Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Extreme Web Data Integration

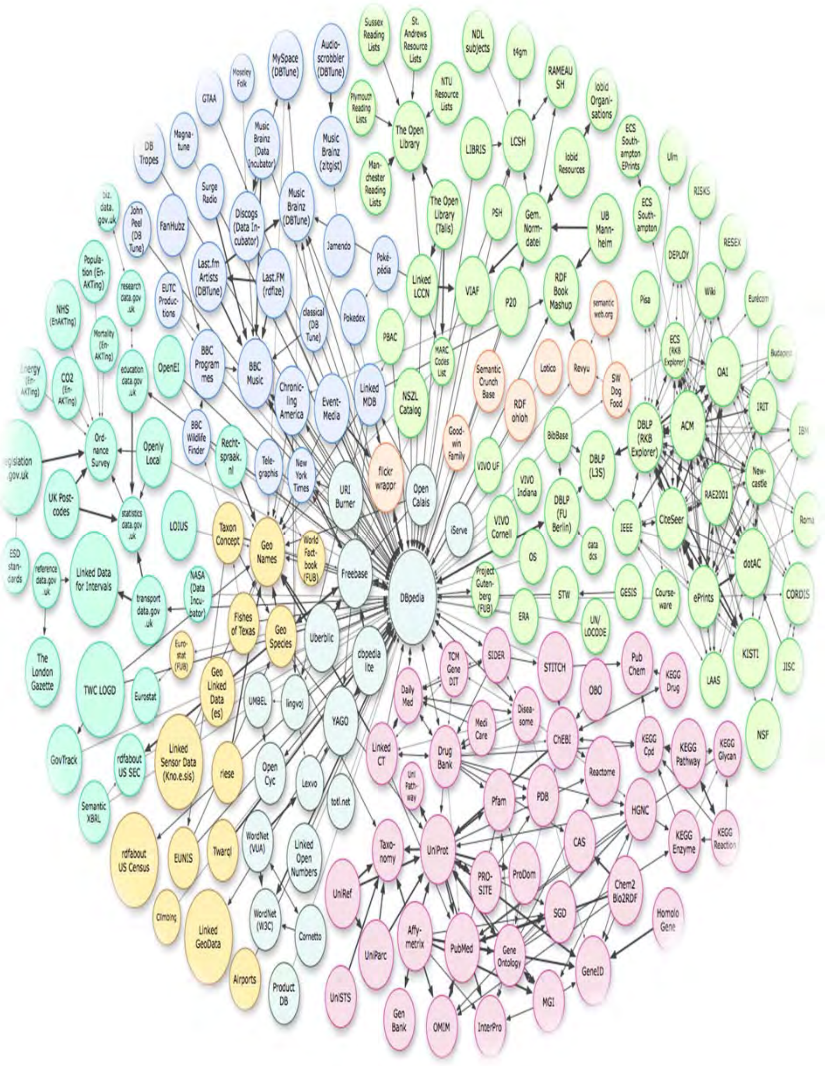
Keynote @ ICWE 2012

26.7.2012

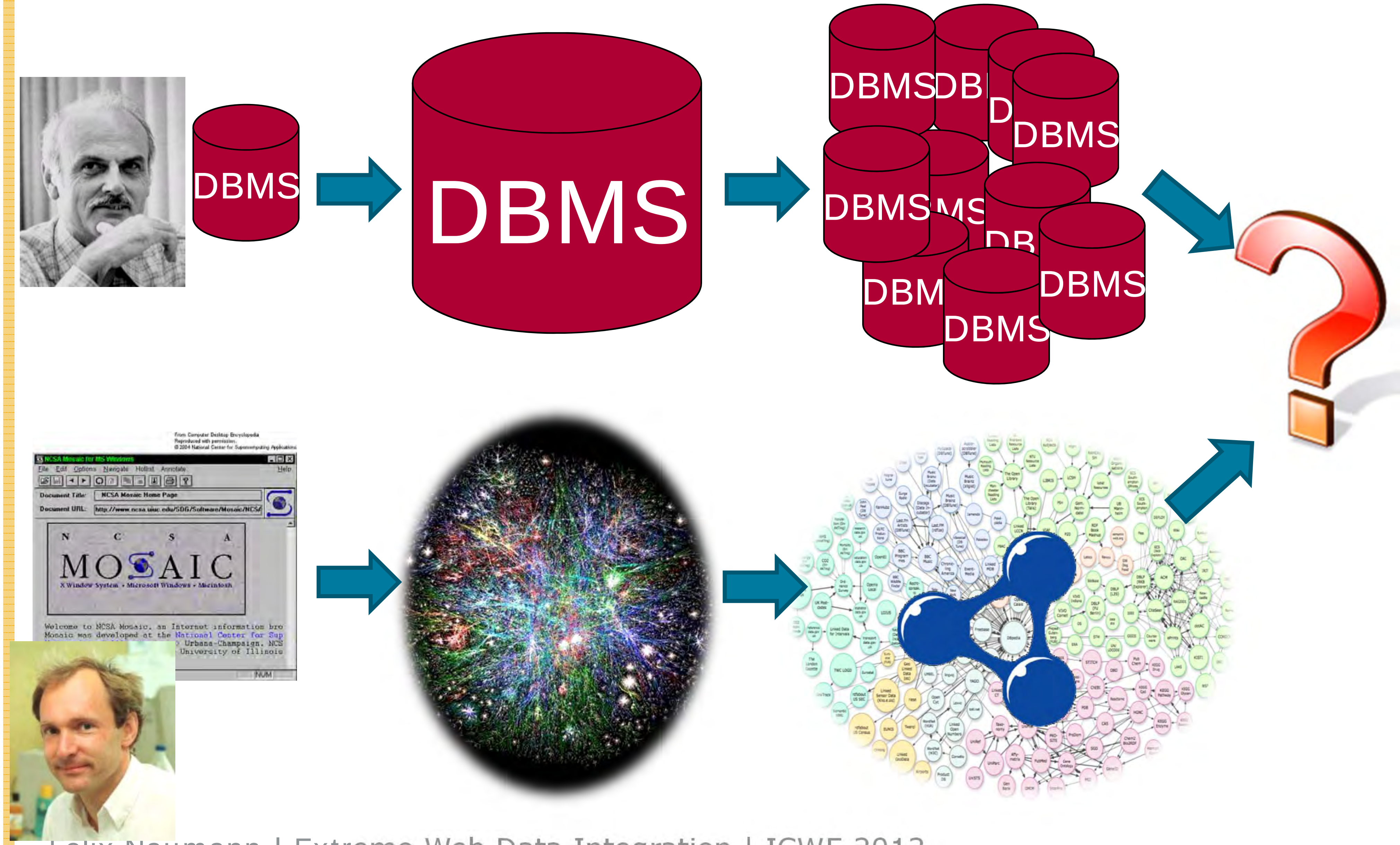
Felix Naumann

Overview

- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



A brief history of data


DBMS
DBMS
DBMS
DB
D
DBMS
DBMS
MS
DB
DBM
DBMS
DBMS
?

From Computer Desktop Encyclopedia
Reproduced with permission.
© 2004 National Center for Supercomputing Applications

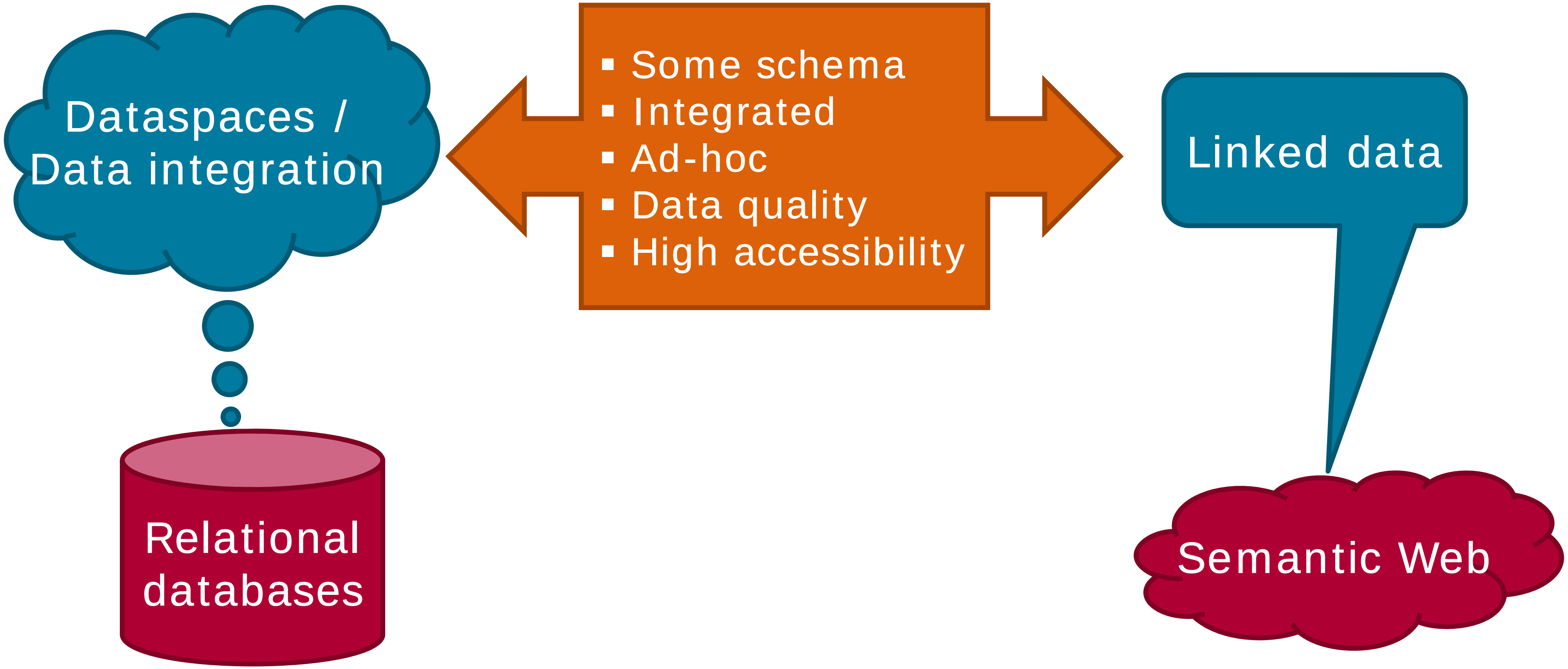
NCSA Mosaic for MS Windows
File Edit Options Navigate Hotlist Annotate Help

Document Title: NCSA Mosaic Home Page
Document URL: http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCS

N C S A
MOSAIC
X Window System • Microsoft Windows • Macintosh

Welcome to NCSA Mosaic, an Internet information bro
Mosaic was developed at the National Center for Sup
Urbana-Champaign. NCS
University of Illinois

Linked Data & Data Spaces: A database guy‘s point-of-view

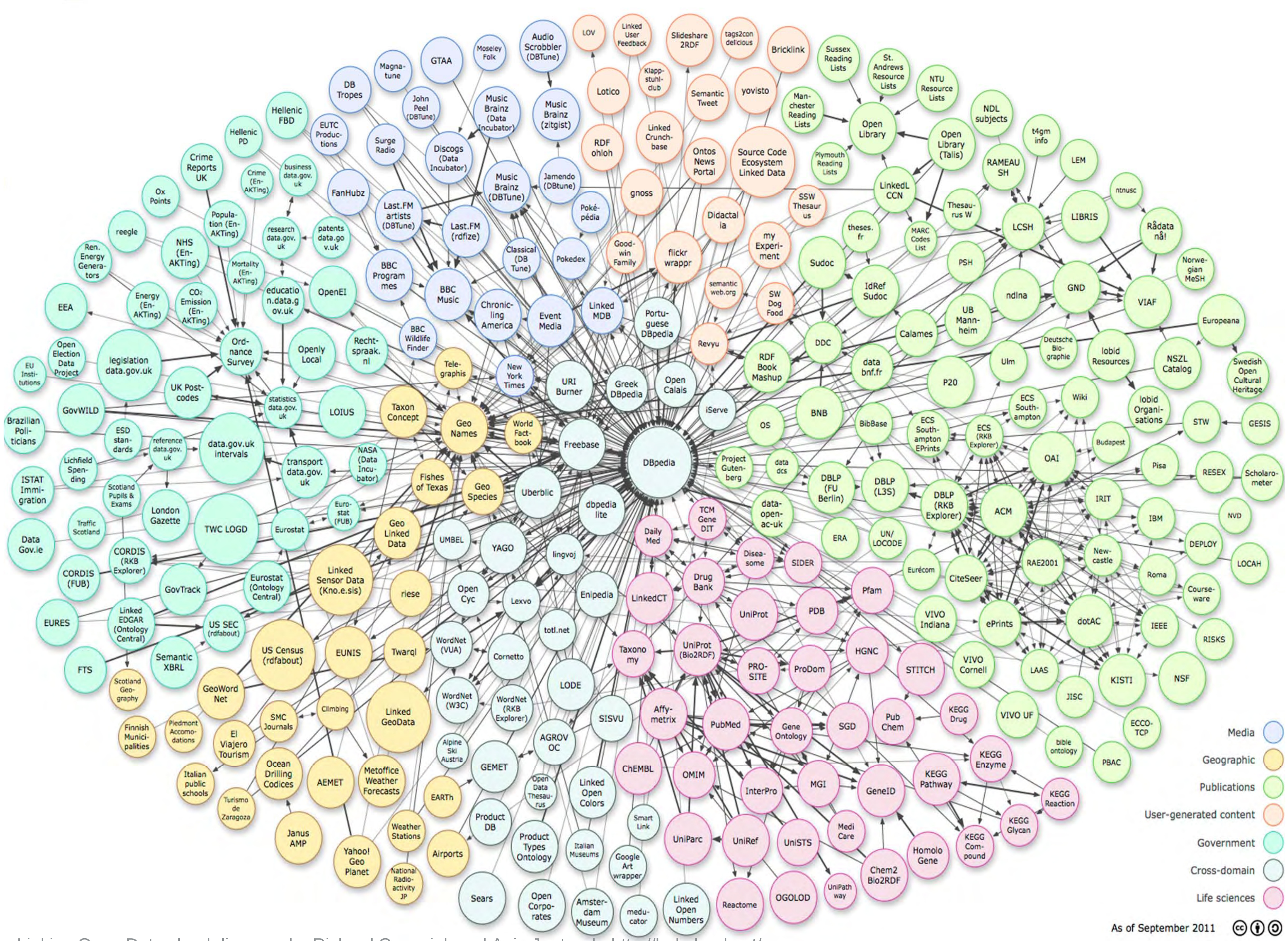


- Some schema
- Integrated
- Ad-hoc
- Data quality
- High accessibility

Linked data – 4 Principles, 7 Properties

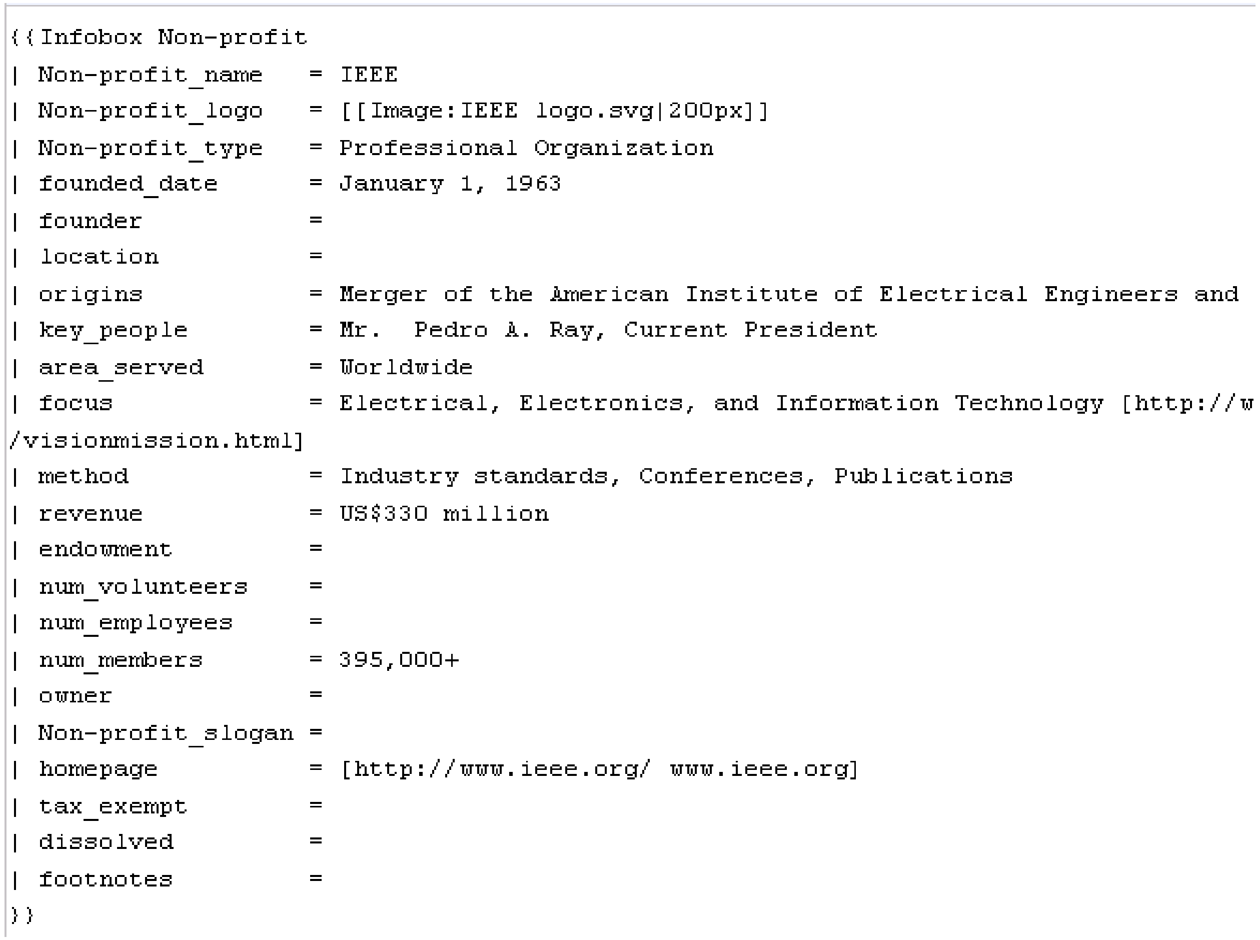
1. Use **URIs as names** for things.
2. Use **HTTP URIs** so that people can look up those names.
3. When someone looks up a URI, **provide useful information**.
4. Include **links to other URIs**, so that they can discover more things.
   - Many common things are represented in multiple data sets!

- The Good
  - Comes as triples
    S: http://.../Berlin
    P: location
    O: http://.../Germany
  - Often user generated
  - Nice domains
  - Free
- The Bad
  - Voluminous
  - Heterogeneous
- The Ugly
  - Dirty, inconsistent, sparse

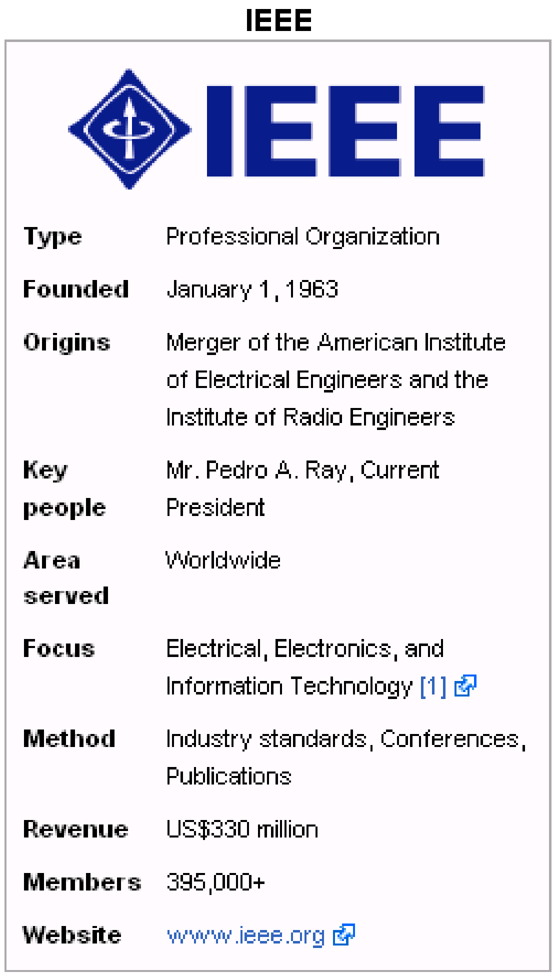


Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

DBpedia – Extraction



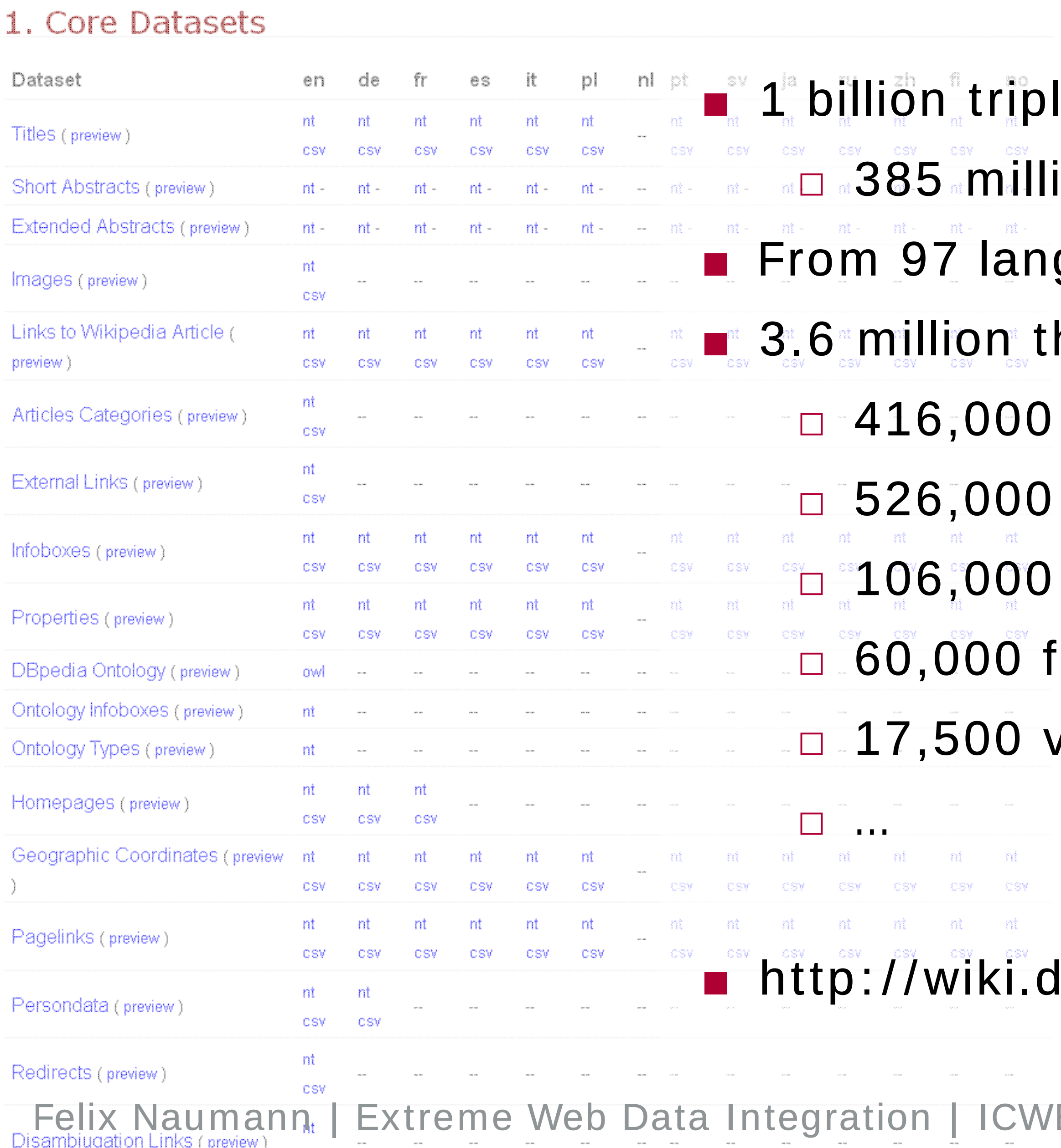
```
{{Infobox Non-profit
| Non-profit_name    = IEEE
| Non-profit_logo    = [[Image:IEEE logo.svg|200px]]
| Non-profit_type    = Professional Organization
| founded_date       = January 1, 1963
| founder            =
| location           =
| origins            = Merger of the American Institute of Electrical Engineers and
| key_people         = Mr.  Pedro A. Ray, Current President
| area_served        = Worldwide
| focus              = Electrical, Electronics, and Information Technology [http://w
/visionmission.html]
| method             = Industry standards, Conferences, Publications
| revenue            = US$330 million
| endowment          =
| num_volunteers     =
| num_employees      =
| num_members        = 395,000+
| owner              =
| Non-profit_slogan =
| homepage           = [http://www.ieee.org/ www.ieee.org]
| tax_exempt         =
| dissolved          =
| footnotes          =
}}
```



IEEE

Type	Professional Organization
Founded	January 1, 1963
Origins	Merger of the American Institute of Electrical Engineers and the Institute of Radio Engineers
Key people	Mr. Pedro A. Ray, Current President
Area served	Worldwide
Focus	Electrical, Electronics, and Information Technology [1]
Method	Industry standards, Conferences, Publications
Revenue	US$330 million
Members	395,000+
Website	www.ieee.org

