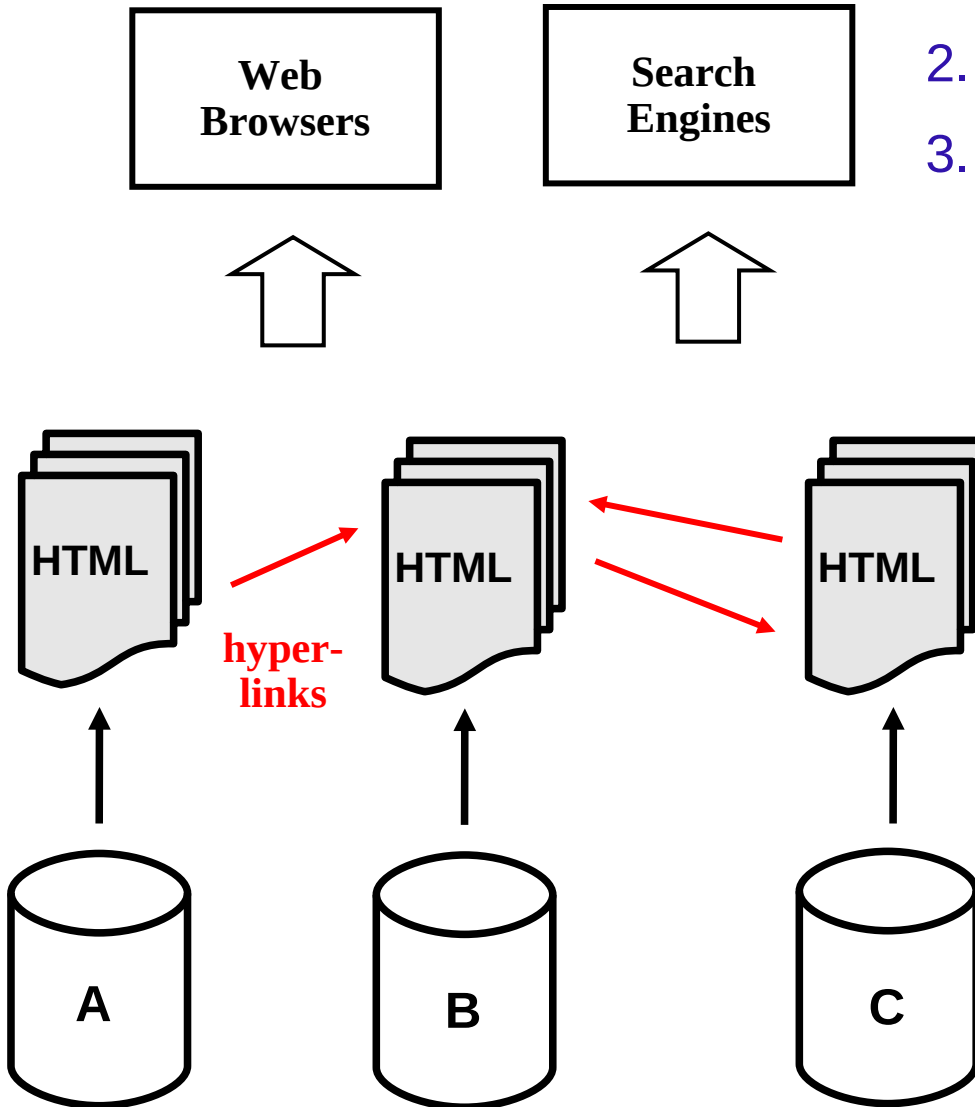
Problem in rešitev

Problem

Ker je vsebina spleta zelo šibko strukturirana, aplikacije težko implementirajo pametne operacije.

