Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

The Quality of Web Data

Keynote @ ICIQ 2012

17.11.2012

Felix Naumann (on sabbatical at QCRI)

The Quality of Web Data

Keynote @ ICIQ 2012

17.11.2012

Felix Naumann (on leave from HPI)

My path through ICIQ

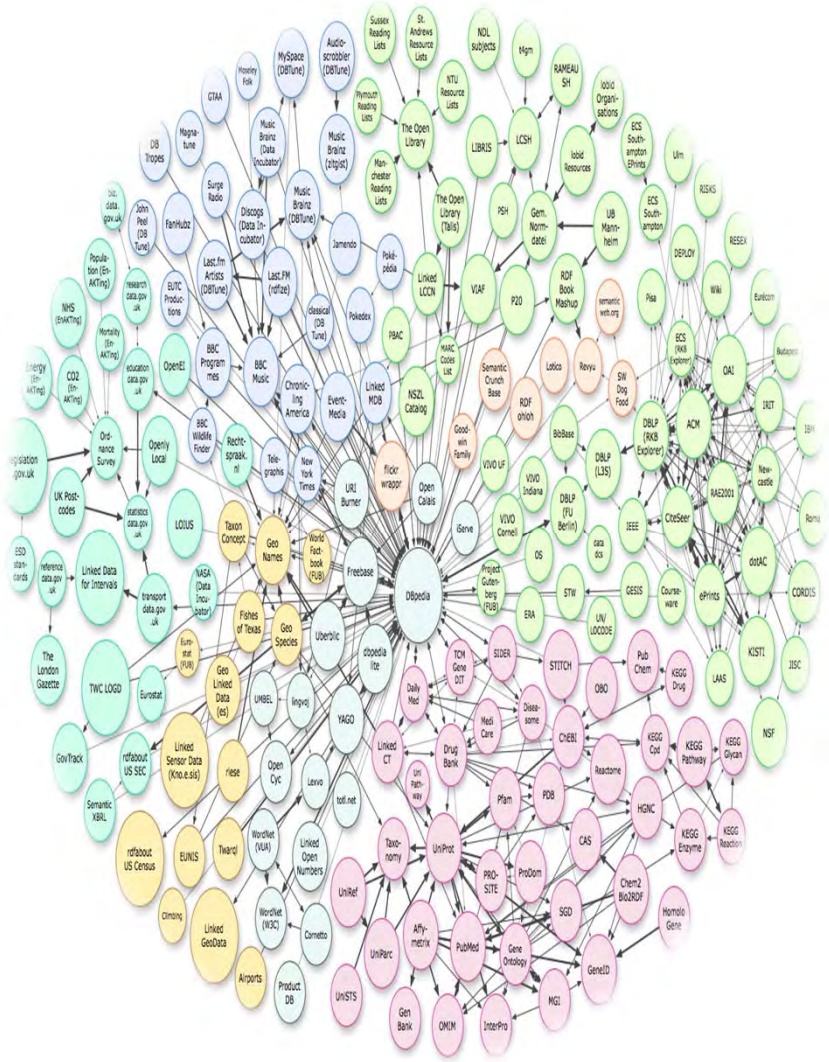
- Attendant 1998, 2000 - 2002, 2004 - 2006, 2009, 2012
- Author 1998 - 2004, 2007
- Presenter 1998, 2000 - 2002, 2004 – 2006, 2012
- Session chair every time…
- PC chair 2005 with Michael Gertz and Stu Madnick
- General chair 2009 at HPI in Potsdam
- Keynote speaker 2012

- Additional job „Waiting staff at the conference banquet“ (page 10)

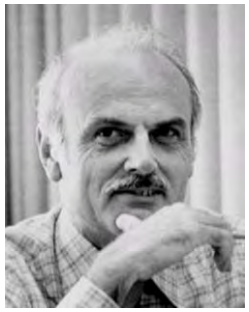
Thanks to the ICIQ community for a great ride!