DBpedia statistics



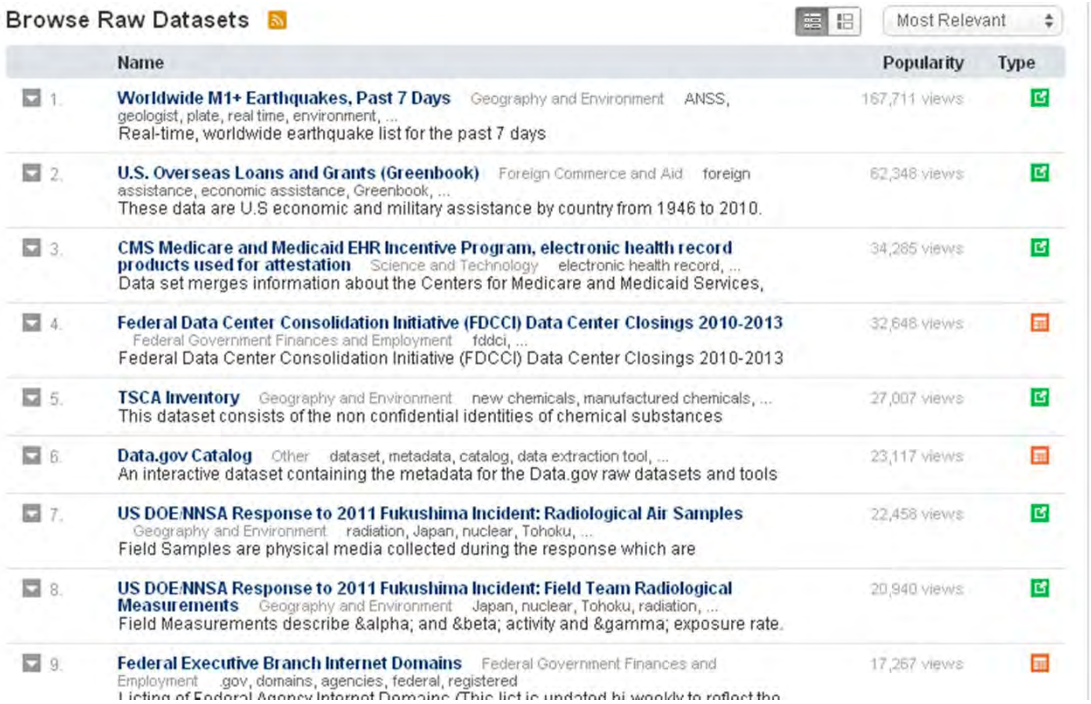
1. Core Datasets

Dataset	en	de	fr	es	it	pl	nl	pt	sv	ja	ru	zh	fi	no
Titles (preview)	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	--	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv
Short Abstracts (preview)	nt -	nt -	nt -	nt -	nt -	nt -	--	nt -	nt -	nt -	nt -	nt -	nt -	nt -
Extended Abstracts (preview)	nt -	nt -	nt -	nt -	nt -	nt -	--	nt -	nt -	nt -	nt -	nt -	nt -	nt -
Images (preview)	nt csv	--	--	--	--	--	--	--	--	--	--	--	--	--
Links to Wikipedia Article (preview)	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	--	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv
Articles Categories (preview)	nt csv	--	--	--	--	--	--	--	--	--	--	--	--	--
External Links (preview)	nt csv	--	--	--	--	--	--	--	--	--	--	--	--	--
Infoboxes (preview)	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	--	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv
Properties (preview)	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	--	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv
DBpedia Ontology (preview)	owl	--	--	--	--	--	--	--	--	--	--	--	--	--
Ontology Infoboxes (preview)	nt	--	--	--	--	--	--	--	--	--	--	--	--	--
Ontology Types (preview)	nt	--	--	--	--	--	--	--	--	--	--	--	--	--
Homepages (preview)	nt csv	nt csv	nt csv	--	--	--	--	--	--	--	--	--	--	--
Geographic Coordinates (preview)	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	--	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv
Pagelinks (preview)	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	--	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv	nt csv
Persondata (preview)	nt csv	nt csv	--	--	--	--	--	--	--	--	--	--	--	--
Redirects (preview)	nt csv	--	--	--	--	--	--	--	--	--	--	--	--	--
Disambiguation Links (preview)	nt	--	--	--	--	--	--	--	--	--	--	--	--	--

- 1 billion triples
  - 385 million English
- From 97 languages of Wikipedia
- 3.6 million things
  - 416,000 persons
  - 526,000 places
  - 106,000 music albums
  - 60,000 films
  - 17,500 video games
  - ...
- http://wiki.dbpedia.org/Datasets

And more sources

- Government data
  - □ www.data.gov
450k data sets
  - □ data.gov.uk
  - □ ec.europa.eu/eurostat
- Finance / business data
- Scientific databases
  - □ www.uniprot.org
  - □ skyserver.sdss.org
- The Web
  - □ HTML tables and lists
  - □ General sources: DBpedia, freebase, ...
  - □ Domain-specific sources: IMDB, Gracenote, isbndb, ...



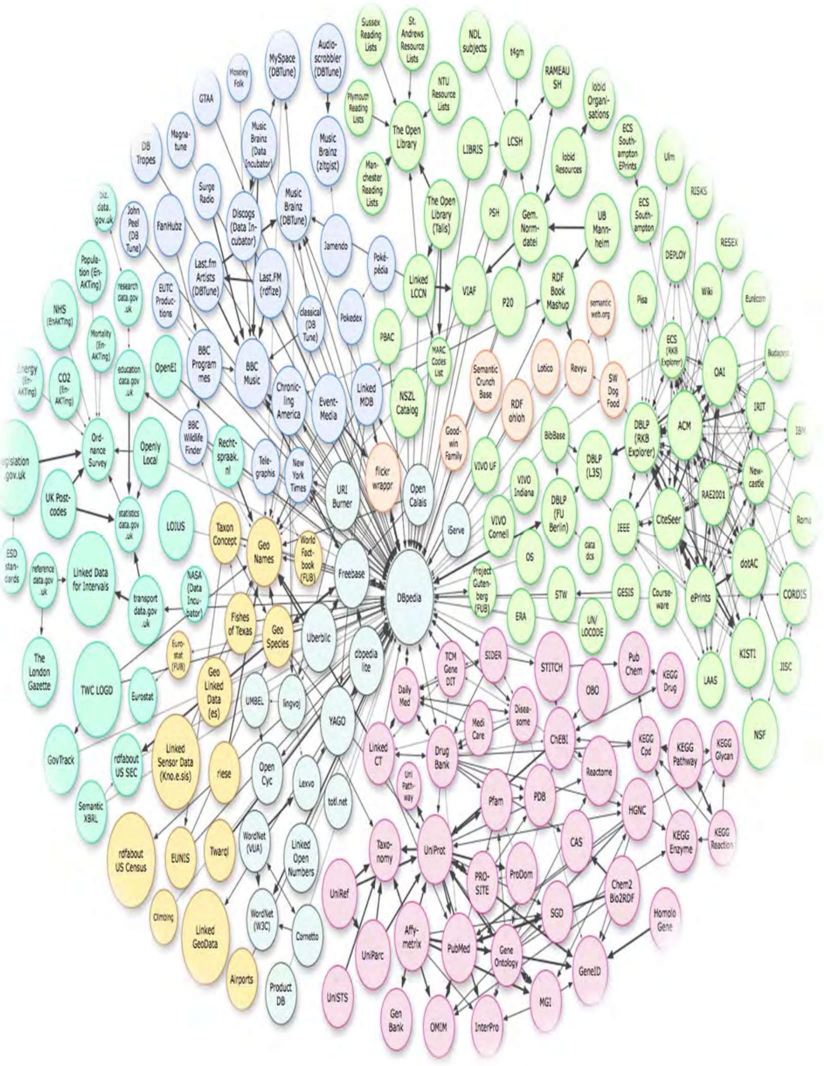
Browse Raw Datasets

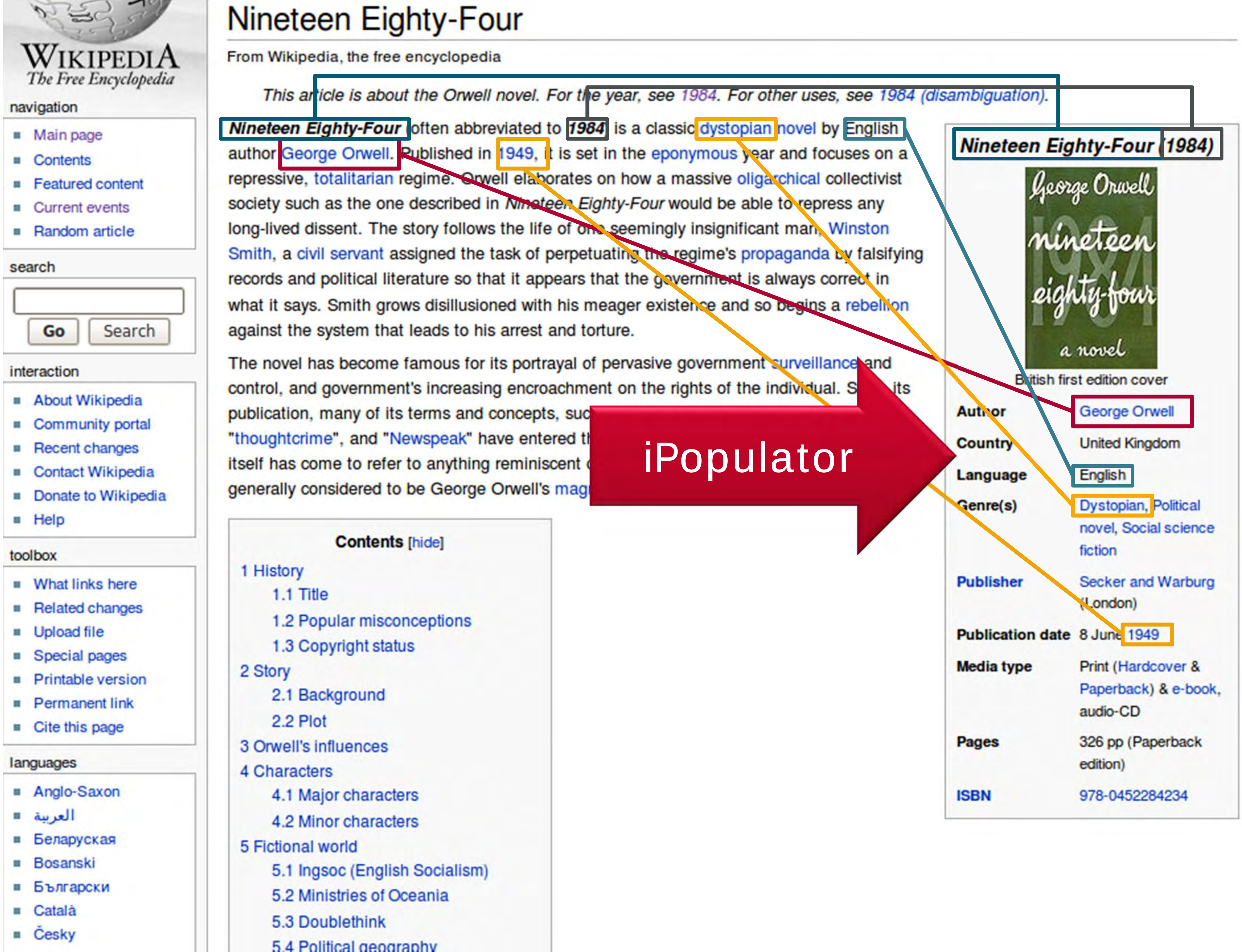
	Name	Popularity	Type
1.	**Worldwide M1+ Earthquakes, Past 7 Days** Geography and Environment ANSS, geologist, plate, real time, environment, ... Real-time, worldwide earthquake list for the past 7 days	167,711 views	↗
2.	**U.S. Overseas Loans and Grants (Greenbook)** Foreign Commerce and Aid foreign assistance, economic assistance, Greenbook, ... These data are U.S economic and military assistance by country from 1946 to 2010.	62,348 views	↗
3.	**CMS Medicare and Medicaid EHR Incentive Program, electronic health record products used for attestation** Science and Technology electronic health record, ... Data set merges information about the Centers for Medicare and Medicaid Services,	34,285 views	↗
4.	**Federal Data Center Consolidation Initiative (FDCCI) Data Center Closings 2010-2013** Federal Government Finances and Employment fddci, ... Federal Data Center Consolidation Initiative (FDCCI) Data Center Closings 2010-2013	32,648 views	▦
5.	**TSCA Inventory** Geography and Environment new chemicals, manufactured chemicals, ... This dataset consists of the non confidential identities of chemical substances	27,007 views	↗
6.	**Data.gov Catalog** Other dataset, metadata, catalog, data extraction tool, ... An interactive dataset containing the metadata for the Data.gov raw datasets and tools	23,117 views	▦
7.	**US DOE/NNSA Response to 2011 Fukushima Incident: Radiological Air Samples** Geography and Environment radiation, Japan, nuclear, Tohoku, ... Field Samples are physical media collected during the response which are	22,458 views	↗
8.	**US DOE/NNSA Response to 2011 Fukushima Incident: Field Team Radiological Measurements** Geography and Environment Japan, nuclear, Tohoku, radiation, ... Field Measurements describe &alpha; and &beta; activity and &gamma; exposure rate.	20,940 views	↗
9.	**Federal Executive Branch Internet Domains** Federal Government Finances and Employment .gov, domains, agencies, federal, registered Listing of Federal Agency Internet Domains (This list is updated bi weekly to reflect the	17,267 views	▦

Killer app?

Overview

- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience





WIKIPEDIA
The Free Encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages

- Anglo-Saxon
- العربية
- Беларуская
- Bosanski
- Български
- Català
- Česky

Nineteen Eighty-Four

From Wikipedia, the free encyclopedia

This article is about the Orwell novel. For the year, see 1984. For other uses, see 1984 (disambiguation).

Nineteen Eighty-Four (often abbreviated to ***1984***) is a classic dystopian novel by English author George Orwell. Published in 1949, it is set in the eponymous year and focuses on a repressive, totalitarian regime. Orwell elaborates on how a massive oligarchical collectivist society such as the one described in *Nineteen Eighty-Four* would be able to repress any long-lived dissent. The story follows the life of one seemingly insignificant man, Winston Smith, a civil servant assigned the task of perpetuating the regime's propaganda by falsifying records and political literature so that it appears that the government is always correct in what it says. Smith grows disillusioned with his meager existence and so begins a rebellion against the system that leads to his arrest and torture.

The novel has become famous for its portrayal of pervasive government surveillance and control, and government's increasing encroachment on the rights of the individual. S its publication, many of its terms and concepts, suc "thoughtcrime", and "Newspeak" have entered t itself has come to refer to anything reminiscent generally considered to be George Orwell's mag

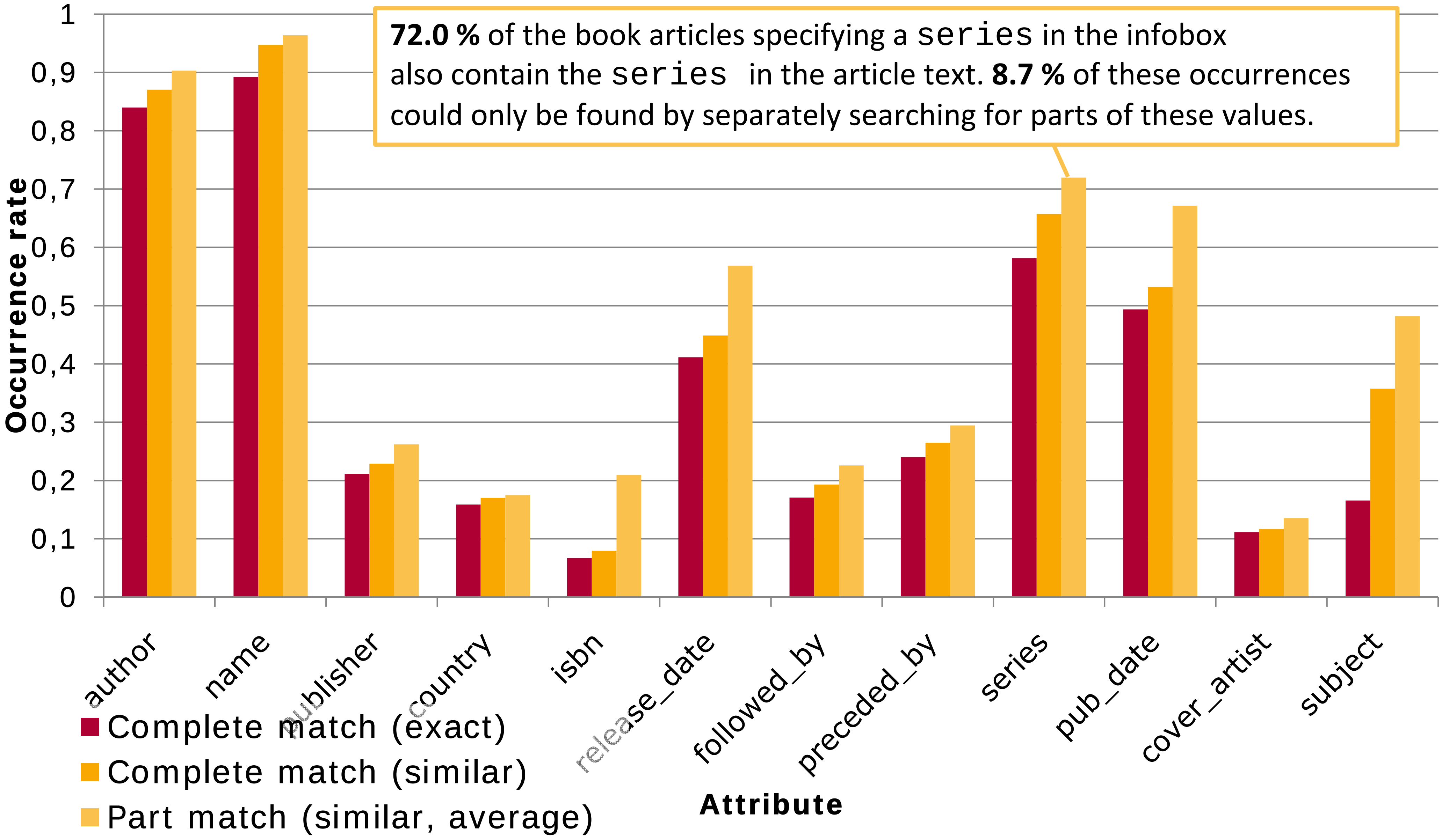
Contents [hide]

***Nineteen Eighty-Four* (1984)**

British first edition cover

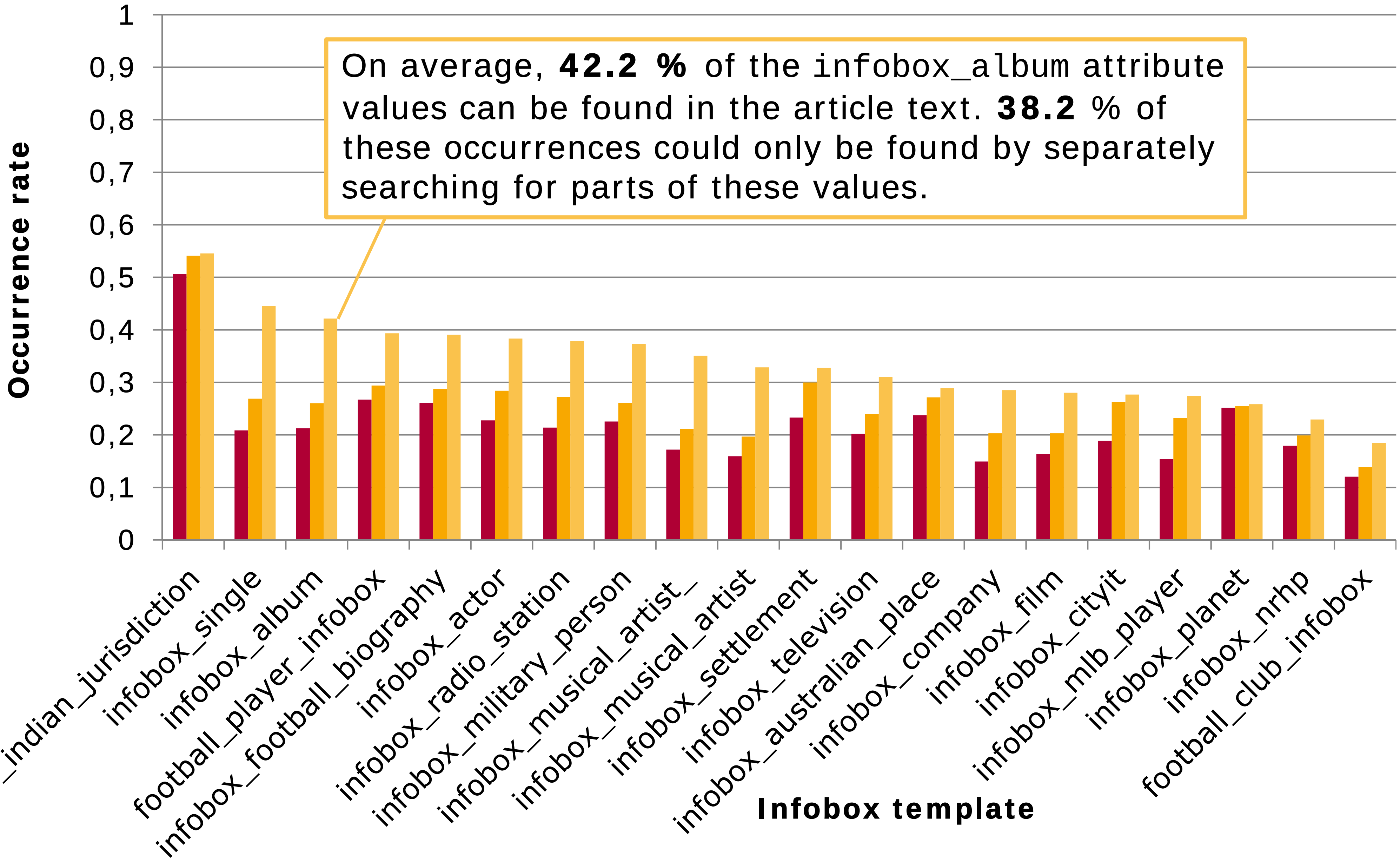
Author	George Orwell
Country	United Kingdom
Language	English
Genre(s)	Dystopian, Political novel, Social science fiction
Publisher	Secker and Warburg (London)
Publication date	8 June 1949
Media type	Print (Hardcover & Paperback) & e-book, audio-CD
Pages	326 pp (Paperback edition)
ISBN	978-0452284234

Occurrence of values in article text: 12 most frequent attributes in infobox_book



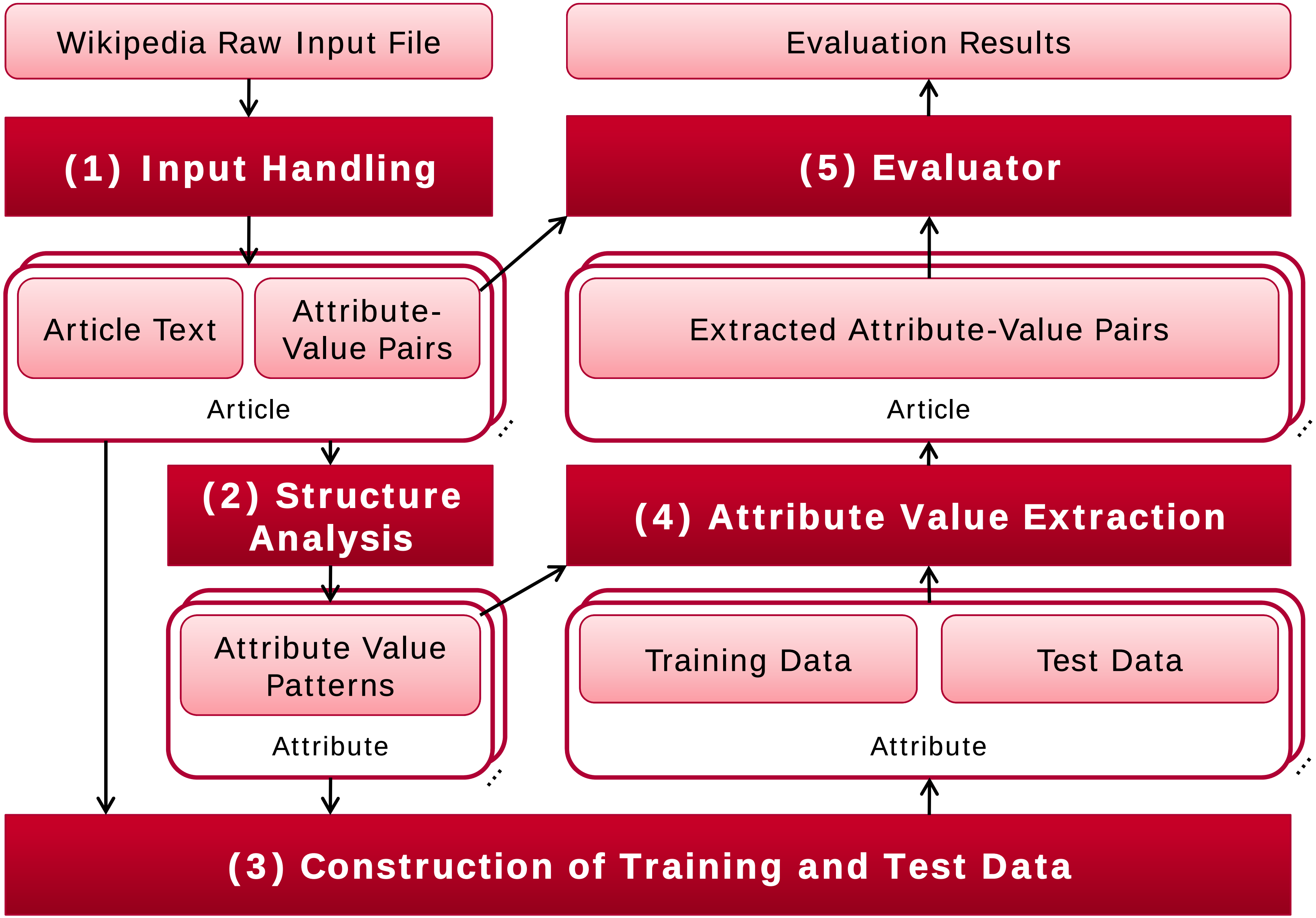
72.0 % of the book articles specifying a `series` in the infobox also contain the `series` in the article text. **8.7 %** of these occurrences could only be found by separately searching for parts of these values.