Rešitev

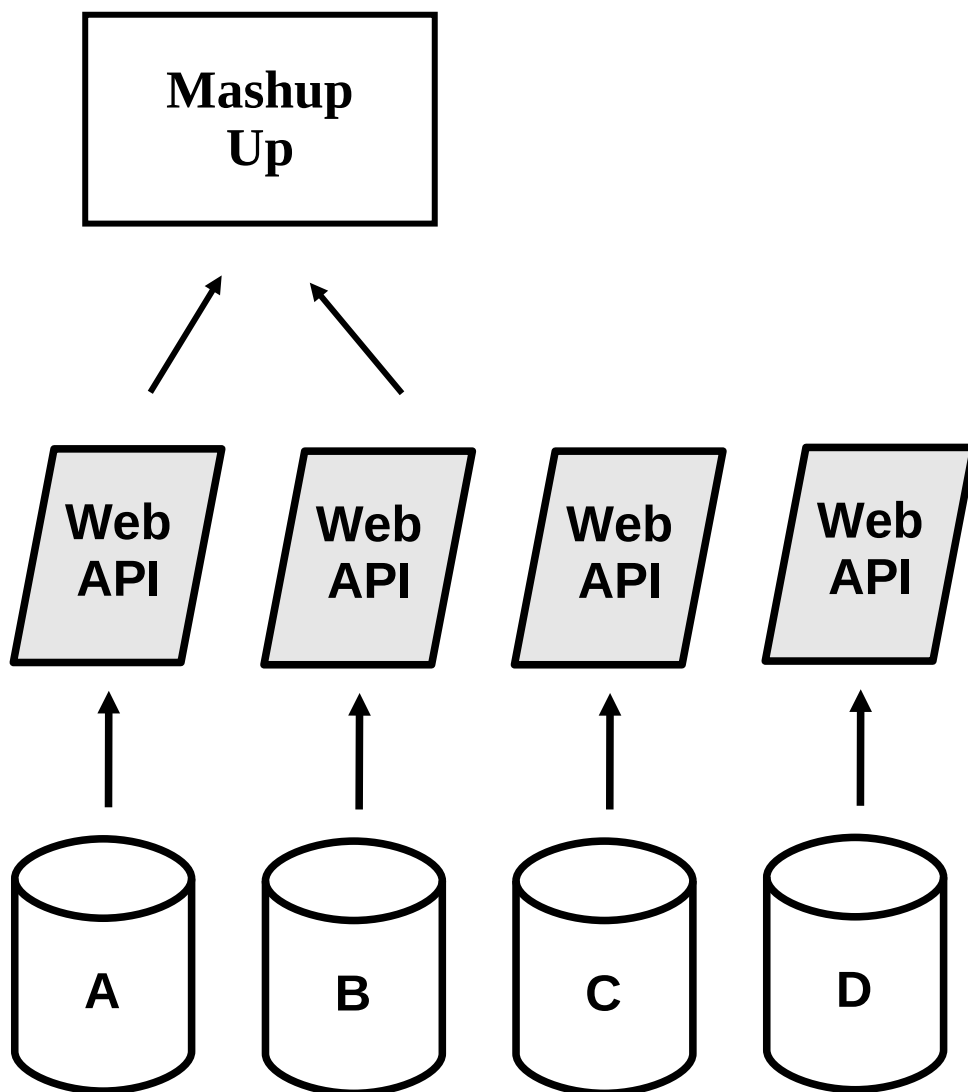Povečaj strukturo vsebine spleta.

Spletni API-ji in prepletene storitve

Spletni API-ji in mashup-i

Slabe lastnosti:

0) API-ji nudijo privatne vmesnike.

1) Mashup-i temeljijo na fiksni množici podatkovnih virov.

2) Ne moremo definirati povezav med podatkovnimi objekti.