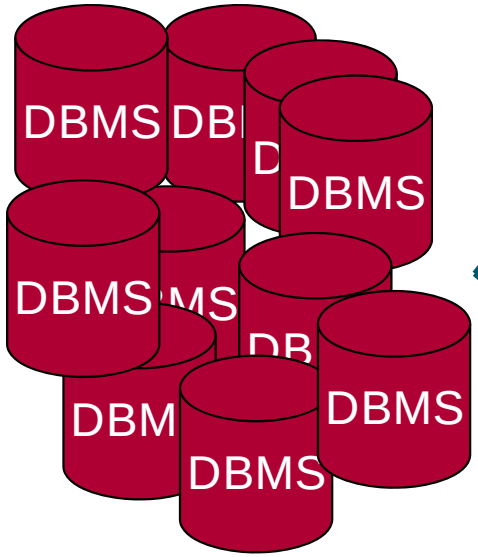
Overview

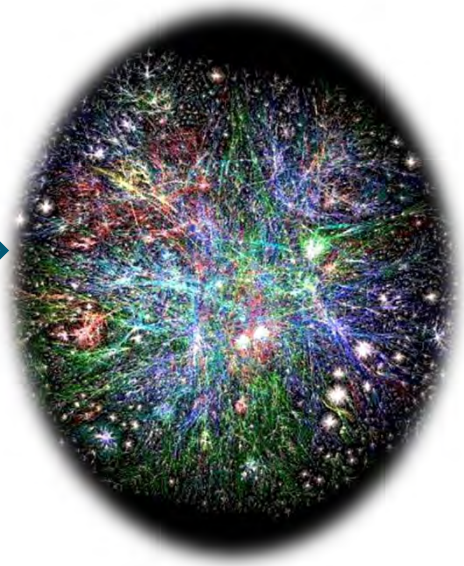
- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - …of mops and brooms
  - Cross-language integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience

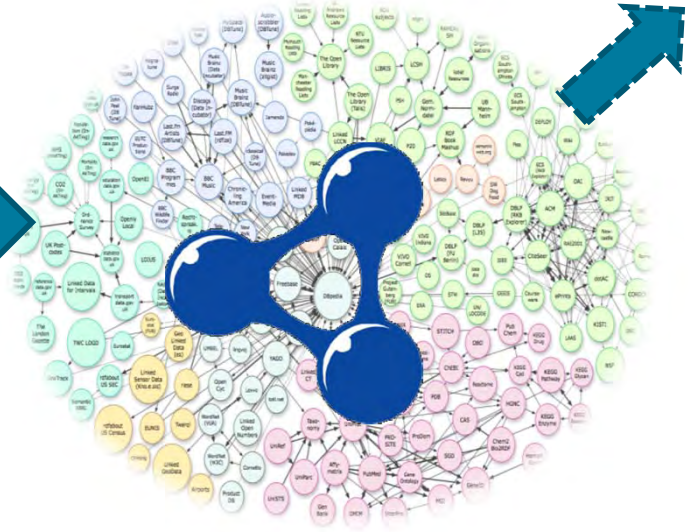


A brief history of data







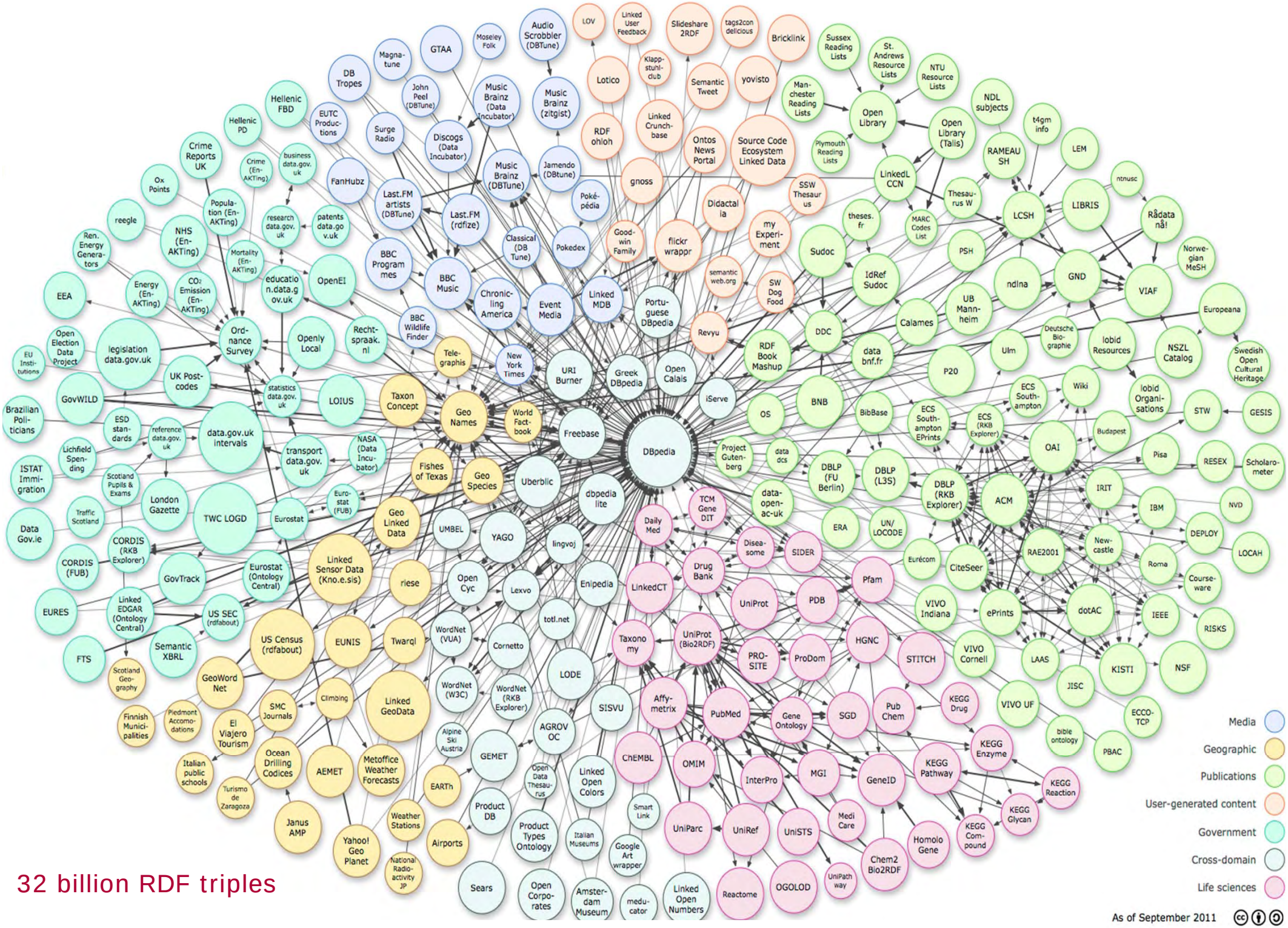


Linked Data & Data Spaces: A database guy‘s point-of-view

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



32 billion RDF triples

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

Wikipedia Infoboxes


```
{{Infobox company
|name             = International Business Machines Corporation
|logo             = <br />[[File:IBM logo.svg|200px]]<br />
|caption          = Logo since 1972, designed by [[Paul Rand]]
|type             = [[Public company|Public]]
|traded_as        = {{New York Stock Exchange|IBM}}<br />[[Dow Jones Industrial Average|Dow Jones Compon
Component]]
|industry         = [[Personal computer hardware|Computer hardware]], [[Software|Computer software]], [[
services]], [[Information technology consulting|IT consulting]]
|products         = [[List of IBM products|See IBM products]]
|founder          = [[Charles Ranlett Flint]]
|foundation       = [[Endicott, New York|Endicott]], New York, U.S.<br />({{Start date|1911|06|16}})
|location_city    = [[Armonk, New York|Armonk]], New York
|location_country = U.S.
|area_served      = Worldwide
|key_people       = [[Ginni Rometty]]<br />({{small|Chairman, President, and CEO}})