20 most frequent templates

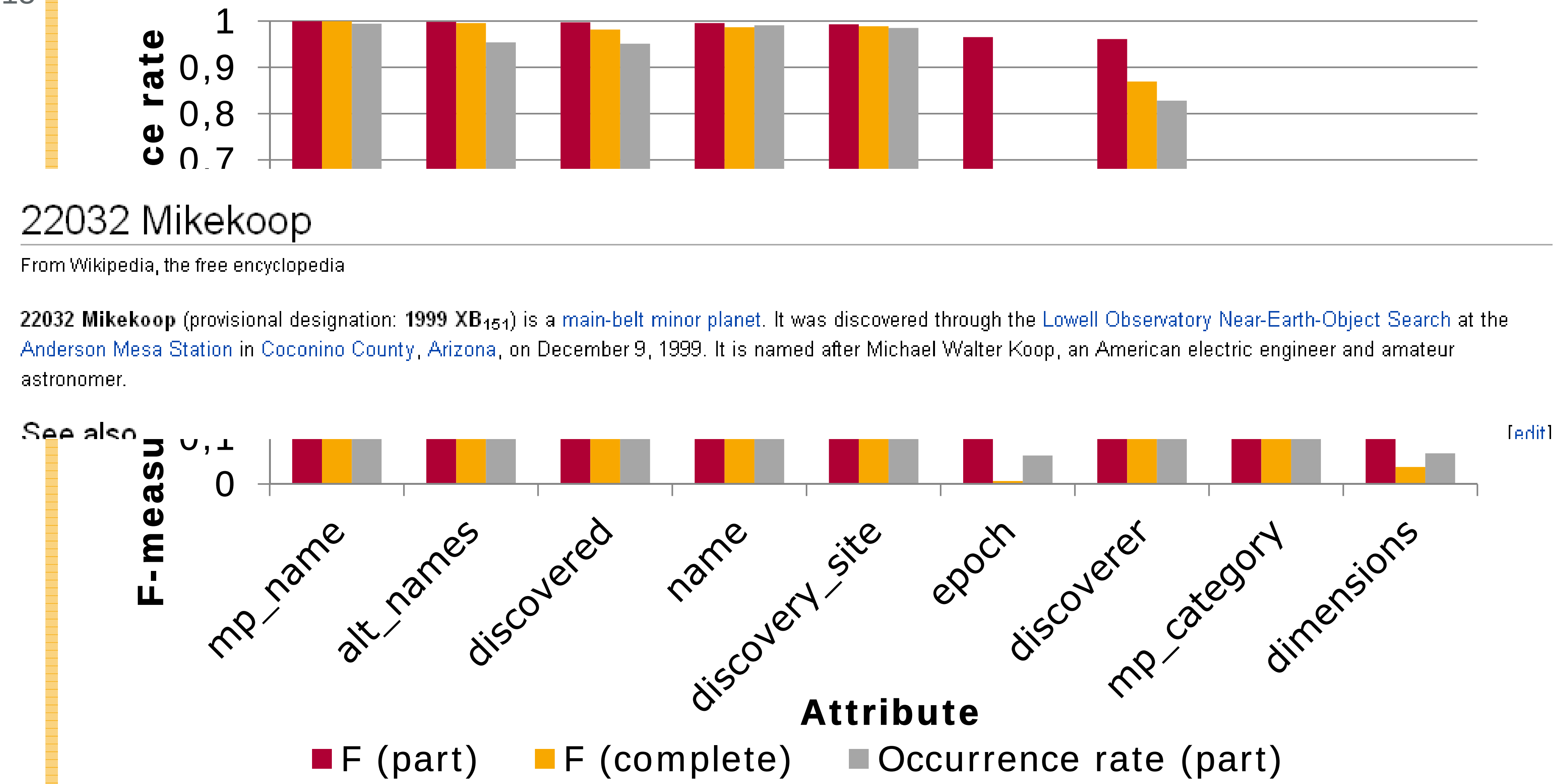


On average, **42.2 %** of the `infobox_album` attribute values can be found in the article text. **38.2** % of these occurrences could only be found by separately searching for parts of these values.

Architecture of iPopulator



Evaluation: infobox_planet



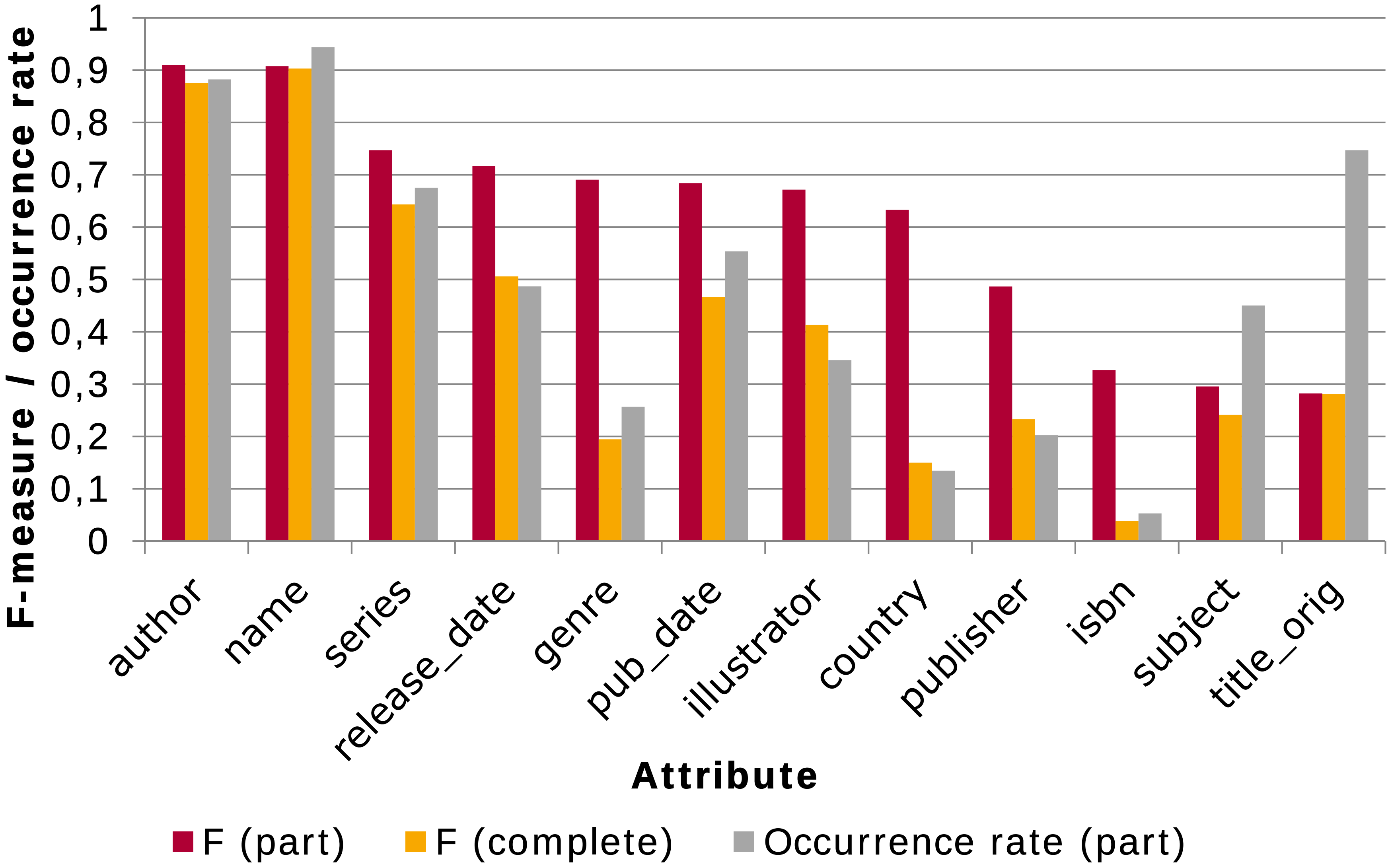
22032 Mikekoop

From Wikipedia, the free encyclopedia

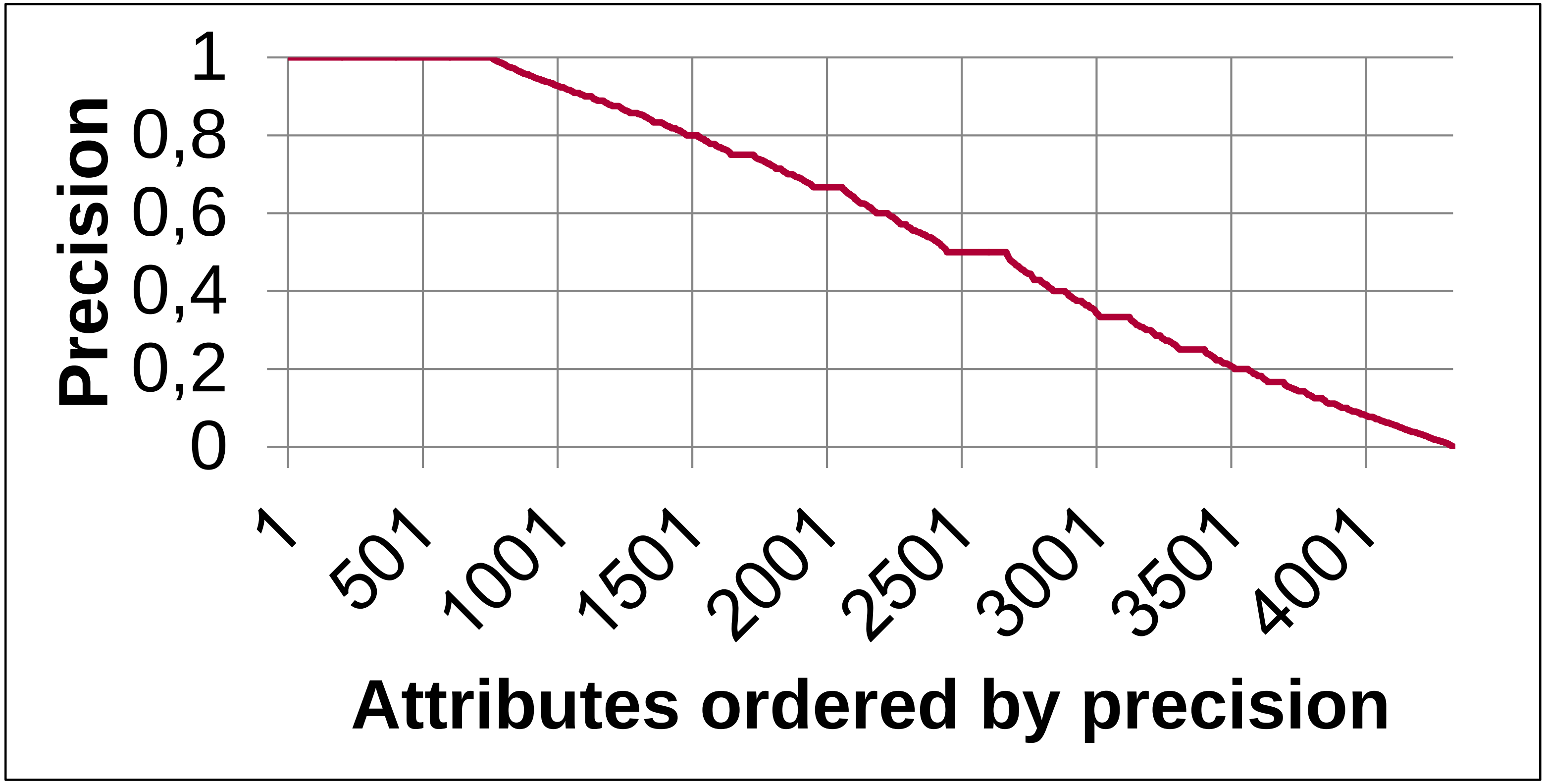
22032 Mikekoop (provisional designation: **1999 XB$_{151}$**) is a main-belt minor planet. It was discovered through the Lowell Observatory Near-Earth-Object Search at the Anderson Mesa Station in Coconino County, Arizona, on December 9, 1999. It is named after Michael Walter Koop, an American electric engineer and amateur astronomer.

See also [edit]

Evaluation: infobox_book



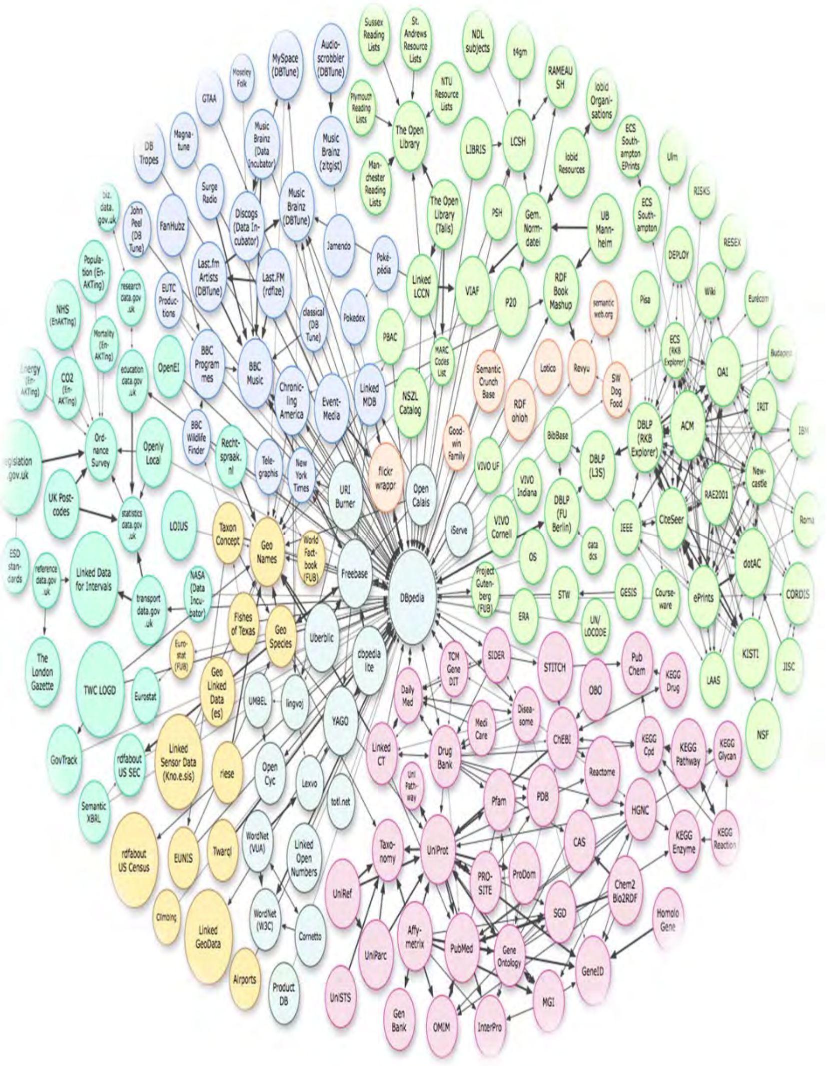
Evaluation on all attributes (>4000) of all infobox templates (>800)



http://www.hpi.uni-potsdam.de/naumann/projekte/completed_projects/ipopulator.html

Overview

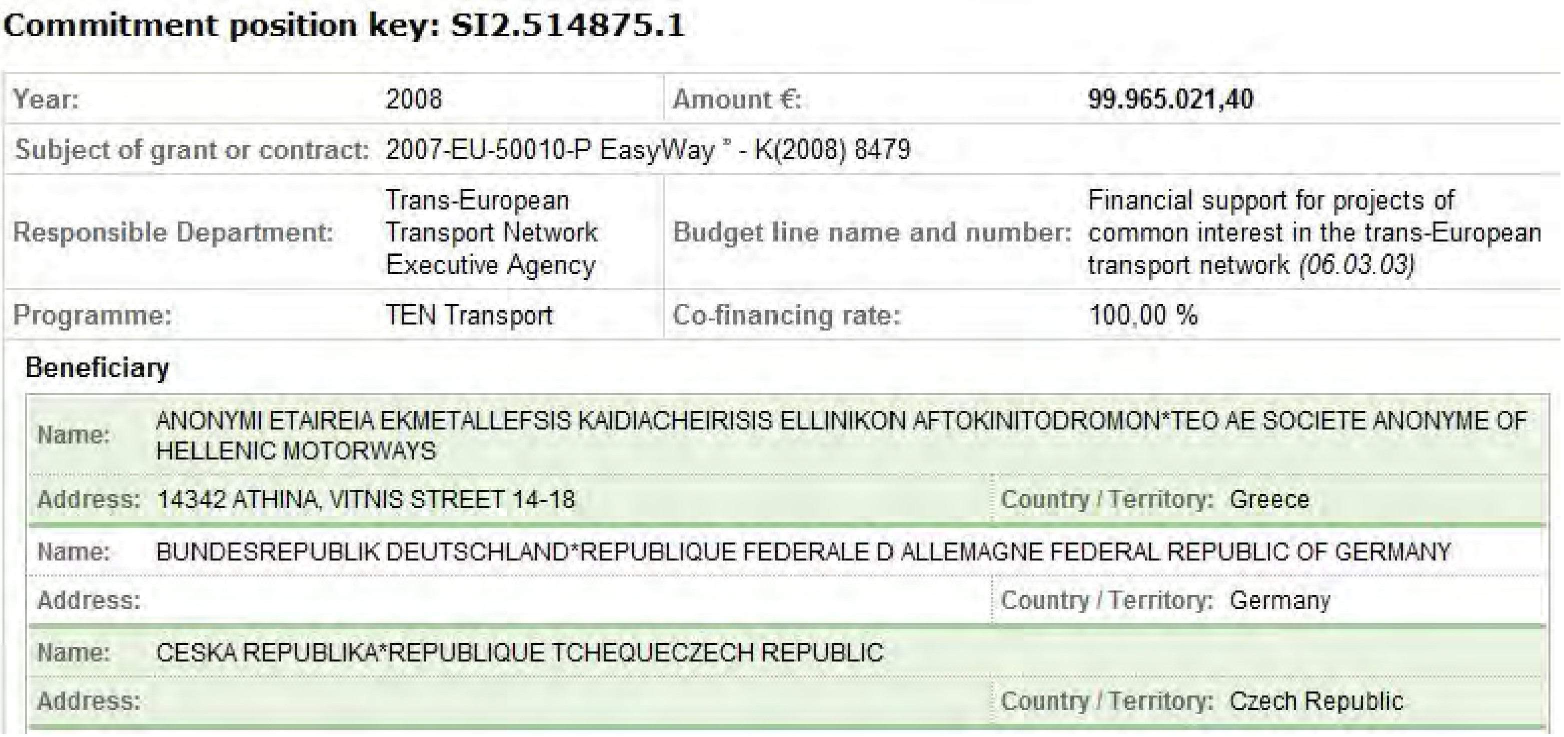
- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



Challenges: Heterogeneity at all levels

- Source
  - □ Formats ↔ □ File converters
  - □ Domain ↔ □ Clustering, topic mining
  - □ Bandwidth ↔ □ Patience
- Schema
  - □ Structure ↔ □ Schema Mapping
  - □ Semantics ↔ □ Domain knowledge
- Data
  - □ Formatting ↔ □ Scrubbing
  - □ Duplicates ↔ □ Entity Matching

The problem – a format mess



Commitment position key: SI2.514875.1

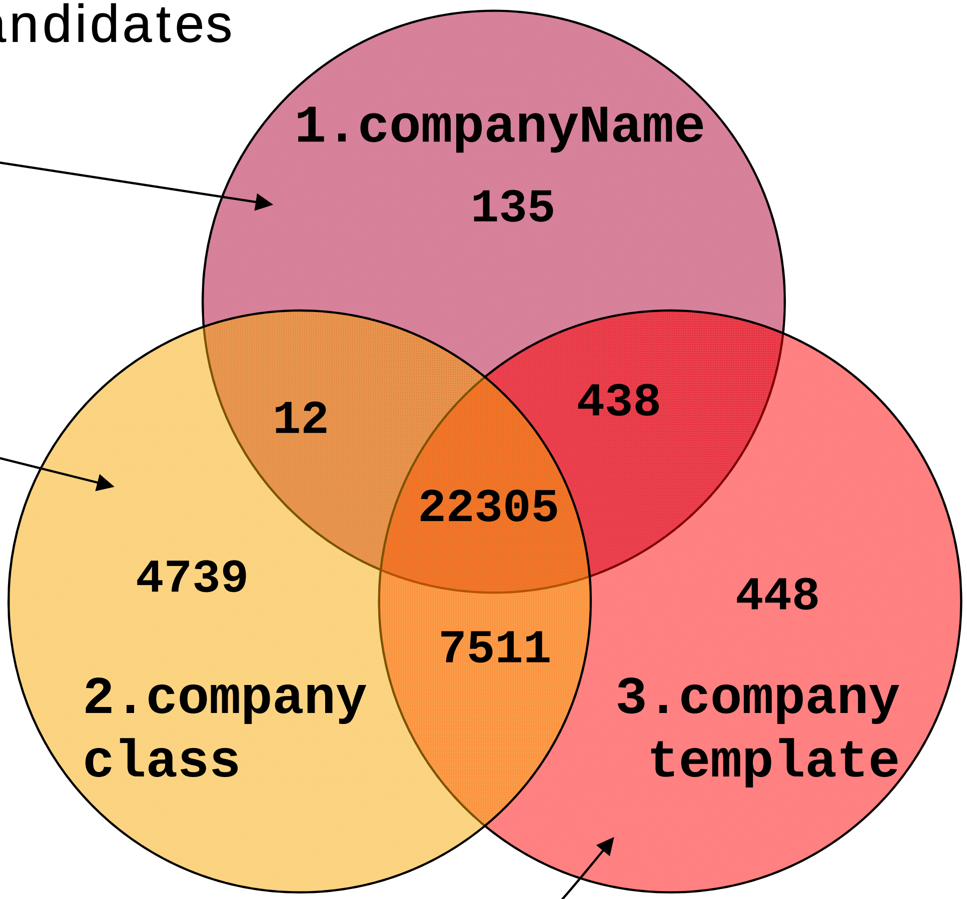
Year:	2008	Amount €:	99.965.021,40
Subject of grant or contract:	2007-EU-50010-P EasyWay " - K(2008) 8479		
Responsible Department:	Trans-European Transport Network Executive Agency	Budget line name and number:	Financial support for projects of common interest in the trans-European transport network *(06.03.03)*
Programme:	TEN Transport	Co-financing rate:	100,00 %

Beneficiary

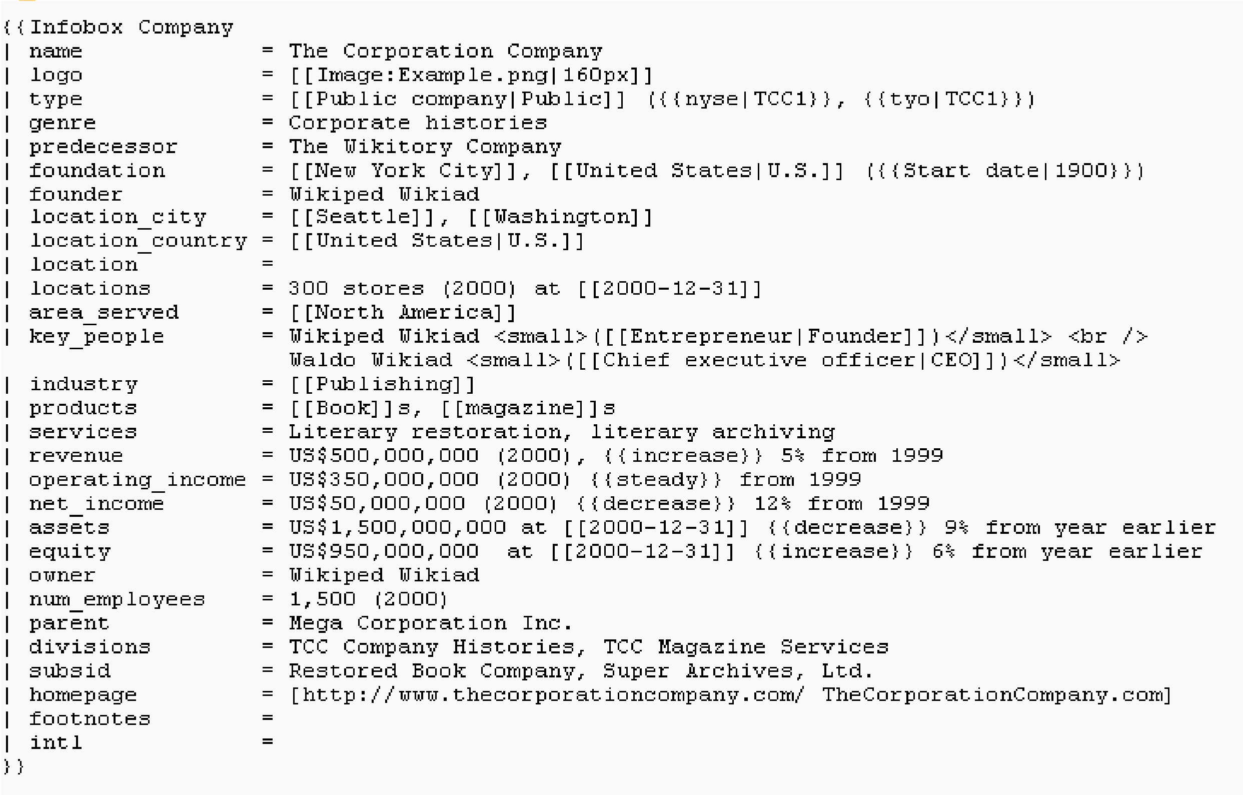
Name:	ANONYMI ETAIREIA EKMETALLEFSIS KAIDIACHEIRISIS ELLINIKON AFTOKINITODROMON*TEO AE SOCIETE ANONYME OF HELLENIC MOTORWAYS		
Address:	14342 ATHINA, VITNIS STREET 14-18	Country / Territory:	Greece
Name:	BUNDESREPUBLIK DEUTSCHLAND*REPUBLIQUE FEDERALE D ALLEMAGNE FEDERAL REPUBLIC OF GERMANY		
Address:		Country / Territory:	Germany
Name:	CESKA REPUBLIKA*REPUBLIQUE TCHEQUECZECH REPUBLIC		
Address:		Country / Territory:	Czech Republic