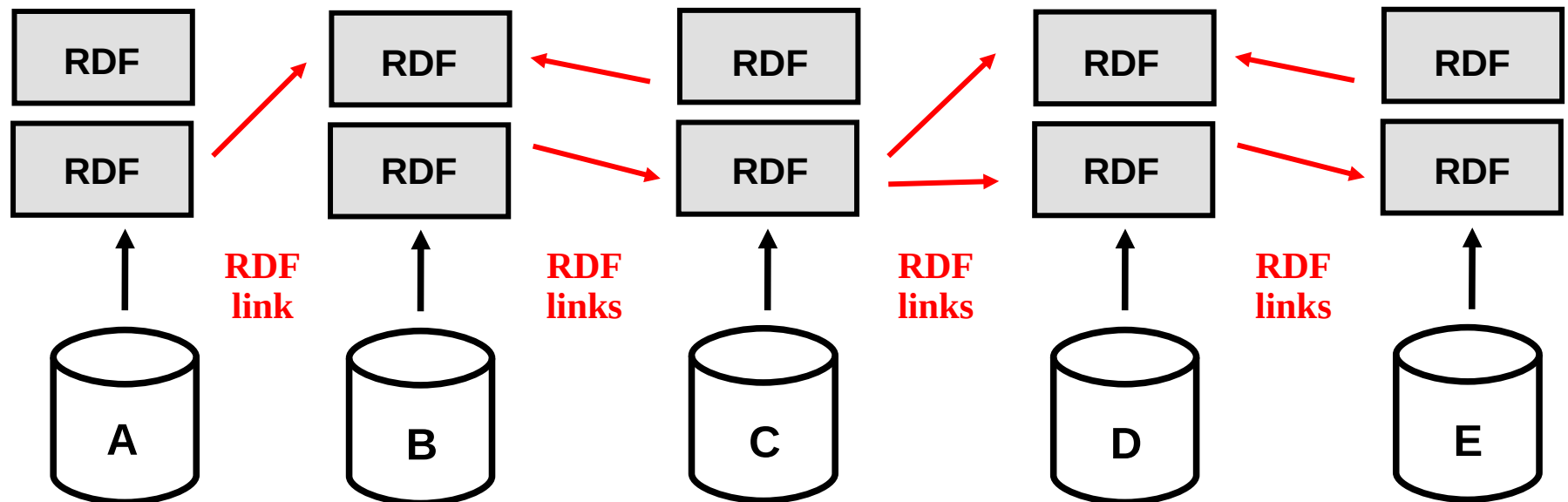# Spletni API-ji razdelijo splet na vrtove z ograjami
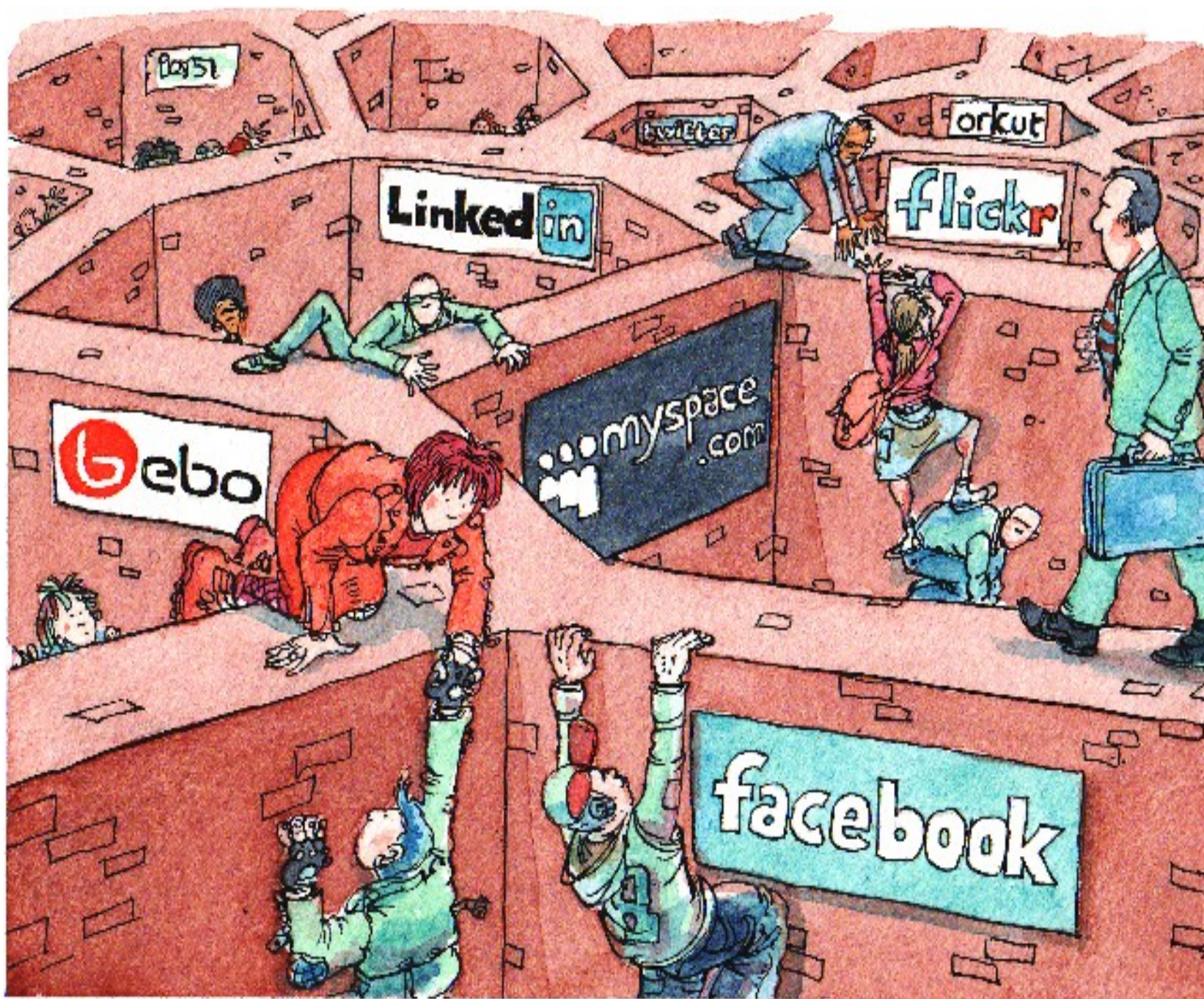
Povezani podatki