|revenue          = {{Increase}} US$ 106.91 [[1000000000 (number)|billion]] <small>(2011)</small><ref na
|url=http://rcpmag.com/articles/2012/01/20/intel-ibm-exceed-earnings-estimates-google-falls-short.aspx|
International Business Machines Corporation |work=United States Securities and Exchange Commission}}</re
|operating_income = {{Increase}} US$ {{0|0}}20.28 billion <small>(2011)</small><ref name=10K/>
|net_income       = {{Increase}} US$ {{0|0}}15.85 billion <small>(2011)</small><ref name=10K/>
|assets           = {{Increase}} US$ 116.43 billion <small>(2011)</small><ref name=10K/>
|equity           = {{Decrease}} US$ {{0|0}}20.13 billion <small>(2011)</small><ref name=10K/>
|num_employees    = 433,362 <small>(2012)</small><ref name="Fortune 500: IBM employees"/>
```

International Business Machines Corporation

Logo since 1972, designed by Paul Rand

Type	Public
Traded as	NYSE: IBM Dow Jones Component S&P 500 Component
Industry	Computer hardware, Computer software, IT services, IT consulting
Founded	Endicott, New York, U.S. (June 16, 1911)
Founder(s)	Charles Ranlett Flint
Headquarters	Armonk, New York, U.S.
Area served	Worldwide
Key people	Ginni Rometty (Chairman, President, and CEO)
Products	See IBM products
Revenue	▲ US$ 106.91 billion (2011)[1]
Operating income	▲ US$ 20.28 billion (2011)[1]
Net income	▲ US$ 15.85 billion (2011)[1]