The problem – a domain mess

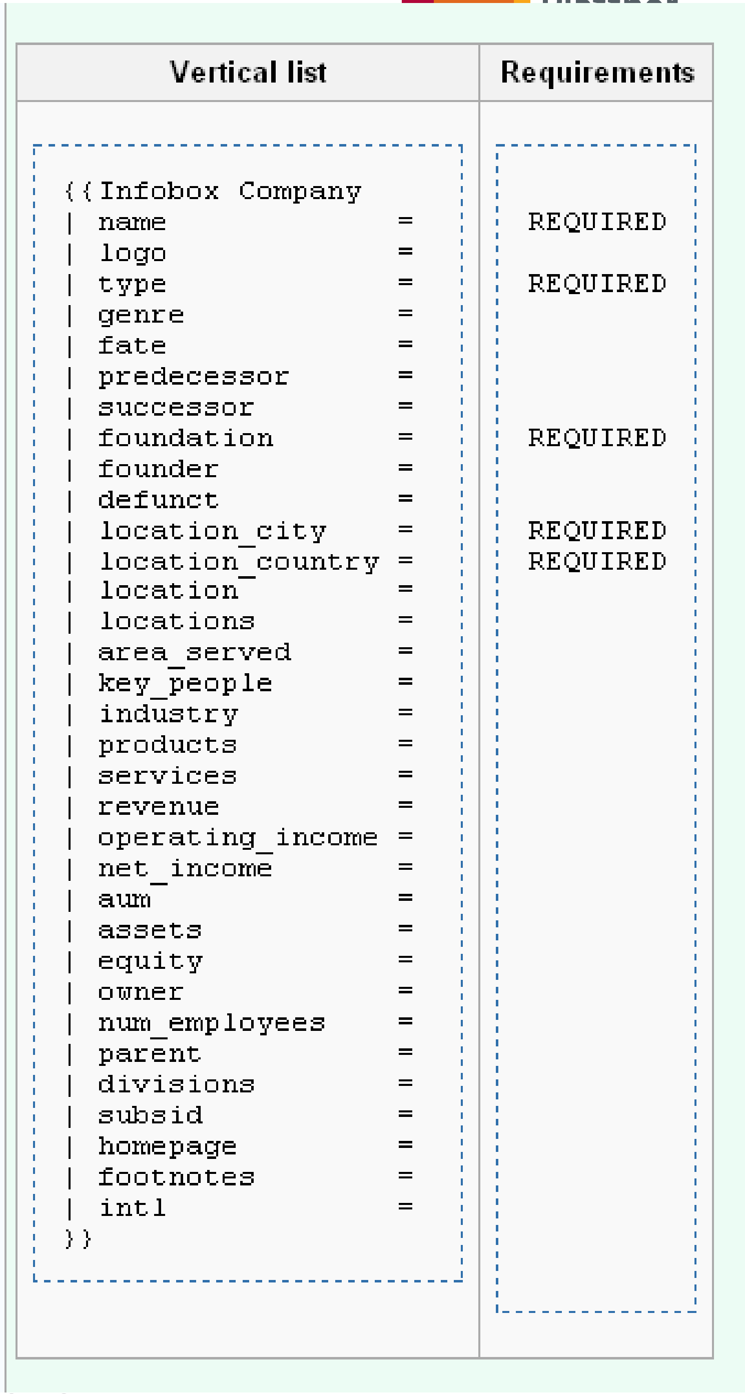
- What is a company? 35,588 candidates
- Def. 1: Entities having a %companyName%
  - □ 22,890
- Def. 2: “Company” according to DBpedia ontology
  - □ 34,567
- Def. 3: Entities having a wikiPageUsesTemplate with value %compan%
  - □ 30,702



Company Template



```
{{Infobox Company
| name             = The Corporation Company
| logo             = [[Image:Example.png|160px]]
| type             = [[Public company|Public]] ({{nyse|TCC1}}, {{tyo|TCC1}})
| genre            = Corporate histories
| predecessor      = The Wikitory Company
| foundation       = [[New York City]], [[United States|U.S.]] ({{Start date|1900}})
| founder          = Wikiped Wikiad
| location_city    = [[Seattle]], [[Washington]]
| location_country = [[United States|U.S.]]
| location         =
| locations        = 300 stores (2000) at [[2000-12-31]]
| area_served      = [[North America]]
| key_people       = Wikiped Wikiad <small>([[Entrepreneur|Founder]])</small> <br />
                     Waldo Wikiad <small>([[Chief executive officer|CEO]])</small>
| industry         = [[Publishing]]
| products         = [[Book]]s, [[magazine]]s
| services         = Literary restoration, literary archiving
| revenue          = US$500,000,000 (2000), {{increase}} 5% from 1999
| operating_income = US$350,000,000 (2000) {{steady}} from 1999
| net_income       = US$50,000,000 (2000) {{decrease}} 12% from 1999
| assets           = US$1,500,000,000 at [[2000-12-31]] {{decrease}} 9% from year earlier
| equity           = US$950,000,000  at [[2000-12-31]] {{increase}} 6% from year earlier
| owner            = Wikiped Wikiad
| num_employees    = 1,500 (2000)
| parent           = Mega Corporation Inc.
| divisions        = TCC Company Histories, TCC Magazine Services
| subsid           = Restored Book Company, Super Archives, Ltd.
| homepage         = [http://www.thecorporationcompany.com/ TheCorporationCompany.com]
| footnotes        =
| intl             =
}}
```



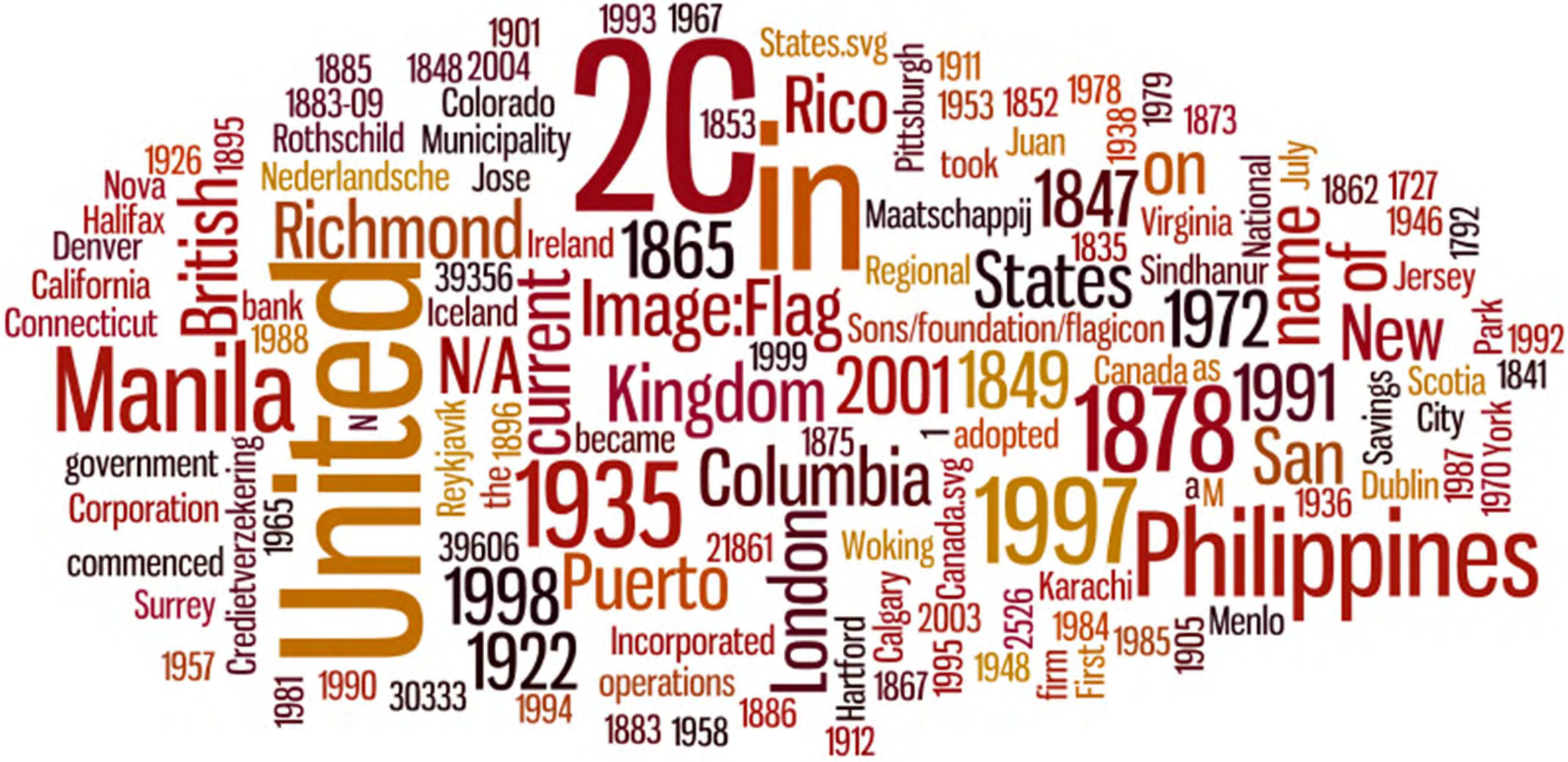
Vertical list	Requirements
`{{Infobox Company`	
`| name             =`	REQUIRED
`| logo             =`	
`| type             =`	REQUIRED
`| genre            =`	
`| fate             =`	
`| predecessor      =`	
`| successor        =`	
`| foundation       =`	REQUIRED
`| founder          =`	
`| defunct          =`	
`| location_city    =`	REQUIRED
`| location_country =`	REQUIRED
`| location         =`	
`| locations        =`	
`| area_served      =`	
`| key_people       =`	
`| industry         =`	
`| products         =`	
`| services         =`	
`| revenue          =`	
`| operating_income =`	
`| net_income       =`	
`| aum              =`	
`| assets           =`	
`| equity           =`	
`| owner            =`	
`| num_employees    =`	
`| parent           =`	
`| divisions        =`	
`| subsid           =`	
`| homepage         =`	
`| footnotes        =`	
`| intl             =`	
`}}`	

The problem – a schema mess

- Triples and ill-defined templates invite disaster.
- Schema chaos: Many attribute synonyms
  - □ Hundreds of different attributes
- Schema misuse: Many attribute homonyms
  - □ **Foundation** attribute in DBpedia may contain
    - ◇ Person who founded the company
    - ◇ Year/Date company was founded
    - ◇ Location where the company was found

- _percent_27_percent_27_percent_27companyName
- _percent_3Cbr/_percent_3ECompanyName
- automatedImagingAssociationCompanyName
- bTcgvuvCompanyName
- bellFoundryCompanyName
- companyNameLocal
- companyNameZh
- companyName_percent_E3_percent_80_percent_80
- companyNames
- dvdEuroCompanyName
- europeanTradeAssociationCompanyName
- iceCreamCompanyName
- itIsExpensiveCompanyName
- publicCompanyName
- companyNameEn
- companyNamesBigBum
- companyName

The **foundation** attribute



Infoboxes in Company class

- 34567 companies with 455821 triples
- 1729 different attributes
  - □ 894 appear only once

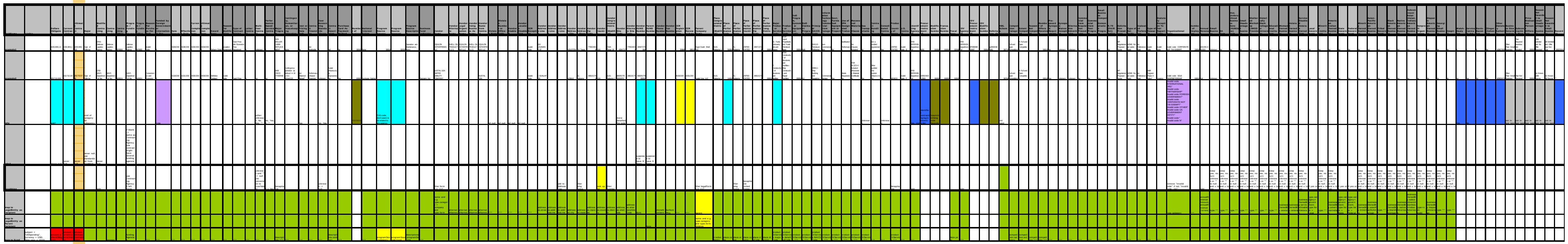
- After cleansing by DBpedia
  - □ 34711 companies with 368185 triples
  - □ Only 50 different attributes

keyPeople	34100	headquarters	3191
industry	28720	airline	2686
foundation	26875	services	2568
products	26486	callsign	2391
homepage	25982	icao	2386
location	24094	iata	2363
companyName	23297	owner	2303
companyType	19591	fleetSize	2246
companyLogo	14644	operatingIncome	2246
numEmployees	11395	hubs	2244
locationCity	9210	website	2104
name	8700	intl	1996
locationCountry	7985	defunct	1987
founder	7867	fate	1944
revenue	7391	slogan	1807
parent	6468	country	1734
type	6358	destinations	1712
areaServed	5842	assets	1591
logo	5434	url	1505
founded	4107	locations	1384
companySlogan	4053	divisions	1227
netIncome	3528	logoSize	1217
genre	3369	successor	1211
subsid	3288	distributor	1125

fieldName	<info>	Dollars Obligated	Current Contract Value	Ultimate Contract Value	Major Agency	Modified Contracting Agency	Contracting Agency	Contracting Office	Program / Funding Agency	Program / Funding Office	Reason For Purchas For DoD
example1		$220,989,132	$220,989,132	$220,989,132	Dept. of Defense	97AS: Defense Logistics Agency	Defense Logistics Agency	SP0600	Defense Logistics Agency	SP0600	Invalid code
example2		$33,710,000	$33,710,000	$33,710,000	Dept. of Defense	1700: NAVY, Department of the	NAVY, Department of the	N00024	NAVY, Department of the	N00024	Convenienc and Econon
info		add?			kind of category for subagency						
info2		never null	never null	never null	never null, use standardized from modified	never null			contracting Agency, one contract might have several funding agencies		
scrubbing						split			use Contracting Agency if left blank		
map to LegalEntity as recipient											
map to LegalEntity as Parent recipient											
	subject = "USSpending",		amount.curr	amount.ulti							

Phew!

Schema mapping and data transformation



1pt font!

The problem – a data mess

- Poor schemata: No types, no constraints
- Sloppy data entry:
  - Data value are neither standardized nor normalized
- **Revenue** attribute may contain different units, different currencies, and different number-formats.
  - 1.64 billion USD vs. $1640 m vs. 1,6 vs. more than one million Euro in 2006
  - And lots of other stuff:

Wal-Mart

Undisclosed

?

Assets exceed £4 billion GBP

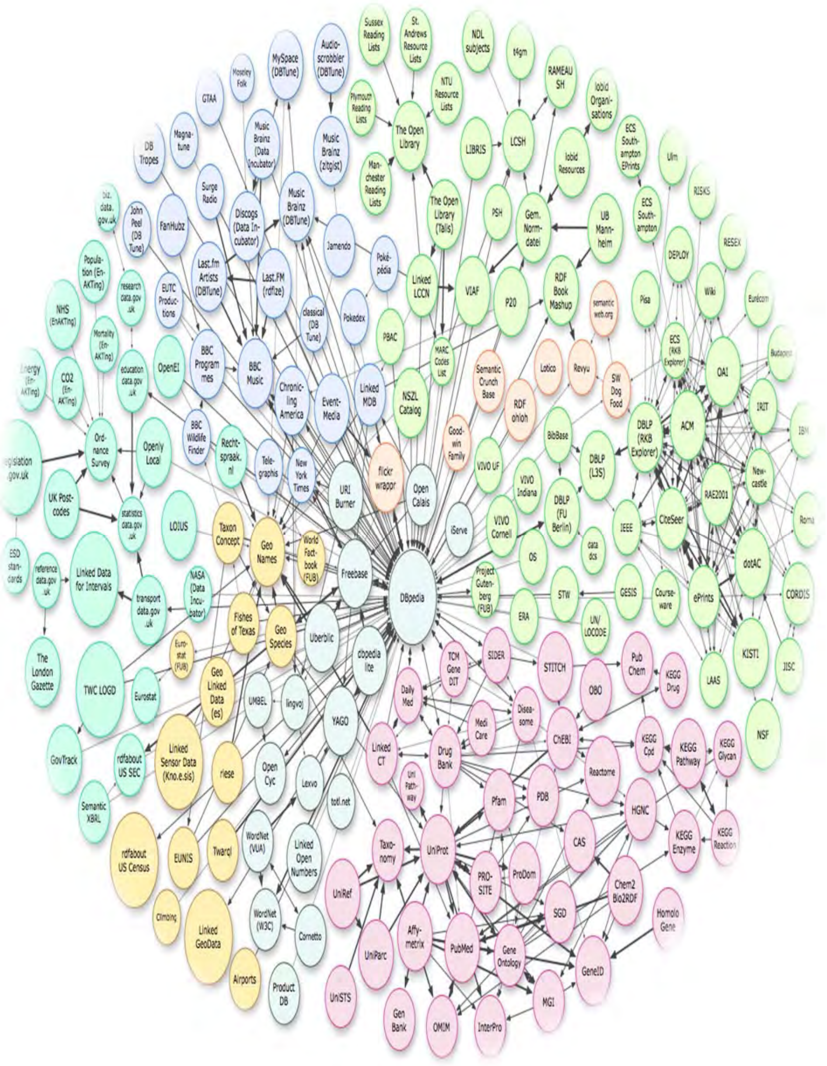
http://www.credit-suisse.com/investors/en/reports/2007_results_q4.jsp

Image:green_up.png → ▲

€ bn (as of 2004)

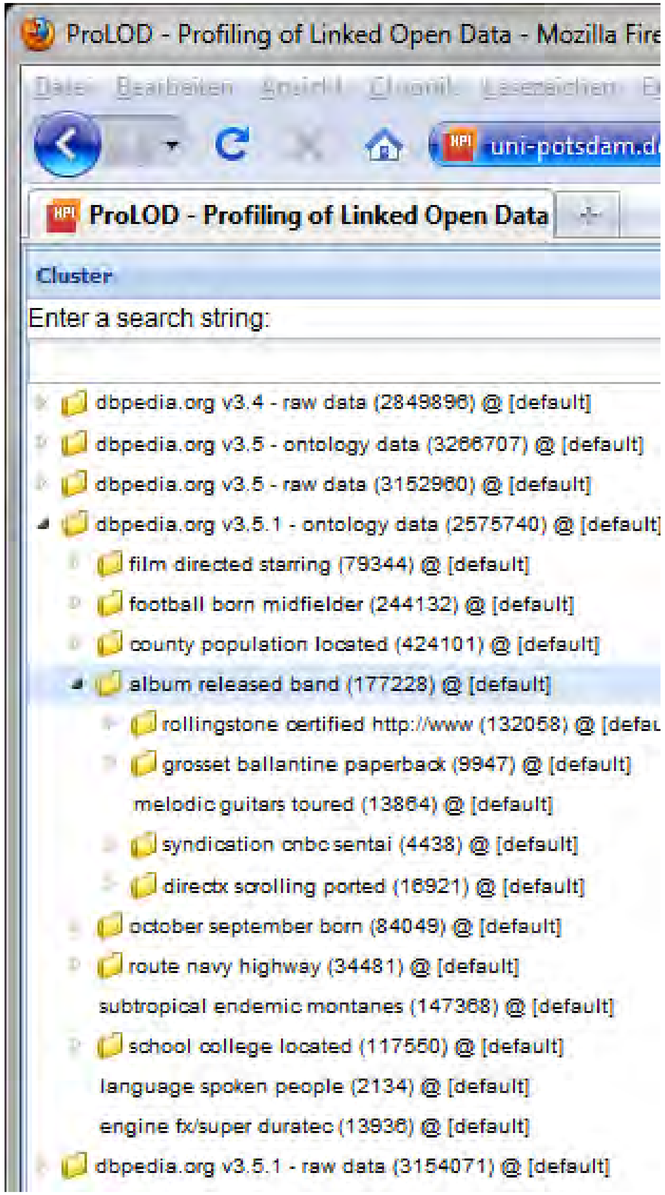
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - …of mops and brooms
  - Cross-language integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience



ProLOD profiling tasks

- Clustering
  - Hierarchical, based on schema
  - Labeling
- Predicate statistics
  - State-of-the-art profiling for attribute values
  - Value types: literals, internal and external links
  - Data types (String, Text, Integer, Decimal, Date)
  - Strings → determine (normalized) patterns
  - Integers, Decimals → display value ranges



ProLOD - Profiling of Linked Open Data - Mozilla Fire

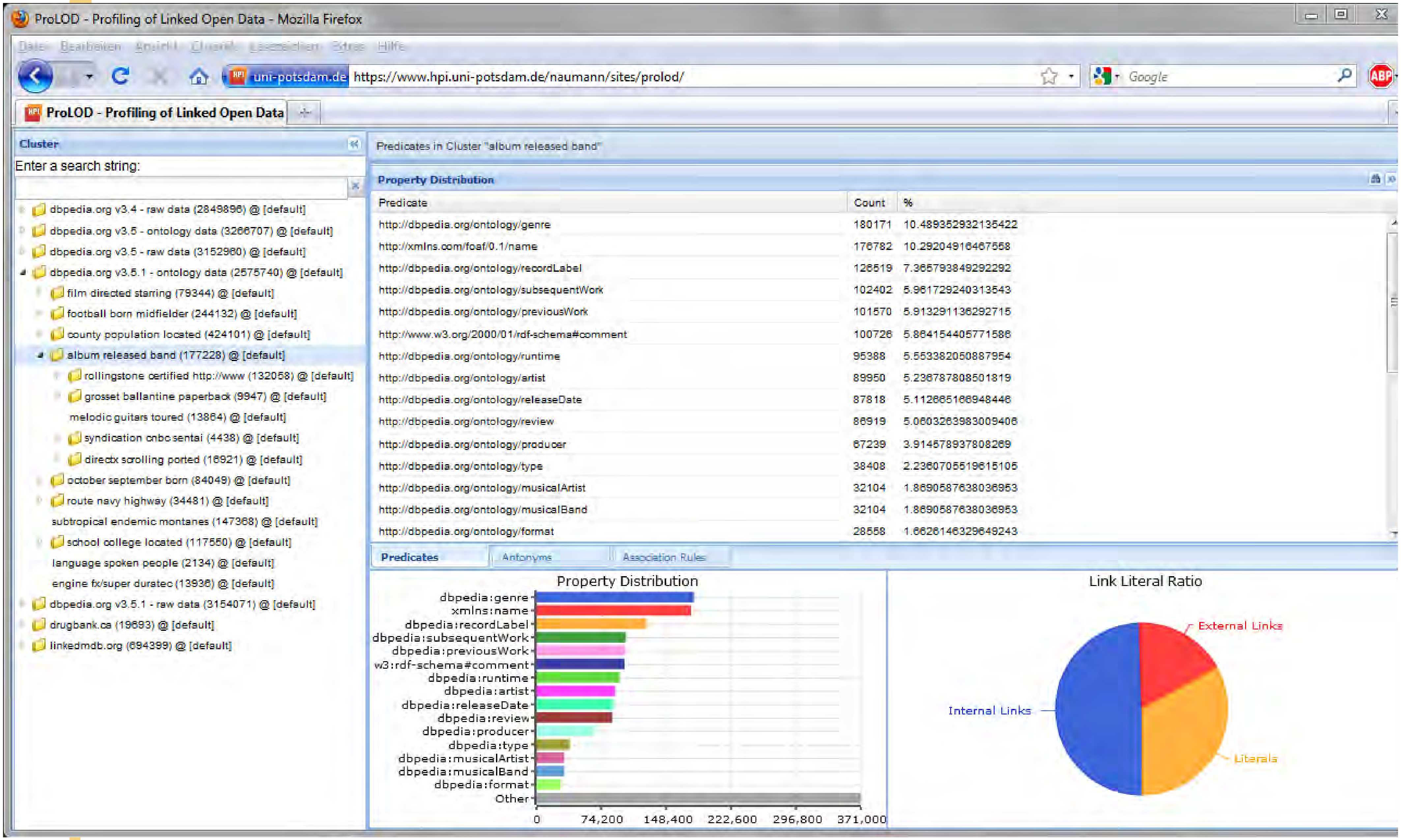
ProLOD - Profiling of Linked Open Data

Cluster

Enter a search string:

- dbpedia.org v3.4 - raw data (2849896) @ [default]
- dbpedia.org v3.5 - ontology data (3266707) @ [default]
- dbpedia.org v3.5 - raw data (3152960) @ [default]
- dbpedia.org v3.5.1 - ontology data (2575740) @ [default]
  - film directed starring (79344) @ [default]
  - football born midfielder (244132) @ [default]
  - county population located (424101) @ [default]
  - album released band (177228) @ [default]
    - rollingstone certified http://www (132058) @ [defau
    - grosset ballantine paperback (9947) @ [default]
    - melodic guitars toured (13864) @ [default]
    - syndication cnbc sentai (4438) @ [default]
    - directx scrolling ported (16921) @ [default]
  - october september born (84049) @ [default]
  - route navy highway (34481) @ [default]
  - subtropical endemic montanes (147368) @ [default]
  - school college located (117550) @ [default]
  - language spoken people (2134) @ [default]
  - engine fx/super duratec (13936) @ [default]
- dbpedia.org v3.5.1 - raw data (3154071) @ [default]