Uporabi tehnologije semantičnega spleta za:
- Publiciranje strukturiranih podatkov na splet,
- Definicijo povezav med podatki iz enega spletnega mesta s podatki v drugih spletnih mestih.

Principi povezanih podatkov

1. Uporabi URI-je kot imena za objekte

2. Uporabi HTTP URI, da lahko ljudje dostopajo do teh imen

3. Ko nekdo dostopa do URI zagotovi uporabne RDF podatke

4. Vključi RDF stavke drugih URI, da lahko ljudje odkrijejo sorodne objekte.

Tim Berners-Lee 2007

http://www.w3.org/DesignIssues/LinkedData.html

Lastnosti spleta povezanih podatkov

- Kdorkoli lahko publicira podatke na spletu povezanih podatkov

- Entitete so povezane s povezavami

  - Kreacija globalnega podatkovnega grafa, ki povezuje podatkovne vire in omogoča odkrivanje novih virov.

- Podatki so samo-opisni

  - Če aplikacija dobi podatke, ki so predstavljeni z nepoznanim besednjakom, mora aplikacija identificirati URI-je, ki identificirajo slovarje z definicijami RDFS in OWL izrazov.

- Splet podatkov je odprt

  - To pomeni, da lahko aplikacije odkrivajo nove podatkovne vire v času izvajanja.