ProLOD – Profiling Linked Open Data


ProLOD - Profiling of Linked Open Data - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
uni-potsdam.de
https://www.hpi.uni-potsdam.de/naumann/sites/prolod/
Google
ProLOD - Profiling of Linked Open Data
Cluster

Enter a search string:

- dbpedia.org v3.4 - raw data (2849896) @ [default]
- dbpedia.org v3.5 - ontology data (3266707) @ [default]
- dbpedia.org v3.5 - raw data (3152980) @ [default]
- dbpedia.org v3.5.1 - ontology data (2575740) @ [default]
  - film directed starring (79344) @ [default]
  - football born midfielder (244132) @ [default]
  - county population located (424101) @ [default]
  - album released band (177228) @ [default]
    - rollingstone certified http://www (132058) @ [default]
    - grosset ballantine paperback (9947) @ [default]
    - melodic guitars toured (13864) @ [default]
    - syndication cnbc sentai (4438) @ [default]
    - directx scrolling ported (16921) @ [default]
  - october september born (84049) @ [default]
  - route navy highway (34481) @ [default]
  - subtropical endemic montanes (147368) @ [default]
  - school college located (117550) @ [default]
  - language spoken people (2134) @ [default]
  - engine fx/super duratec (13936) @ [default]
- dbpedia.org v3.5.1 - raw data (3154071) @ [default]
- drugbank.ca (19693) @ [default]
- linkedmdb.org (694399) @ [default]

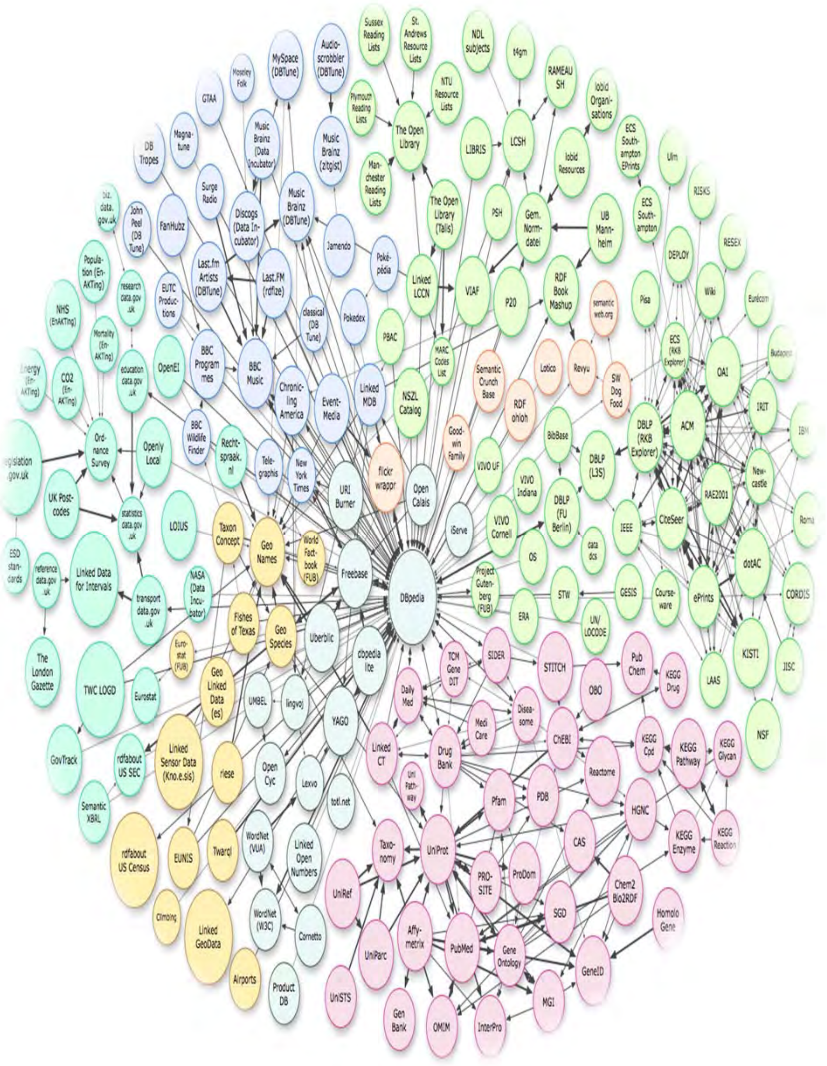
Predicates in Cluster "album released band"

Property Distribution

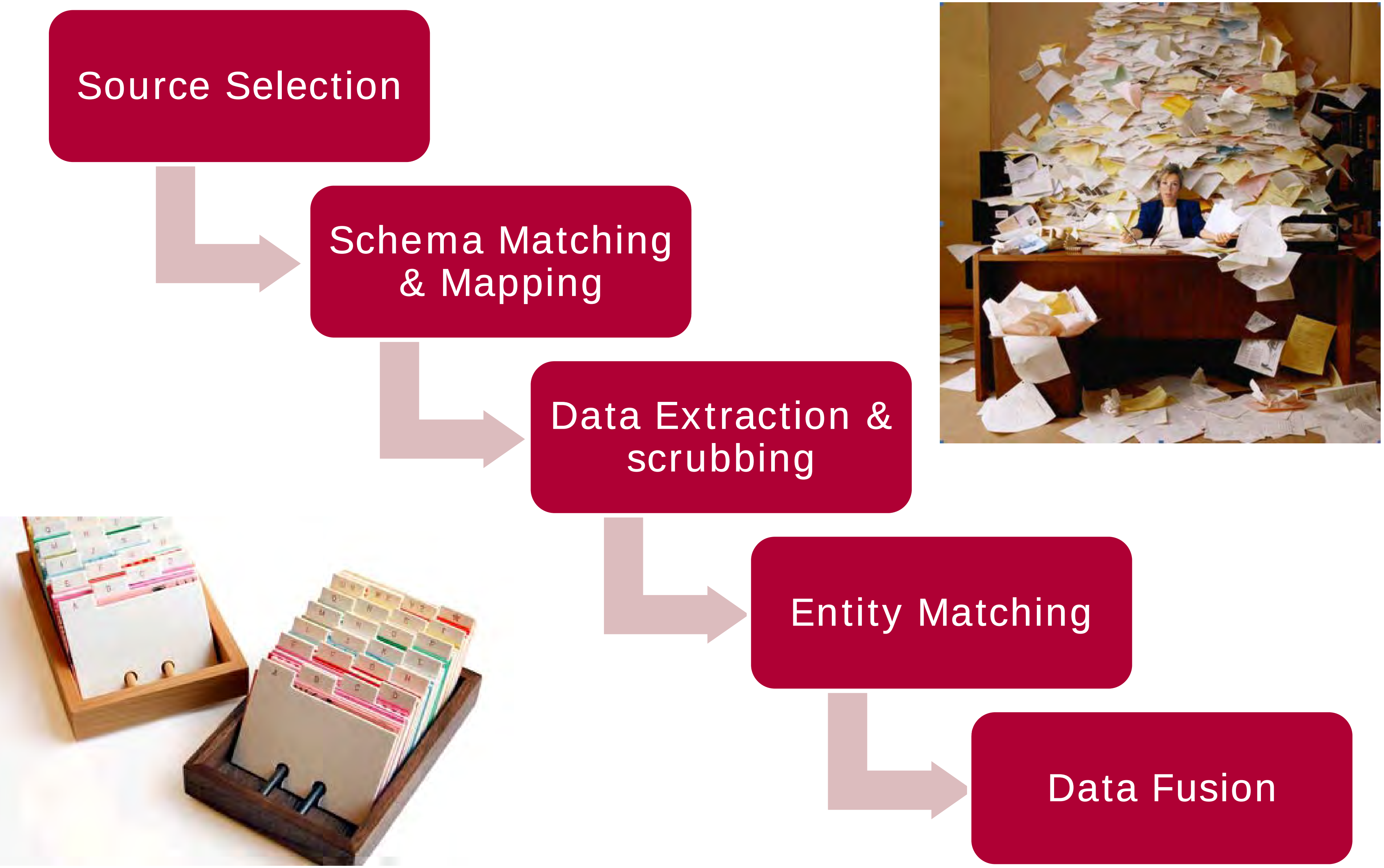
Predicate	Count	%
http://dbpedia.org/ontology/genre	180171	10.489352932135422
http://xmlns.com/foaf/0.1/name	176782	10.29204916467558
http://dbpedia.org/ontology/recordLabel	126519	7.365793849292292
http://dbpedia.org/ontology/subsequentWork	102402	5.961729240313543
http://dbpedia.org/ontology/previousWork	101570	5.913291136292715
http://www.w3.org/2000/01/rdf-schema#comment	100726	5.864154405771586
http://dbpedia.org/ontology/runtime	95388	5.553382050887954
http://dbpedia.org/ontology/artist	89950	5.236787808501819
http://dbpedia.org/ontology/releaseDate	87818	5.112665166948446
http://dbpedia.org/ontology/review	86919	5.0603263983009406
http://dbpedia.org/ontology/producer	67239	3.914578937808269
http://dbpedia.org/ontology/type	38408	2.2360705519615105
http://dbpedia.org/ontology/musicalArtist	32104	1.8690587638036953
http://dbpedia.org/ontology/musicalBand	32104	1.8690587638036953
http://dbpedia.org/ontology/format	28558	1.6626146329649243

Overview

- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



Five steps for integration



Source Selection

Schema Matching & Mapping

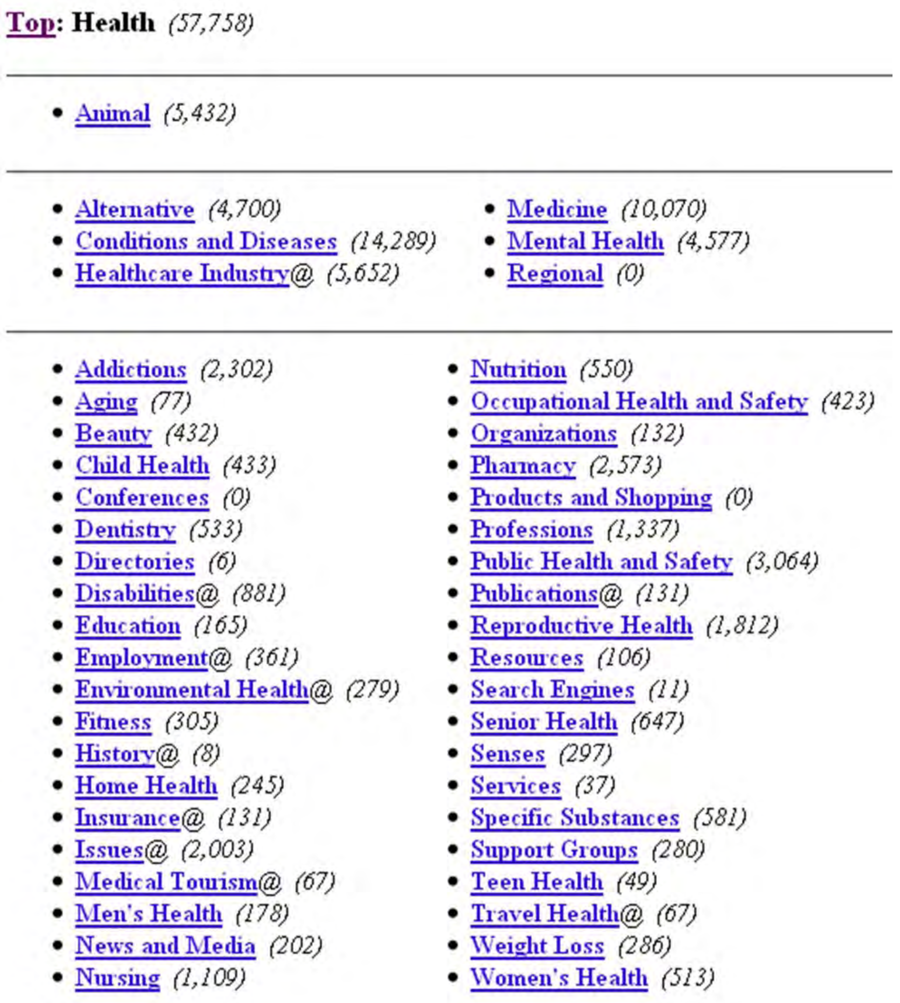
Data Extraction & scrubbing

Entity Matching

Data Fusion

Step 1: Source selection

- Performed by domain experts
- Criteria
  - Availability and downloadability
  - Coverage of domain (completeness)
  - Complementation with other sources
  - Reputation of source
  - Accuracy of data
  - Cost
  - Other data quality criteria…



Top: Health *(57,758)*

- Animal *(5,432)*

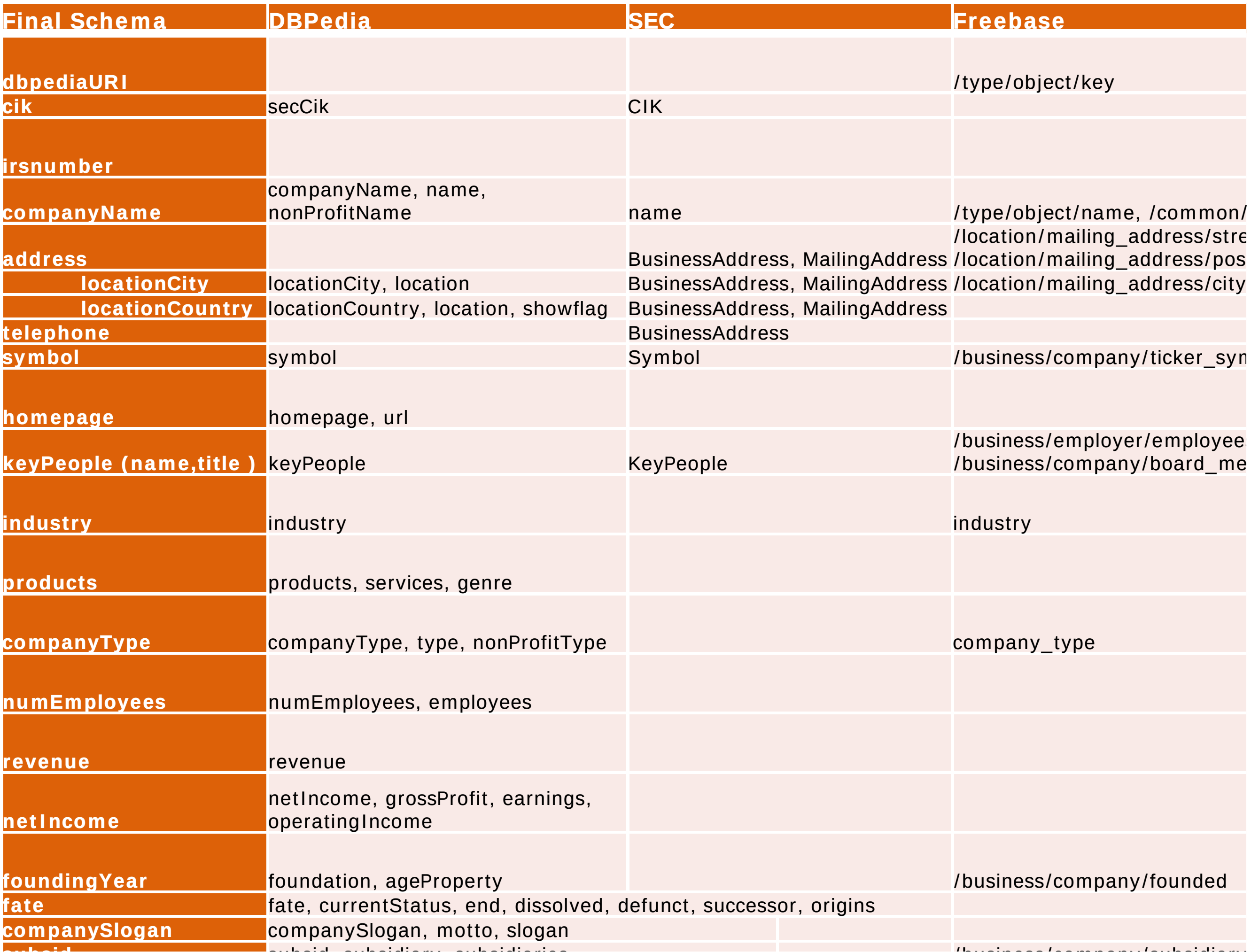
- Alternative *(4,700)*
- Conditions and Diseases *(14,289)*
- Healthcare Industry@ *(5,652)*
- Medicine *(10,070)*
- Mental Health *(4,577)*
- Regional *(0)*

- Addictions *(2,302)*
- Aging *(77)*
- Beauty *(432)*
- Child Health *(433)*
- Conferences *(0)*
- Dentistry *(533)*
- Directories *(6)*
- Disabilities@ *(881)*
- Education *(165)*
- Employment@ *(361)*
- Environmental Health@ *(279)*
- Fitness *(305)*
- History@ *(8)*
- Home Health *(245)*
- Insurance@ *(131)*
- Issues@ *(2,003)*
- Medical Tourism@ *(67)*
- Men's Health *(178)*
- News and Media *(202)*
- Nursing *(1,109)*
- Nutrition *(550)*
- Occupational Health and Safety *(423)*
- Organizations *(132)*
- Pharmacy *(2,573)*
- Products and Shopping *(0)*
- Professions *(1,337)*
- Public Health and Safety *(3,064)*
- Publications@ *(131)*
- Reproductive Health *(1,812)*
- Resources *(106)*
- Search Engines *(11)*
- Senior Health *(647)*
- Senses *(297)*
- Services *(37)*
- Specific Substances *(581)*
- Support Groups *(280)*
- Teen Health *(49)*
- Travel Health@ *(67)*
- Weight Loss *(286)*
- Women's Health *(513)*

dmoz.org

Step 2: Schema matching and mapping

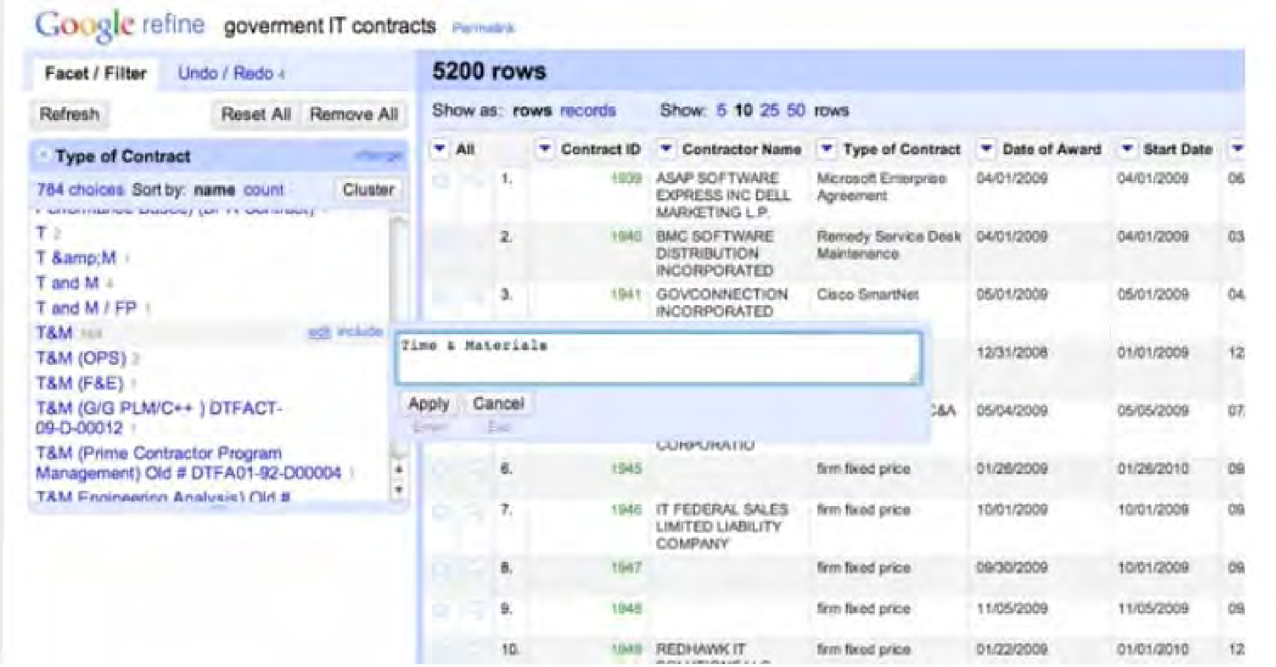
- Semi-automated matching
  - □ Label-based and instance-based
- Challenges:
  - □ Multi-lingual
  - □ Homonyms and Synonyms
  - □ 1:1, 1:n, n:m
- Complex data transformation



Final Schema	DBPedia	SEC	Freebase
dbpediaURI			/type/object/key
cik	secCik	CIK	
irsnumber			
companyName	companyName, name, nonProfitName	name	/type/object/name, /common/
address		BusinessAddress, MailingAddress	/location/mailing_address/stre /location/mailing_address/pos
locationCity	locationCity, location	BusinessAddress, MailingAddress	/location/mailing_address/city
locationCountry	locationCountry, location, showflag	BusinessAddress, MailingAddress	
telephone		BusinessAddress	
symbol	symbol	Symbol	/business/company/ticker_syn
homepage	homepage, url		
keyPeople (name,title)	keyPeople	KeyPeople	/business/employer/employees /business/company/board_me
industry	industry		industry
products	products, services, genre		
companyType	companyType, type, nonProfitType		company_type
numEmployees	numEmployees, employees		
revenue	revenue		
netIncome	netIncome, grossProfit, earnings, operatingIncome		
foundingYear	foundation, ageProperty		/business/company/founded
fate	fate, currentStatus, end, dissolved, defunct, successor, origins		
companySlogan	companySlogan, motto, slogan		
subsid	subsid, subsidiary, subsidiaries		/business/company/subsidiary

Step 3: Data extraction & scrubbing

- Recognize data types
- Regular expressions for multi-valued strings
- Remove spurious values (layout, formatting, …)
- Standardize formats
- Translate from foreign languages
- Many tools



T 2
T &amp;M 1
T and M 4
T and M / FP 1
T&M 163
T&M (OPS) 2
T&M (F&E) 1
T&M (G/G PLM/C++) DTFACT-09-D-00012 1
T&M (Prime Contractor Program Management) Old # DTFA01-92-D00004 1
T&M Engineering Analysis) Old #

Time & Materials

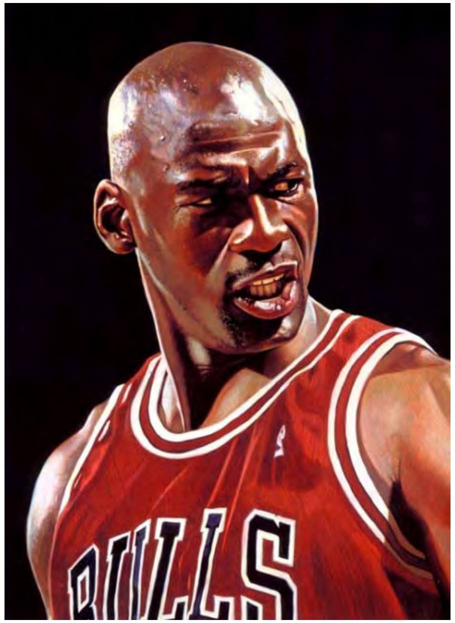
5200 rows

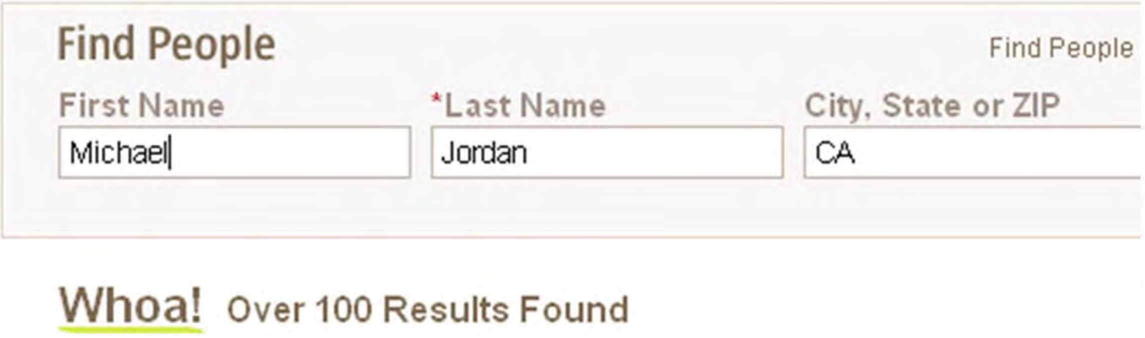
Show as: rows records   Show: 5 10 25 50 rows

All	Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	
1.	1939	ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/2009	04/01/2009	06
2.	1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Remedy Service Desk Maintenance	04/01/2009	04/01/2009	03
3.	1941	GOVCONNECTION INCORPORATED	Cisco SmartNet	05/01/2009	05/01/2009	04
				12/31/2008	01/01/2009	12
		CORPORATIO	C&A	05/04/2009	05/05/2009	07
6.	1945		firm fixed price	01/28/2009	01/28/2010	09
7.	1946	IT FEDERAL SALES LIMITED LIABILITY COMPANY	firm fixed price	10/01/2009	10/01/2009	09
8.	1947		firm fixed price	09/30/2009	10/01/2009	09
9.	1948		firm fixed price	11/05/2009	11/05/2009	09
10.	1949	REDHAWK IT	firm fixed price	01/22/2009	01/01/2010	12

Step 4: Entity matching

- Duplicate entities
- Linking between entities
- Challenges
  - Fuzzy matching: Similarity measures
  - Data volume: Partitioning algorithms
  - Sparse data
    - **Michael Jordan born_in Miami**





Find People

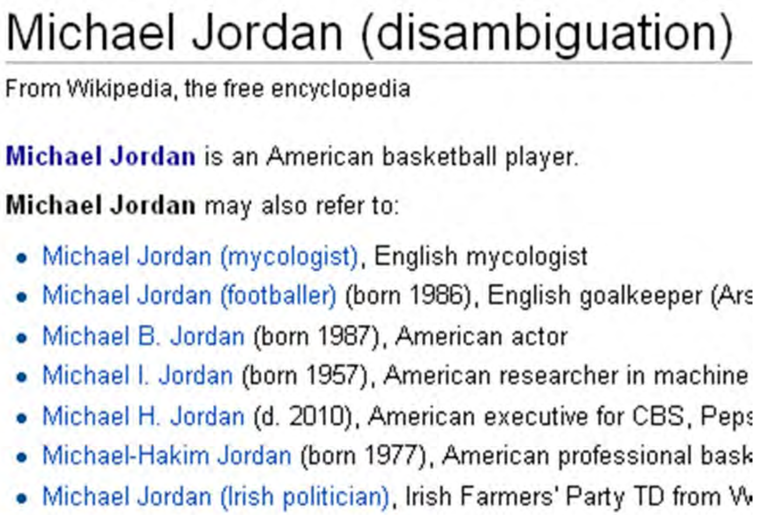
Find People

First Name: Michael

*Last Name: Jordan

City, State or ZIP: CA

Whoa! Over 100 Results Found



Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

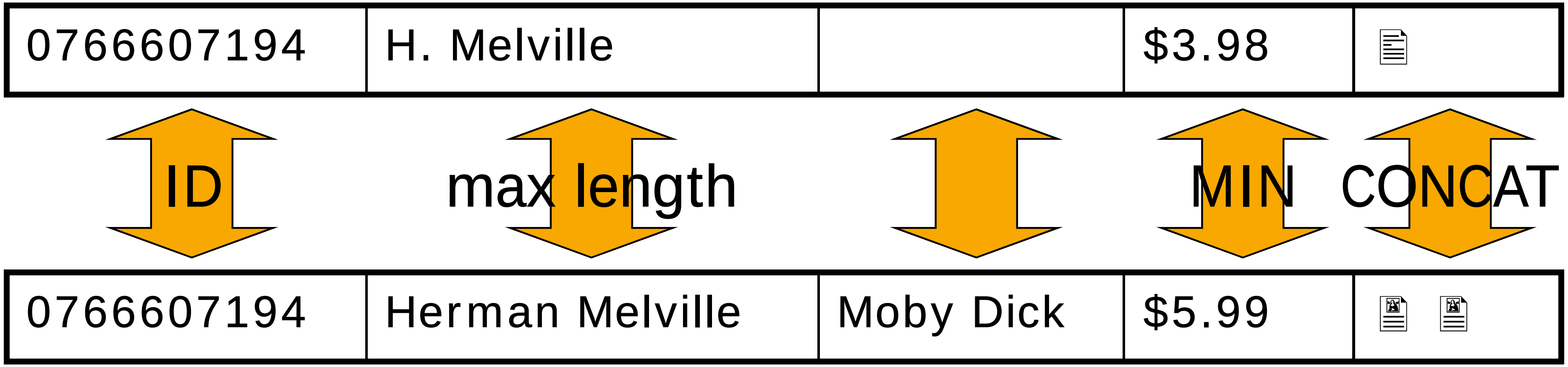
Michael Jordan is an American basketball player.

Michael Jordan may also refer to:

- Michael Jordan (mycologist), English mycologist
- Michael Jordan (footballer) (born 1986), English goalkeeper (Ars
- Michael B. Jordan (born 1987), American actor
- Michael I. Jordan (born 1957), American researcher in machine
- Michael H. Jordan (d. 2010), American executive for CBS, Peps
- Michael-Hakim Jordan (born 1977), American professional bask
- Michael Jordan (Irish politician), Irish Farmers' Party TD from W

Step 5: Data fusion

- Combine multiple representations of real-world entities
  - Survivorship, consolidation, etc.
- Resolve data conflicts
  - Conflict resolution functions
  - Reputation / accuracy / freshness -> “truth discovery”

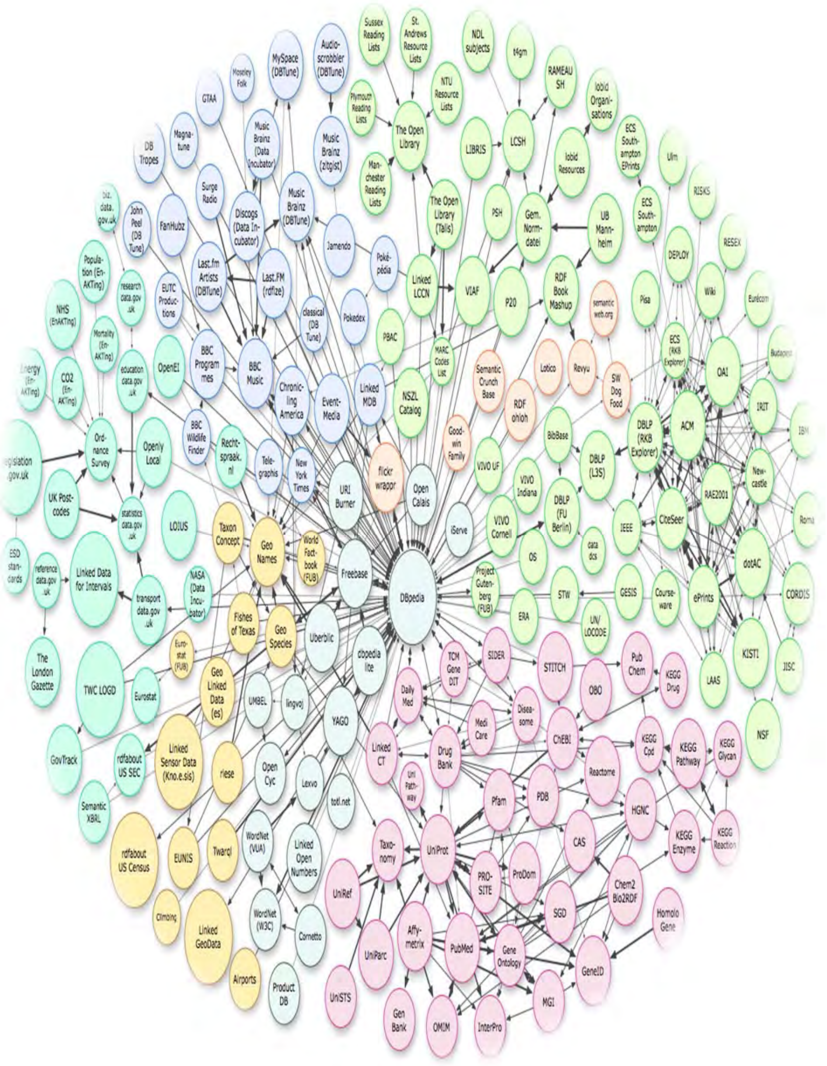


0766607194	H. Melville		$3.98	📄
ID	max length		MIN	CONCAT
0766607194	Herman Melville	Moby Dick	$5.99	📄 📄

- Retain data lineage

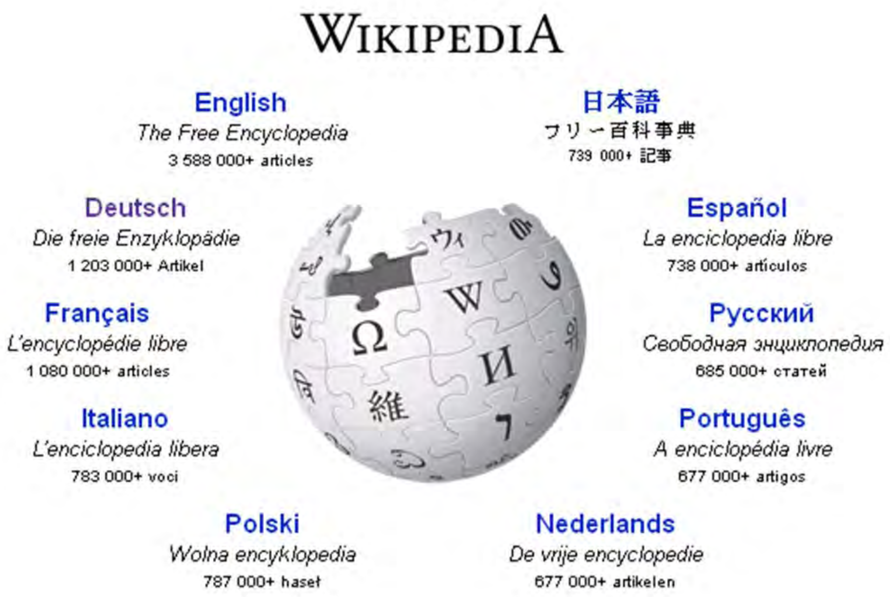
Overview

- ■ Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- ■ Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- ■ Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- ■ Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



Multi-Lingual Wikipedia

- Goal: Schema matching across languages
  - Complement infobox data
  - Autocomplete for authors
  - Detect errors or inconsistencies
  - Keep values up to date
- Idea: Use cross-language links across all 285 languages



WIKIPEDIA

English
The Free Encyclopedia
3 588 000+ articles

日本語
フリー百科事典
739 000+ 記事

Deutsch
Die freie Enzyklopädie
1 203 000+ Artikel

Español
La enciclopedia libre
738 000+ artículos

Français
L'encyclopédie libre
1 080 000+ articles

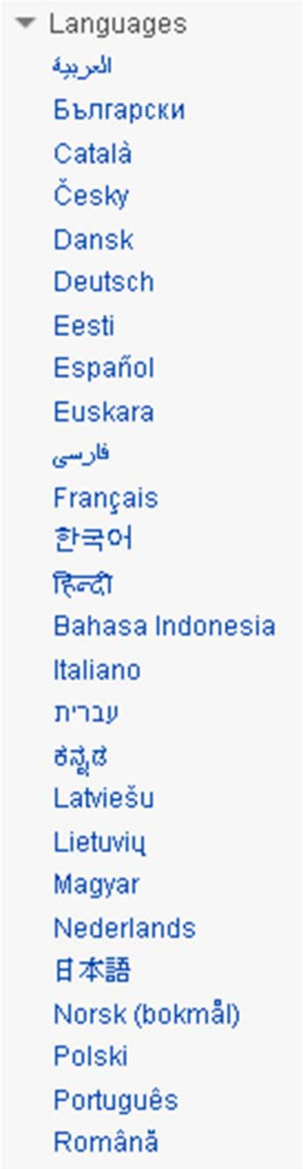
Русский
Свободная энциклопедия
685 000+ статей

Italiano
L'enciclopedia libera
783 000+ voci

Português
A enciclopédia livre
677 000+ artigos

Polski
Wolna encyklopedia
787 000+ haseł

Nederlands
De vrije encyclopedie
677 000+ artikelen

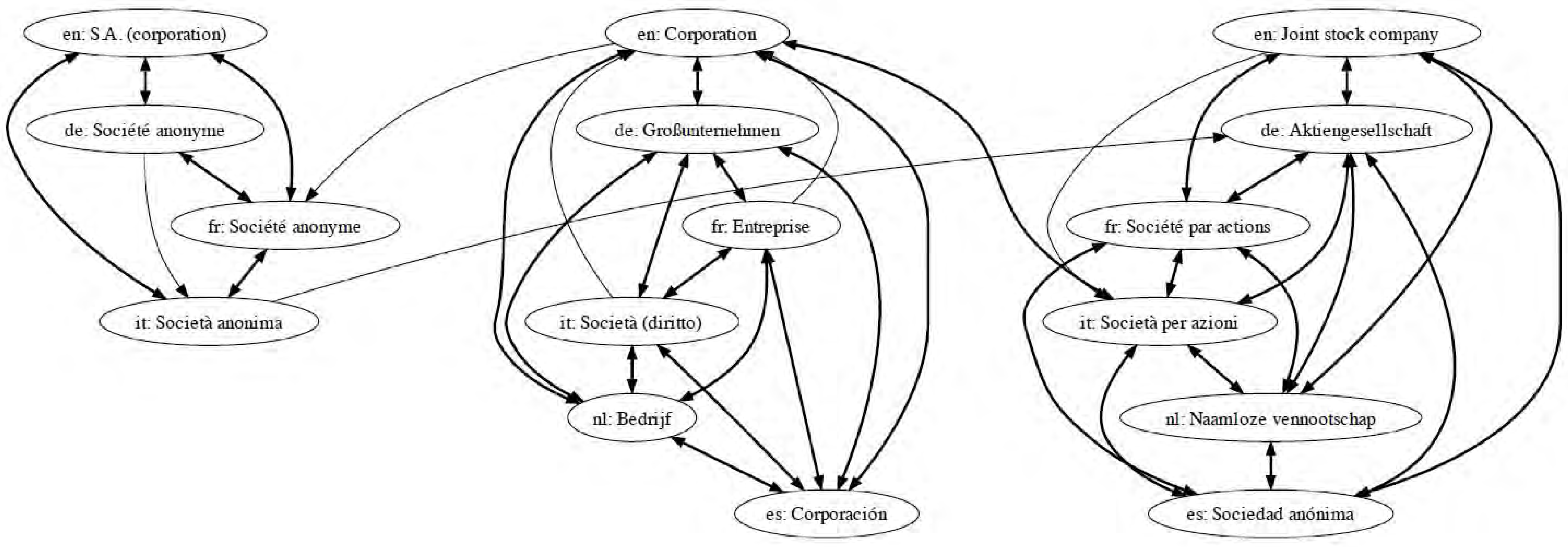


Languages

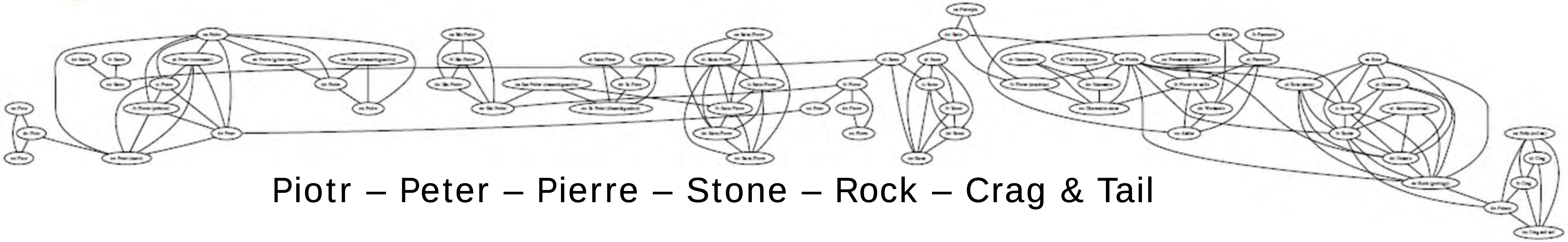
العربية
Български
Català
Česky
Dansk
Deutsch
Eesti
Español
Euskara
فارسی
Français
한국어
हिन्दी
Bahasa Indonesia
Italiano
עברית
ಕನ್ನಡ
Latviešu
Lietuvių
Magyar
Nederlands
日本語
Norsk (bokmål)
Polski
Português
Română

Interlanguage links (ILLs)

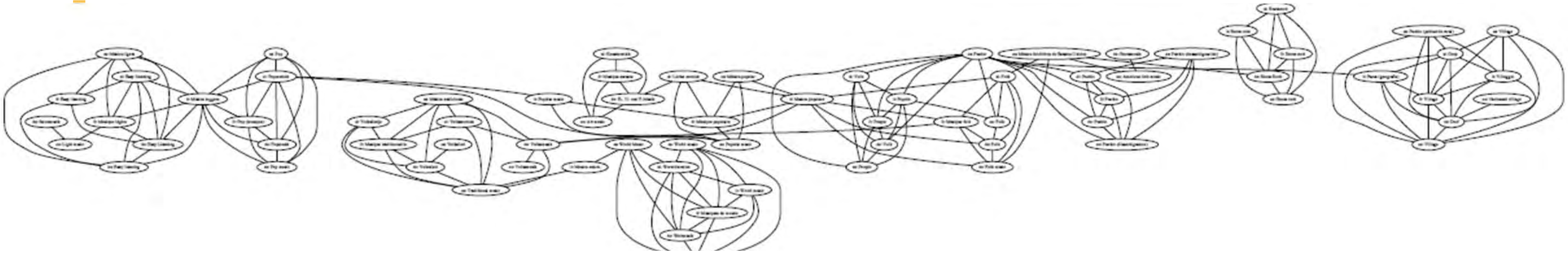
- First, evaluate quality of ILLs and build duplicate clusters
  - Build connected components using cross-language links (on the six largest languages)
- But, largest weakly connected component has 108 articles
  - 26 English, 26 German, 21 French, 13 Italian, 13 Dutch, and 9 Spanish articles



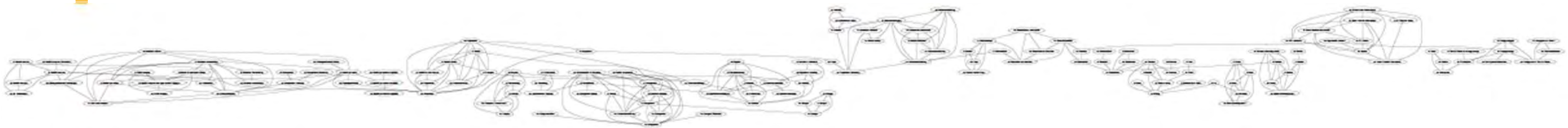
Other large components



Piotr – Peter – Pierre – Stone – Rock – Crag & Tail



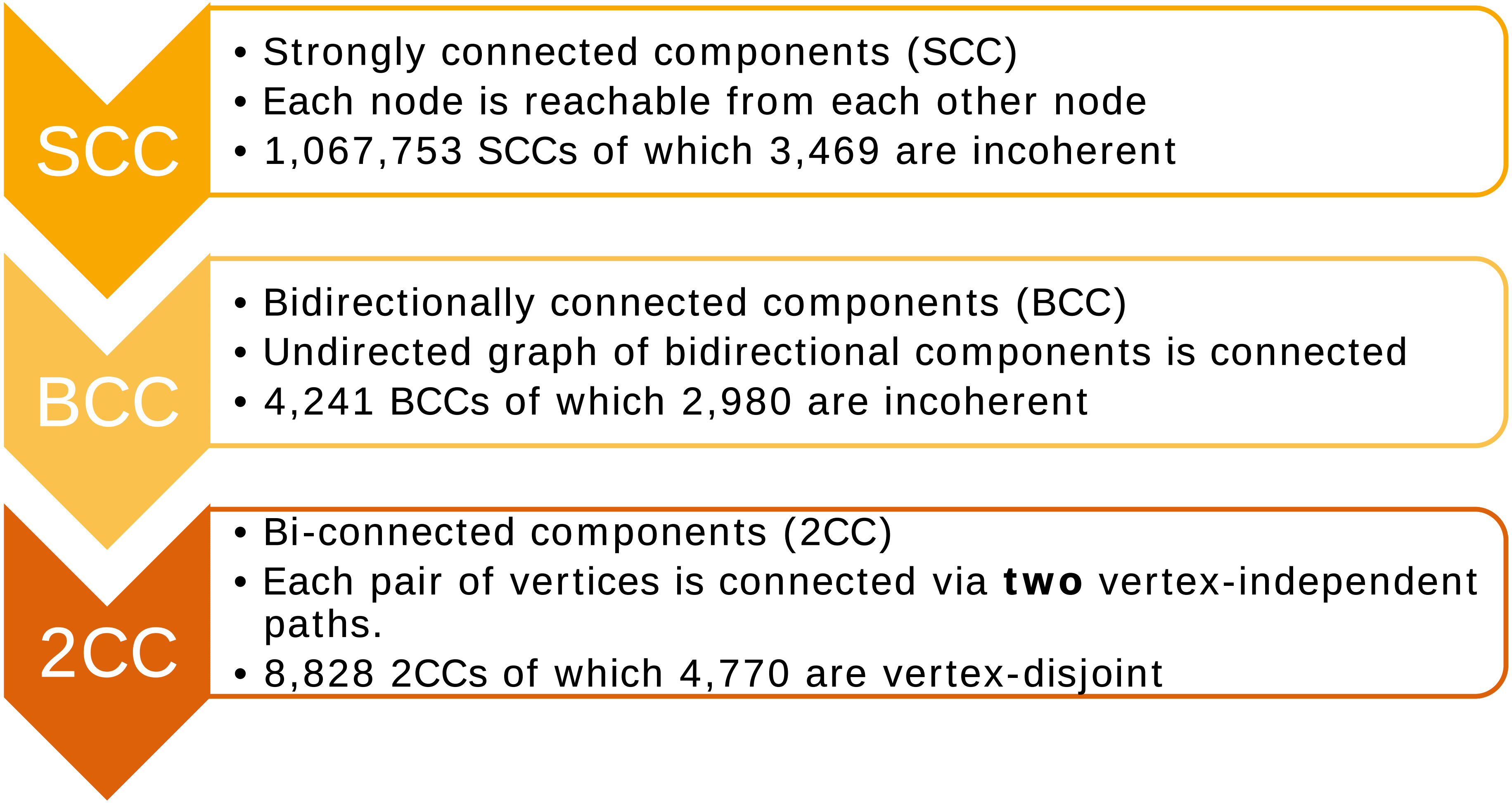
Easy Listening – Pop music – World music – Musique folk – Folk – Pueblo - Village



Joint Stock Company – ... – Brother

Whittling down the ILL set

- A connected component is **incoherent** if it contains more than one node for any language.



SCC

- Strongly connected components (SCC)
- Each node is reachable from each other node
- 1,067,753 SCCs of which 3,469 are incoherent

BCC

- Bidirectionally connected components (BCC)
- Undirected graph of bidirectional components is connected
- 4,241 BCCs of which 2,980 are incoherent

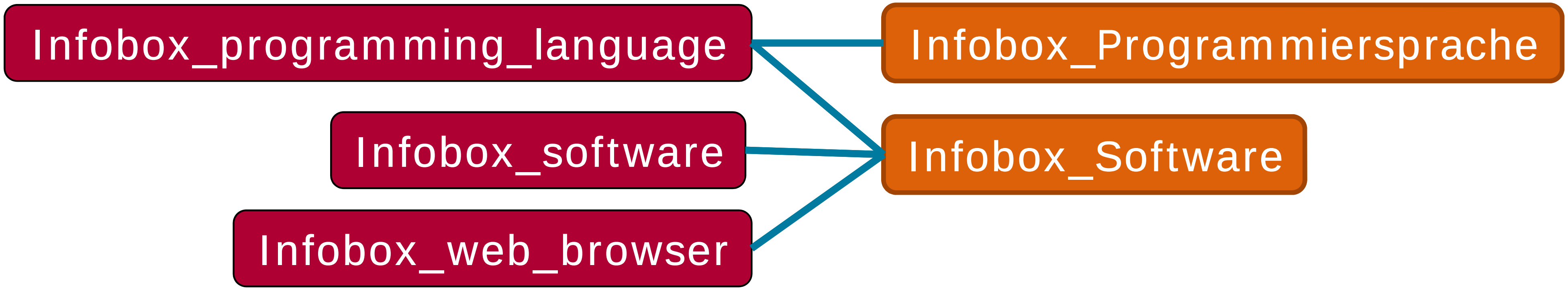
2CC

- Bi-connected components (2CC)
- Each pair of vertices is connected via **two** vertex-independent paths.
- 8,828 2CCs of which 4,770 are vertex-disjoint

- Result: 1,069,948 coherent, connected components

Infobox Template Mapping

- Problem: Match schemata only of **corresponding** templates.
- Different granularities in templates => n:m mapping
- **Idea:** Count co-occurrences of infobox templates in terms of connected components and apply thresholds:
  - □ **Absolute**: at least 5 co-occurrences
  - □ **Relative**: co-occurrence frequency at least 20% of individual occurrences of the templates

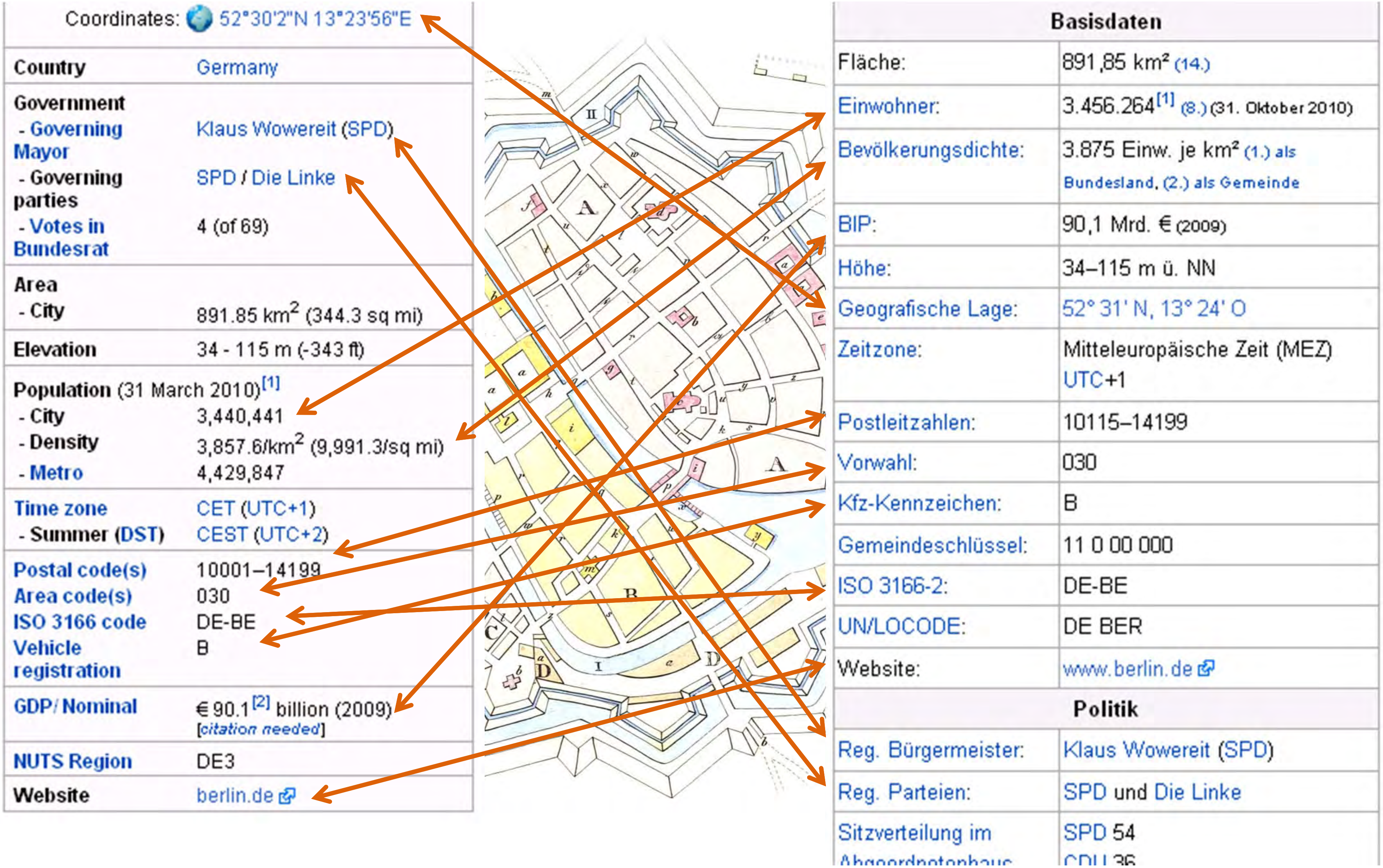


Duplicate-based Schema Matching

- General technique of data is available under both schemas
- Idea: If data coincides for attributes of two schemata, they probably match.