Implementacija povezanih podatkov na spletu

- **Je to realno?**
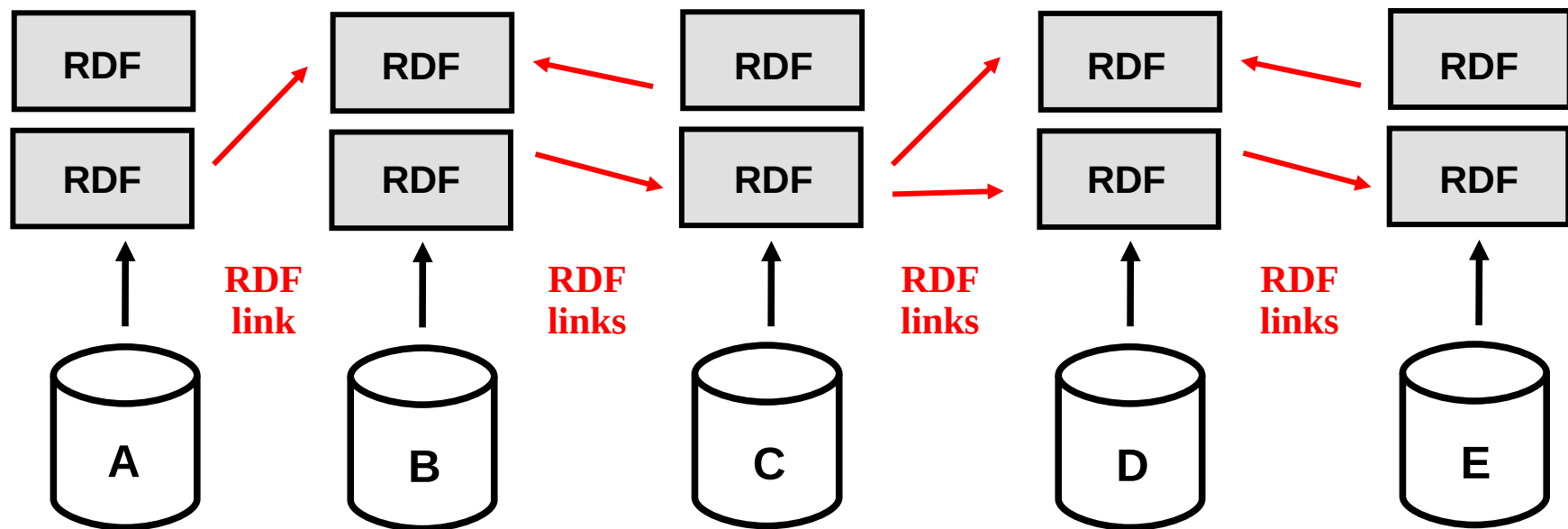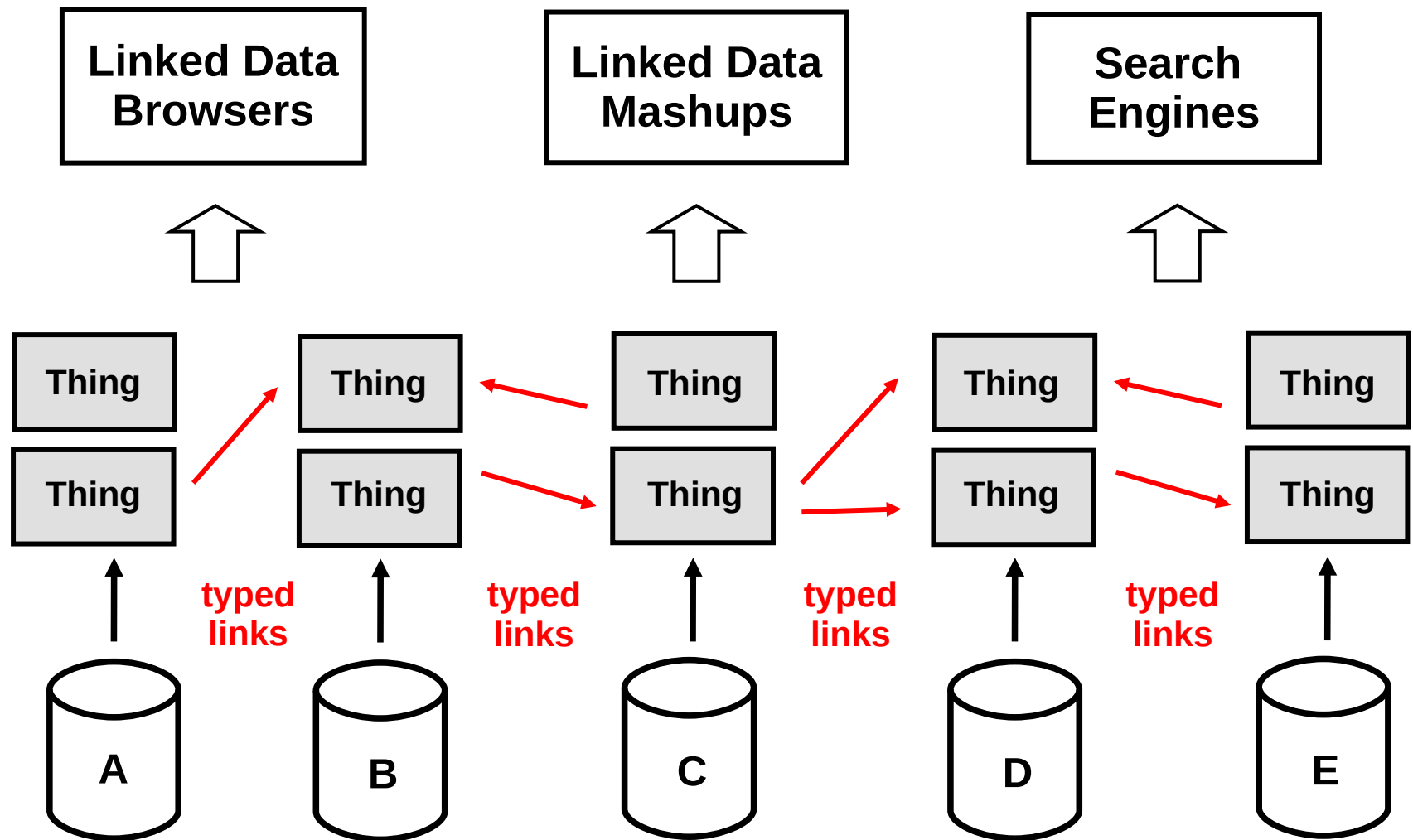
W3C Linking Open Data