- For each infobox template pair
  - For each article pair
    - For each attribute value pair
      - Determine similarity of values (edit-distance)
      - Store in matrix
  - Aggregate similarities across all articles
  - Perform global matching: bipartite assignment

Duplicate-based Schema Matching



Coordinates:	52°30′2″N 13°23′56″E
Country	Germany
Government	
- Governing Mayor	Klaus Wowereit (SPD)
- Governing parties	SPD / Die Linke
- Votes in Bundesrat	4 (of 69)
Area	
- City	891.85 km^2 (344.3 sq mi)
Elevation	34 - 115 m (-343 ft)
Population (31 March 2010)[1]	
- City	3,440,441
- Density	3,857.6/km^2 (9,991.3/sq mi)
- Metro	4,429,847
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postal code(s)	10001–14199
Area code(s)	030
ISO 3166 code	DE-BE
Vehicle registration	B
GDP/ Nominal	€90.1[2] billion (2009) [citation needed]
NUTS Region	DE3
Website	berlin.de

Basisdaten	
Fläche:	891,85 km² (14.)
Einwohner:	3.456.264[1] (8.) (31. Oktober 2010)
Bevölkerungsdichte:	3.875 Einw. je km² (1.) als Bundesland, (2.) als Gemeinde
BIP:	90,1 Mrd. € (2009)
Höhe:	34–115 m ü. NN
Geografische Lage:	52° 31′ N, 13° 24′ O
Zeitzone:	Mitteleuropäische Zeit (MEZ) UTC+1
Postleitzahlen:	10115–14199
Vorwahl:	030
Kfz-Kennzeichen:	B
Gemeindeschlüssel:	11 0 00 000
ISO 3166-2:	DE-BE
UN/LOCODE:	DE BER
Website:	www.berlin.de
Politik	
Reg. Bürgermeister:	Klaus Wowereit (SPD)
Reg. Parteien:	SPD und Die Linke
Sitzverteilung im Abgeordnetenhaus	SPD 54 CDU 36

Evaluation

- Qualitative evaluation via hand-crafted attribute mappings
  - 96 infobox template pairs
  - 1,417 expected attribute pairs

%	en de	en fr	en nl	de fr	de nl	fr nl	Overall
Precision	91.97	92.28	95.15	90.78	91.67	93.85	92.64
Recall	94.17	96.83	94.80	92.06	93.22	92.82	94.21
F_1 Score	93.06	94.50	94.97	91.42	92.44	93.33	93.42

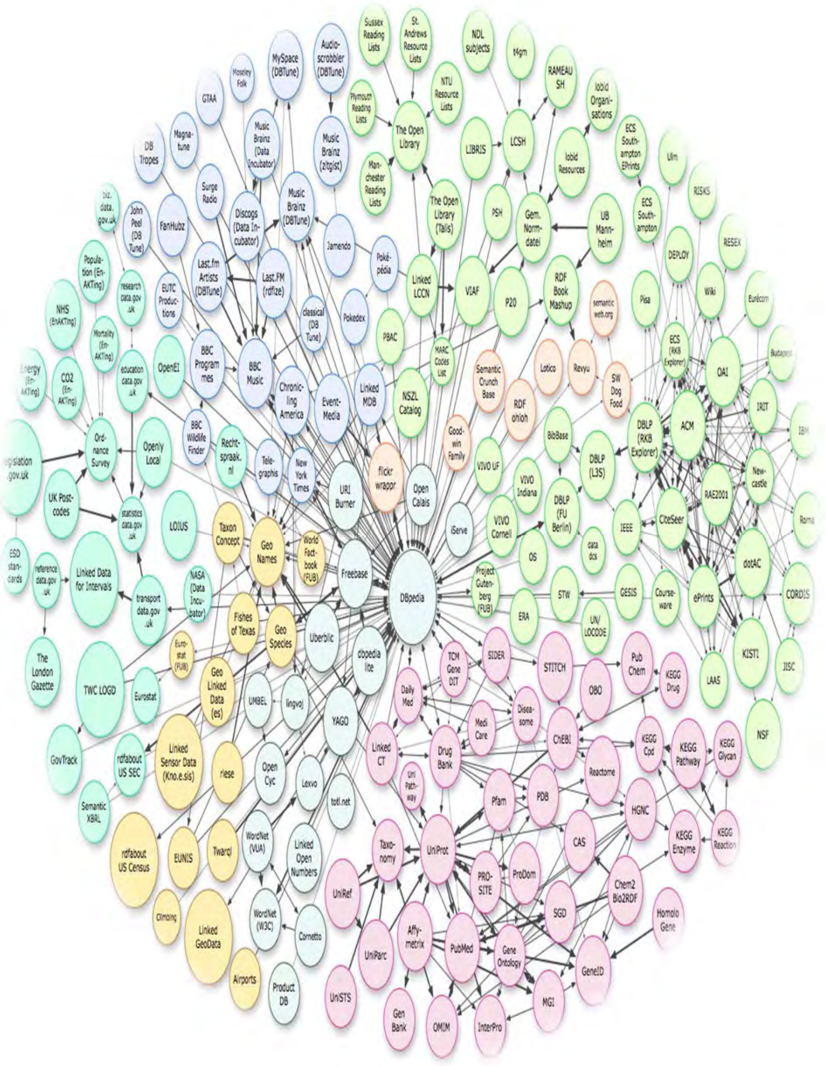
Next step by community: Wikidata

- Free knowledge base about the world
- Read and edited by humans and machines
- Data in all the languages of the Wikimedia projects
  - □ In particular: Wikipedia pages
- Central access to data

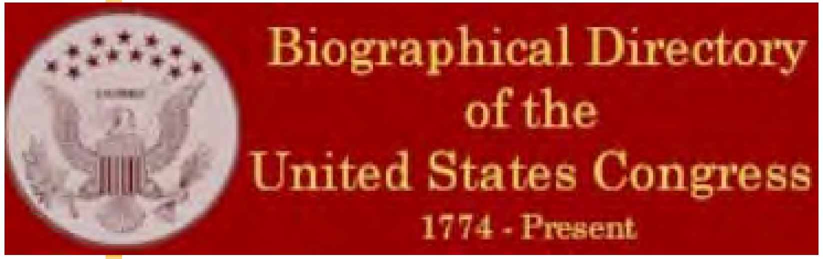
- Begin April 2012 – much to do
- http://meta.wikimedia.org/wiki/Wikidata/de

Overview

- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ ...of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



Motivation – Wealth of Open Gov Data


Biographical Directory of the United States Congress 1774 - Present

European Commission

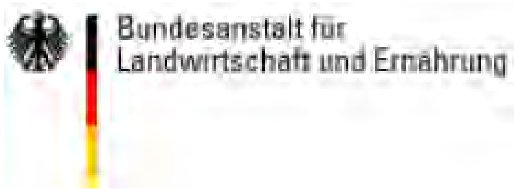
European Commission > Beneficiaries > Fts

Financial Transparency System

USASPENDING.GOV

Government spending at your fingertips

Welcome to the European Parliament

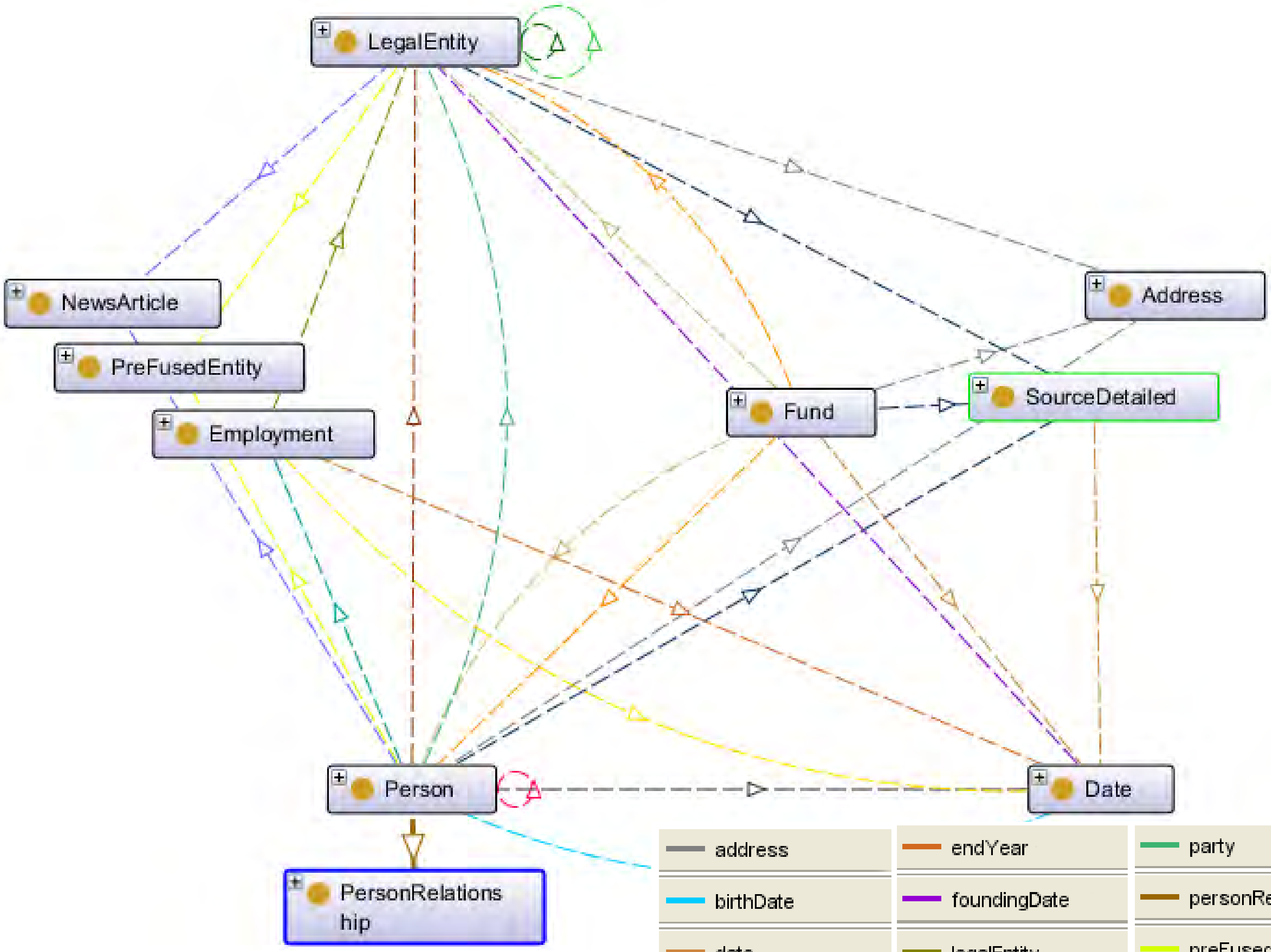

Bundesanstalt für Landwirtschaft und Ernährung

OFFICE OF MANAGEMENT AND BUDGET

Unklarheiten.de

Politische Datenbank

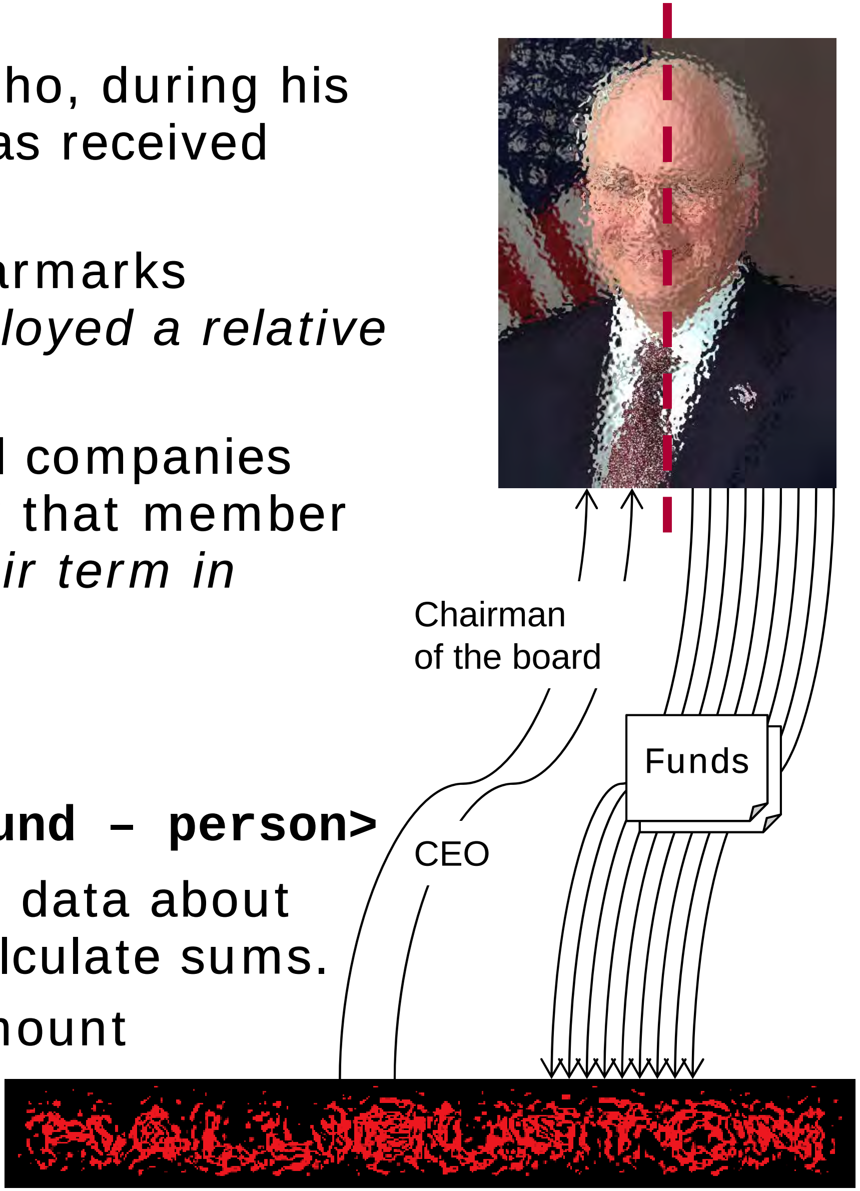
Companies, Agencies, and People



▬ address	▬ endYear	▬ party	▬ sourceDetailed
▬ birthDate	▬ foundingDate	▬ personRelationship	▬ sponsor
▬ date	▬ legalEntity	▬ preFusedEntity	▬ startYear
▬ deathDate	▬ newsArticle	▬ recipient	▬ subsidiary
▬ employment	▬ parent	▬ relatedPerson	▬ worksFor

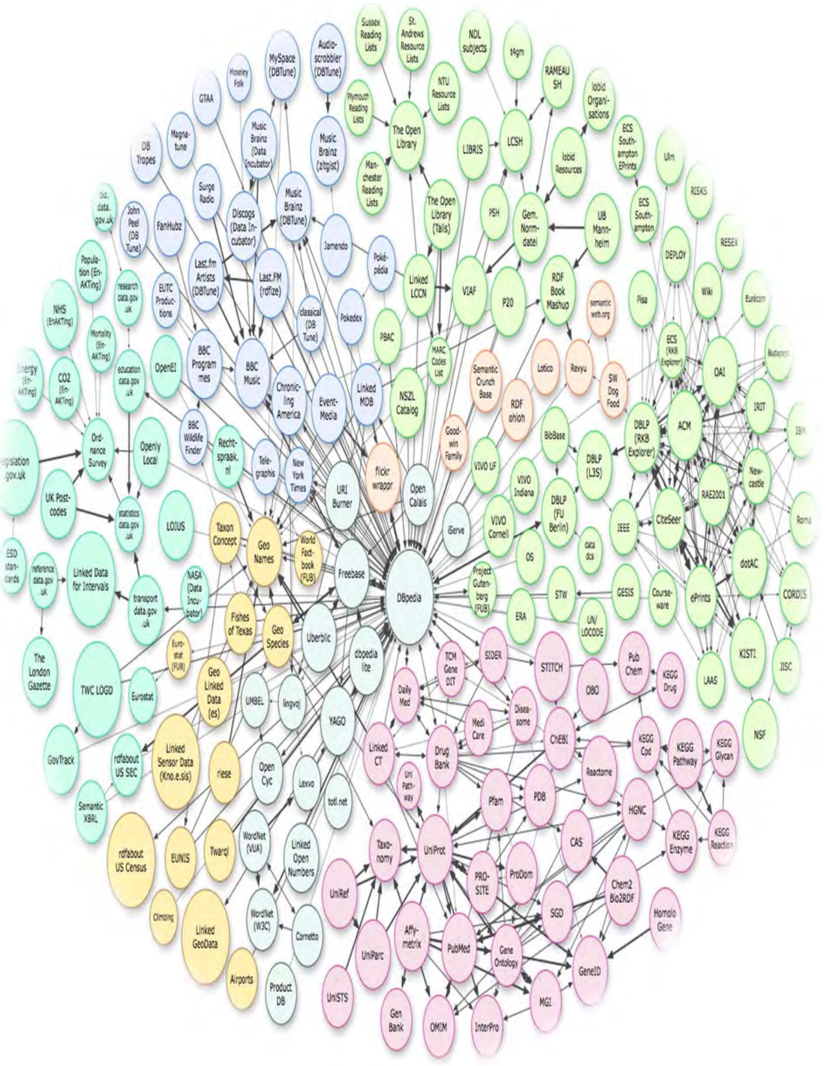
Interesting queries

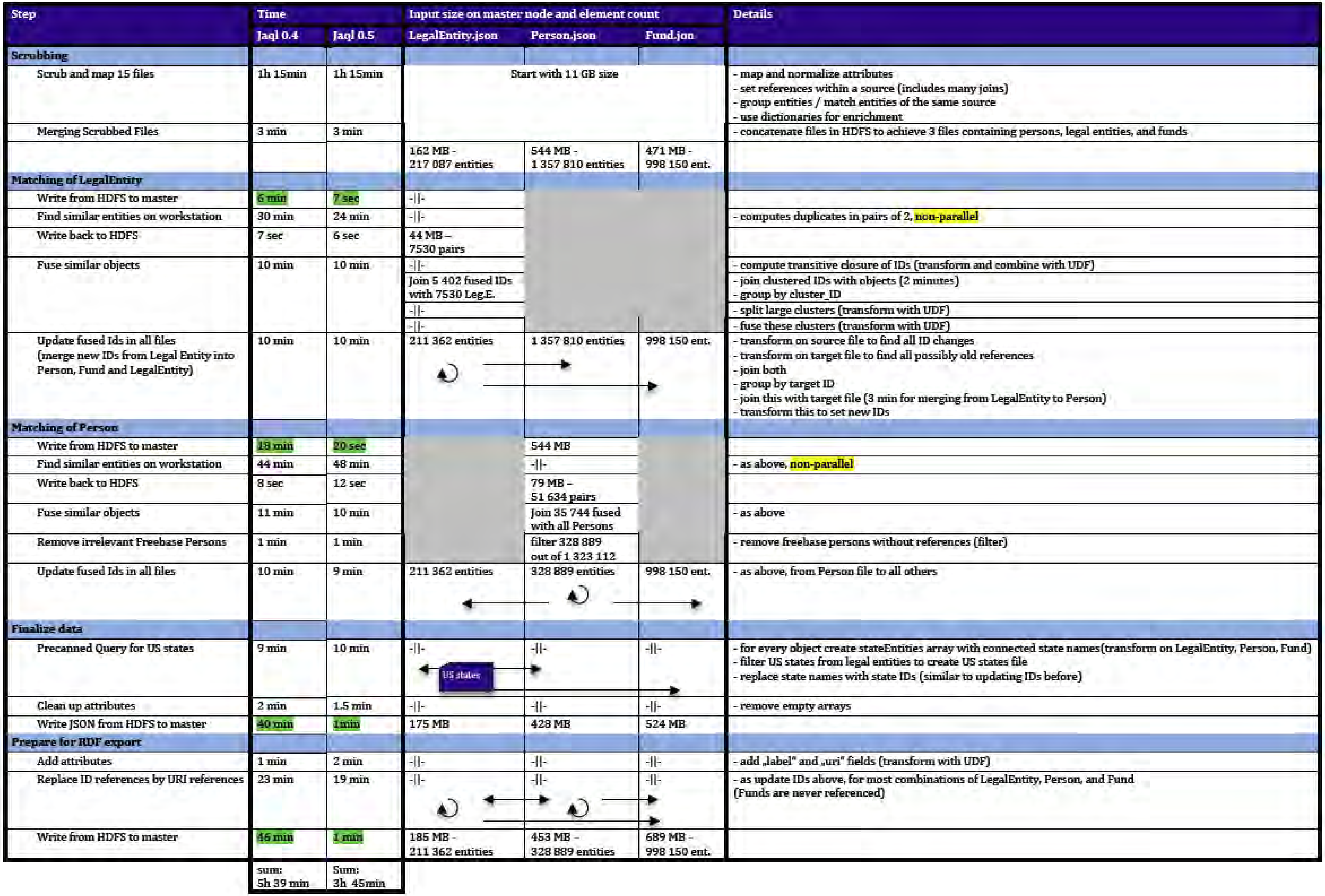
- Find all *classmates* of George W. Bush who, during his term, have worked at a company that has received government funding.
- For each member of congress, find all earmarks awarded to organizations that have *employed a relative* of that member of congress.
- For each government employees, find all companies that have received funding supported by that member and have *employed him after/before their term in congress*.
- Goal: Demonstrate the power of
  - □ *Joins*: Find unknown connections
    `<person – university|company|fund – person>`
  - □ *Grouping* and *aggregation*: Combine data about parties, companies, and persons; calculate sums.
  - □ *Sorting*: Order results by funding amount
  - □ *Sets*: “for each … find all …”



Overview

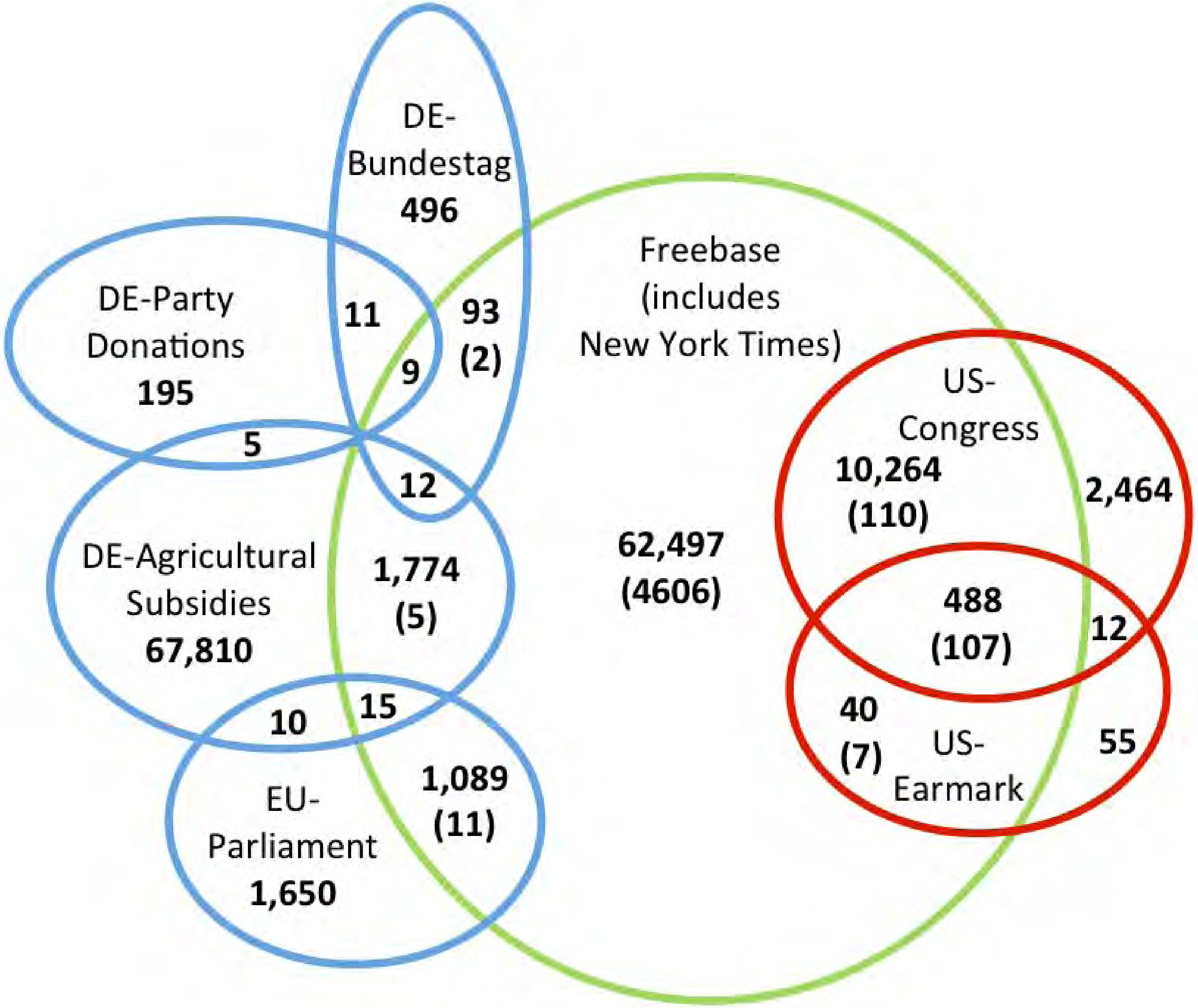
- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



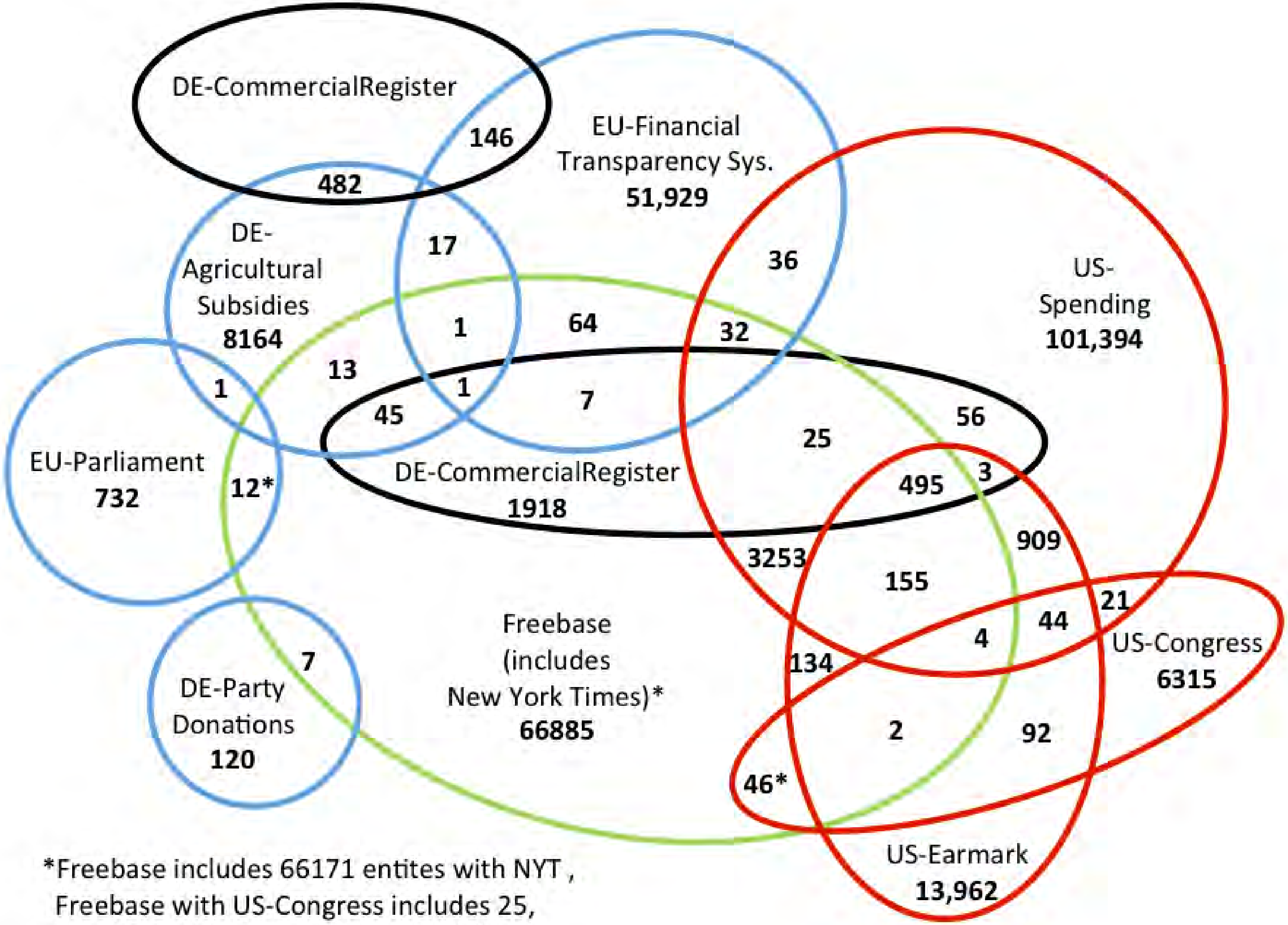


Step	Time		Input size on master node and element count			Details
	Jaql 0.4	Jaql 0.5	LegalEntity.json	Person.json	Fund.jon	
Scrubbing						
Scrub and map 15 files	1h 15min	1h 15min	Start with 11 GB size			- map and normalize attributes - set references within a source (includes many joins) - group entities / match entities of the same source - use dictionaries for enrichment
Merging Scrubbed Files	3 min	3 min				- concatenate files in HDFS to achieve 3 files containing persons, legal entities, and funds
			162 MB - 217 087 entities	544 MB - 1 357 810 entities	471 MB - 998 150 ent.	
Matching of LegalEntity						
Write from HDFS to master	6 min	7 sec	-||-			
Find similar entities on workstation	30 min	24 min	-||-			- computes duplicates in pairs of 2, non-parallel
Write back to HDFS	7 sec	6 sec	44 MB – 7530 pairs			
Fuse similar objects	10 min	10 min	-||- Join 5 402 fused IDs with 7530 Leg.E. -||- -||-			- compute transitive closure of IDs (transform and combine with UDF) - join clustered IDs with objects (2 minutes) - group by cluster_ID - split large clusters (transform with UDF) - fuse these clusters (transform with UDF)
Update fused Ids in all files (merge new IDs from Legal Entity into Person, Fund and LegalEntity)	10 min	10 min	211 362 entities	1 357 810 entities	998 150 ent.	- transform on source file to find all ID changes - transform on target file to find all possibly old references - join both - group by target ID - join this with target file (3 min for merging from LegalEntity to Person) - transform this to set new IDs
Matching of Person						
Write from HDFS to master	18 min	20 sec		544 MB		
Find similar entities on workstation	44 min	48 min		-||-		- as above, non-parallel
Write back to HDFS	8 sec	12 sec		79 MB – 51 634 pairs		
Fuse similar objects	11 min	10 min		Join 35 744 fused with all Persons		- as above
Remove irrelevant Freebase Persons	1 min	1 min		filter 328 889 out of 1 323 112		- remove freebase persons without references (filter)
Update fused Ids in all files	10 min	9 min	211 362 entities	328 889 entities	998 150 ent.	- as above, from Person file to all others
Finalize data						
Precanned Query for US states	9 min	10 min	-||- US states	-||-	-||-	- for every object create stateEntities array with connected state names(transform on LegalEntity, Person, Fund) - filter US states from legal entities to create US states file - replace state names with state IDs (similar to updating IDs before)
Clean up attributes	2 min	1.5 min	-||-	-||-	-||-	- remove empty arrays
Write JSON from HDFS to master	40 min	1min	175 MB	428 MB	524 MB	
Prepare for RDF export						
Add attributes	1 min	2 min	-||-	-||-	-||-	- add „label" and „uri" fields (transform with UDF)
Replace ID references by URI references	23 min	19 min	-||-	-||-	-||-	- as update IDs above, for most combinations of LegalEntity, Person, and Fund (Funds are never referenced)
Write from HDFS to master	46 min	1 min	185 MB - 211 362 entities	453 MB – 328 889 entities	689 MB - 998 150 ent.	
	sum: 5h 39 min	Sum: 3h 45min				

Persons


DE-
Bundestag
496
DE-Party
Donations
195
11
93
(2)
9
Freebase
(includes
New York Times)
5
12
DE-Agricultural
Subsidies
67,810
1,774
(5)
62,497
(4606)
US-
Congress
10,264
(110)
2,464
488
(107)
12
10
15
1,089
(11)
EU-
Parliament
1,650
40
(7)
US-
Earmark
55

Organisations

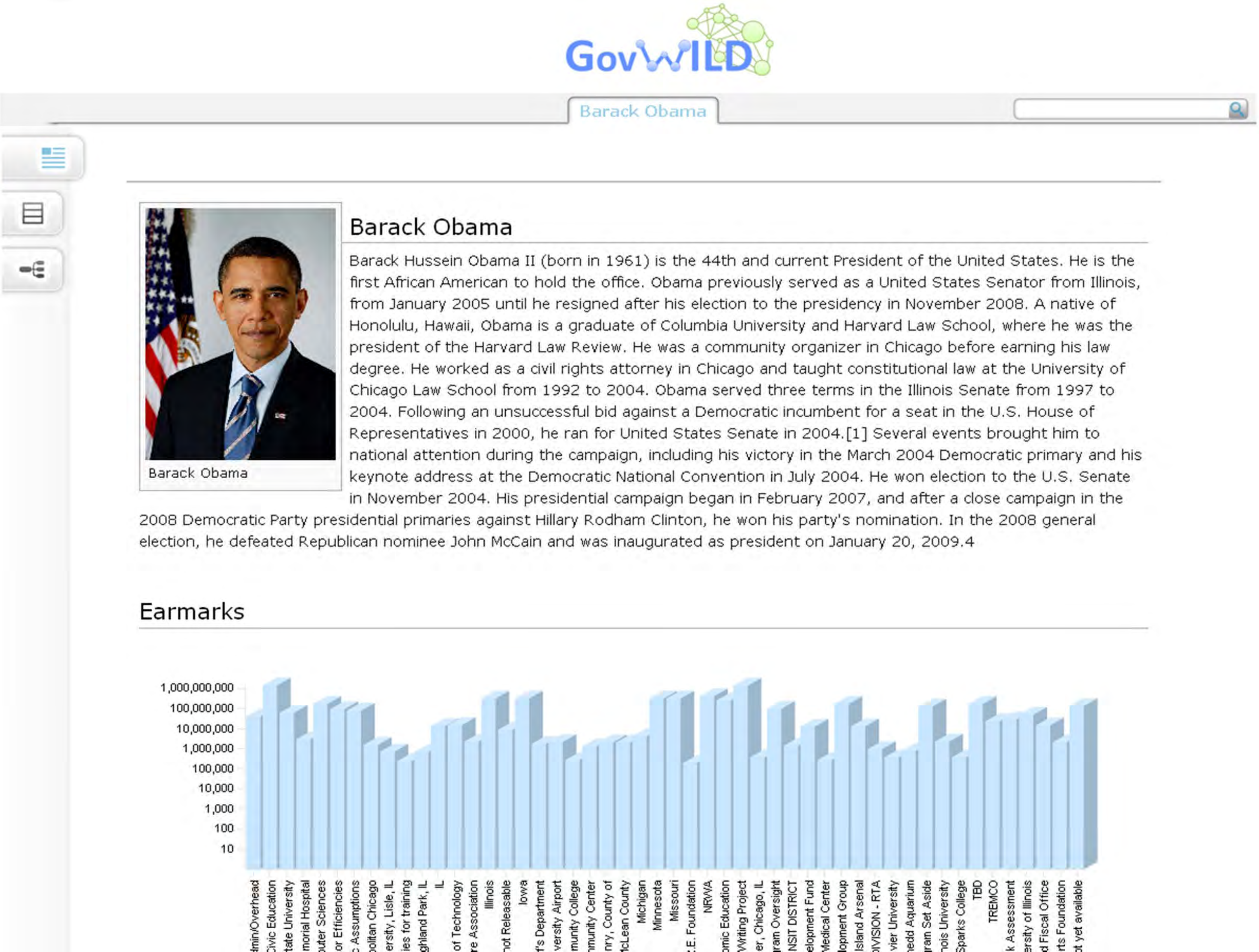


*Freebase includes 66171 entites with NYT ,
Freebase with US-Congress includes 25,
Freebase with EU-Parliament includes 0.

http://govwild.org

- 150,000 persons
- 270,000 legal entities
- 1,100,000 funds
- 43,000,000 triples

- Keyword Queries
- Linked Data Interface (dereference URIs)
- Exploration of entities mentioned in New York Times articles
- Data Download (RDF, SQL Dump, JSON files)



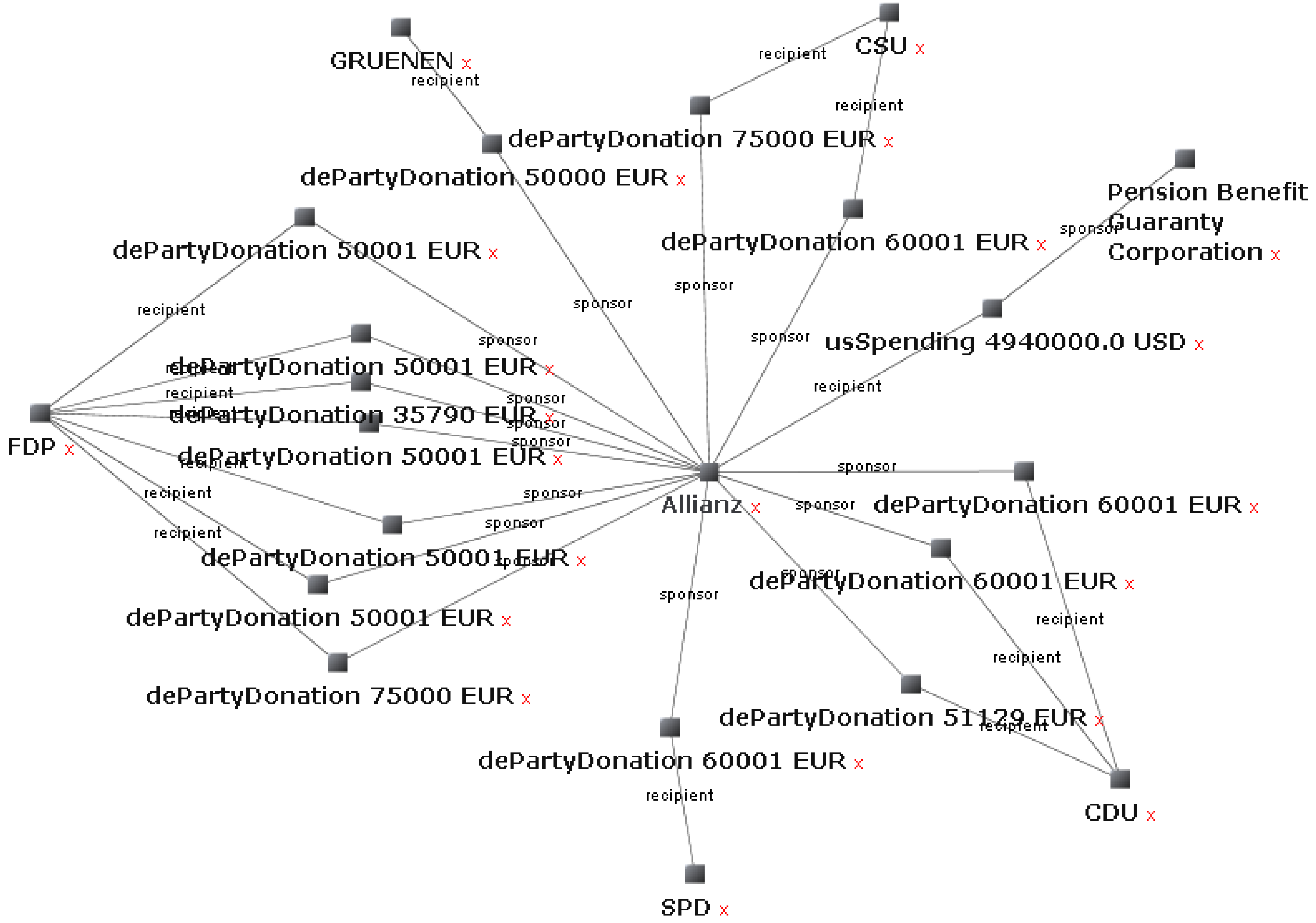
Barack Obama

Barack Obama

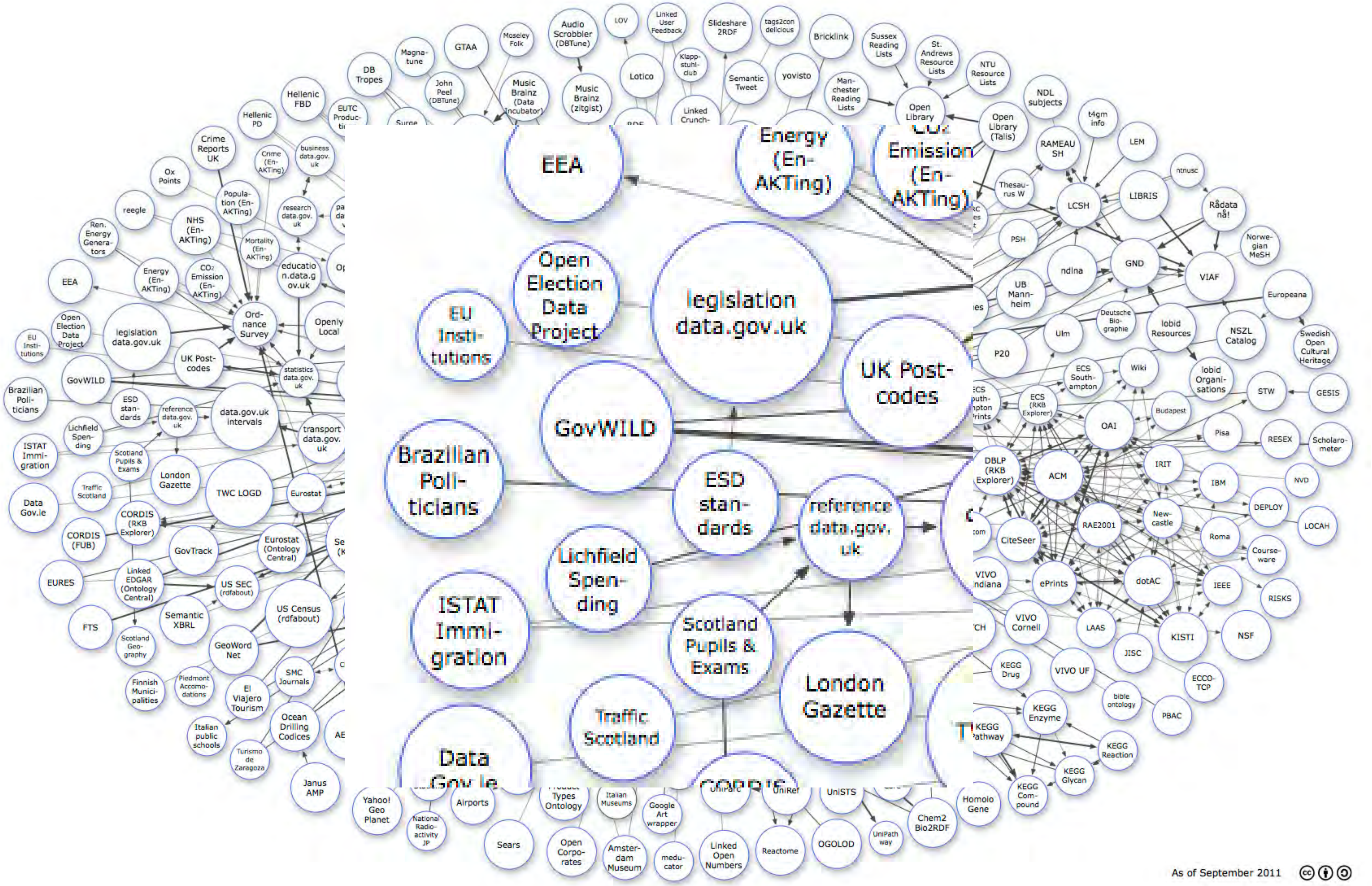
Barack Hussein Obama II (born in 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as a United States Senator from Illinois, from January 2005 until he resigned after his election to the presidency in November 2008. A native of Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. Obama served three terms in the Illinois Senate from 1997 to 2004. Following an unsuccessful bid against a Democratic incumbent for a seat in the U.S. House of Representatives in 2000, he ran for United States Senate in 2004.[1] Several events brought him to national attention during the campaign, including his victory in the March 2004 Democratic primary and his keynote address at the Democratic National Convention in July 2004. He won election to the U.S. Senate in November 2004. His presidential campaign began in February 2007, and after a close campaign in the 2008 Democratic Party presidential primaries against Hillary Rodham Clinton, he won his party's nomination. In the 2008 general election, he defeated Republican nominee John McCain and was inaugurated as president on January 20, 2009.4

Earmarks

Allianz insurance


CSU x
GRUENEN x
recipient
recipient
recipient
dePartyDonation 75000 EUR x
dePartyDonation 50000 EUR x
Pension Benefit Guaranty Corporation x
sponsor
dePartyDonation 60001 EUR x
dePartyDonation 50001 EUR x
sponsor
sponsor
recipient
sponsor
sponsor
usSpending 4940000.0 USD x
dePartyDonation 50001 EUR x
recipient
recipient
dePartyDonation 35790 EUR x
sponsor
recipient
sponsor
sponsor
FDP x
dePartyDonation 50001 EUR x
recipient
sponsor
sponsor
recipient
recipient
sponsor
sponsor
Allianz x
sponsor
dePartyDonation 60001 EUR x
dePartyDonation 50001 EUR x
sponsor
dePartyDonation 60001 EUR x
sponsor
dePartyDonation 50001 EUR x
recipient
recipient
dePartyDonation 75000 EUR x
dePartyDonation 51129 EUR x
recipient
dePartyDonation 60001 EUR x
recipient
CDU x
SPD x

Linked Data Graph

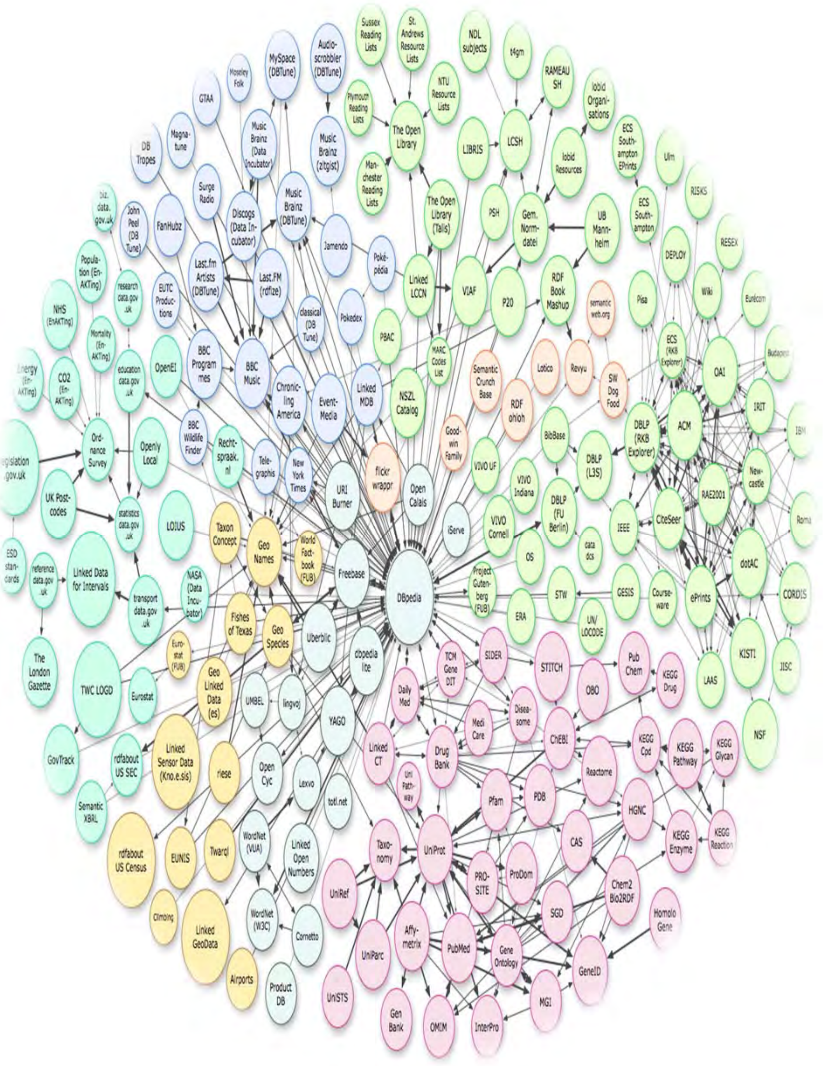


As of September 2011

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

Summary

- Web Data abounds
  - □ Linked, open, and otherwise
  - □ iPopulator
- Web Data stinks
  - □ Dirt, grime, and some surprises
  - □ ProLOD – Profiling LOD
- Cleansing and Integration
  - □ …of mops and brooms
  - □ Cross-language integration
- Government data
  - □ Politicians, friends, and funds
  - □ The GovWILD experience



References

- Extracting Structured Information from Wikipedia Articles to Populate Infoboxes
  Dustin Lange, Christoph Böhm, and Felix Naumann
  Proceedings of the 19th Conference on Information and Knowledge Management (CIKM) 2010, Toronto, Canada
  (Extended version available as technical report)
- Profiling Linked Open Data with ProLOD
  Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, David Sonnabend
  Workshop *New Trends in Information Integration* (NTII) 2010, Long Beach, USA
- Linking Open Government Data: What Journalists Wish They Had Known
  Christoph Böhm, Felix Naumann, Markus Freitag, Stefan George, Norman Höfler, Martin Köppelmann, Claudia Lehmann, Andrina Mascher, and Tobias Schmidt.
  Honorable Mention at Linked Data Triplification Challenge 2010 @ I-Semantics, Graz. (link to GovWILD)
- DuDe: The Duplicate Detection Toolkit
  Uwe Draisbach and Felix Naumann: QDB 2010 Workshop at VLDB, Singapore