■ **Skupnost:**

- Municiranje obstoječih odprtih podatkovnih zbirk na spletu.
- Povezovanje stvari med različnimi podatkovnimi viri

Aplikacije

- Kaj lahko naredimo s tem?

Povezani podatki v prepletenih storitvah (meshups)

Aplikacije s specifično domeno, ki uporabljajo povezane podatke iz spleta

Podatkovna fuzija

Aplikacija

Inegrated view
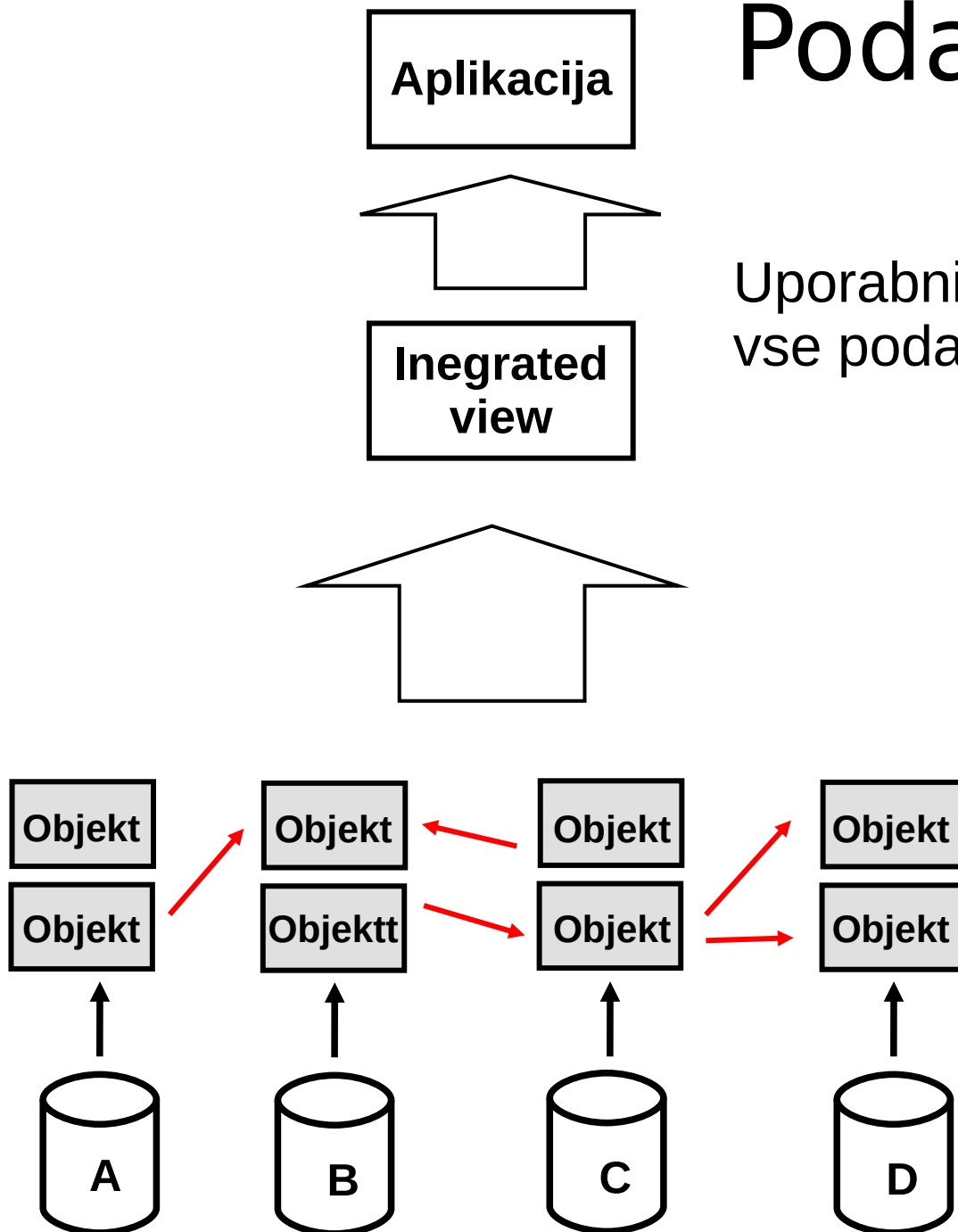
Uporabnik želi imeti integriran pogled na vse podatke, ki so dostopni za objekt.

Znani problemi:

Preslikave shem
Reševanje nekonsistentnosti

Objekt **Objekt** **Objekt** **Objekt**

Objekt **Objektt** **Objekt** **Objekt**

A B C D

Brskalnik za povezane podatke (stari podatki!)

- **Tabulator Browser (MIT, USA)**

- **Marbles (FU Berlin, DE)**

- **OpenLink RDF Browser (OpenLink, UK)**

- **Zitgist RDF Browser (Zitgist, USA)**

- **Humboldt (HP Labs, UK)**

- **Disco Hyperdata Browser (FU Berlin, DE)**

- **Fenfire (DERI, Irland)**

DBpedia Mobile

Geospatial vhodna točka na splet podatkov

Uporablja:
DBpedia, Revyu in Flickr podatke

Podatkovni splet in iskalniki
(stari podatki!)

Falcons (IWS, China)

http://iws.seu.edu.cn/services/falcons/documentsearch/

Sig.ma (DERI, Ireland)

http://www.deri.ie/, http://sig.ma/

Swoogle (UMBC, USA)

http://swoogle.umbc.edu/

Watson (Open University, UK)

http://watson.kmi.open.ac.uk/WatsonWUI/