DBpedia statistics

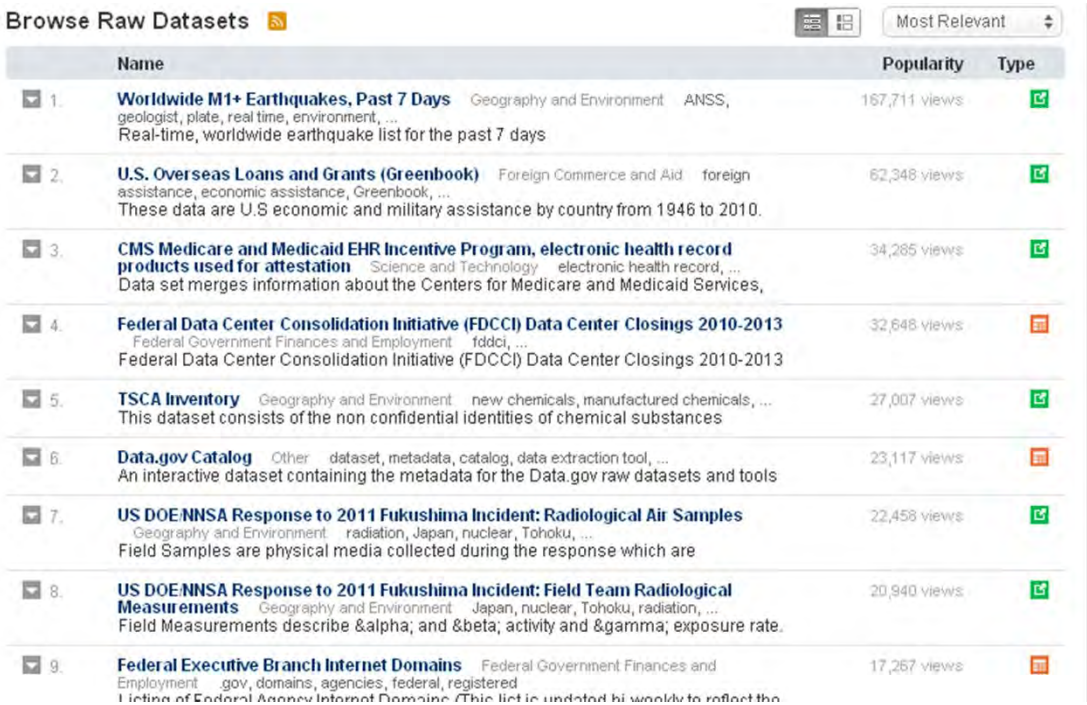
1. Core Datasets

Dataset	en	de	fr	es	it	pl	nl	pt	sv	ja	zh	fi	ro
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

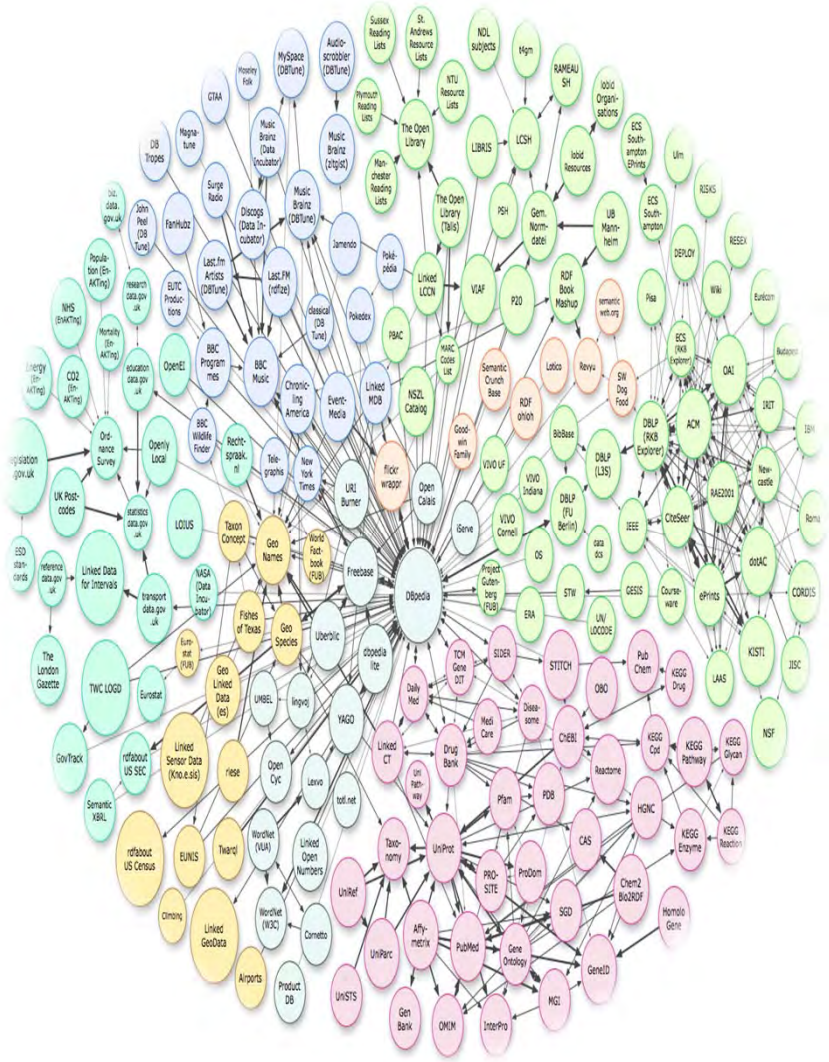
- Government data
  - www.data.gov (380k data sets)
  - data.gov.uk (9k)
  - ec.europa.eu/eurostat
- Finance / business data
- Scientific databases
  - www.uniprot.org
  - skyserver.sdss.org
- The Web
  - HTML tables and lists > 1billion (estimated Feb. 2011)
  - General sources: Dbpedia (3.7m), freebase (23m), ...
  - Domain-specific sources: IMDB, Gracenote, isbndb, ...

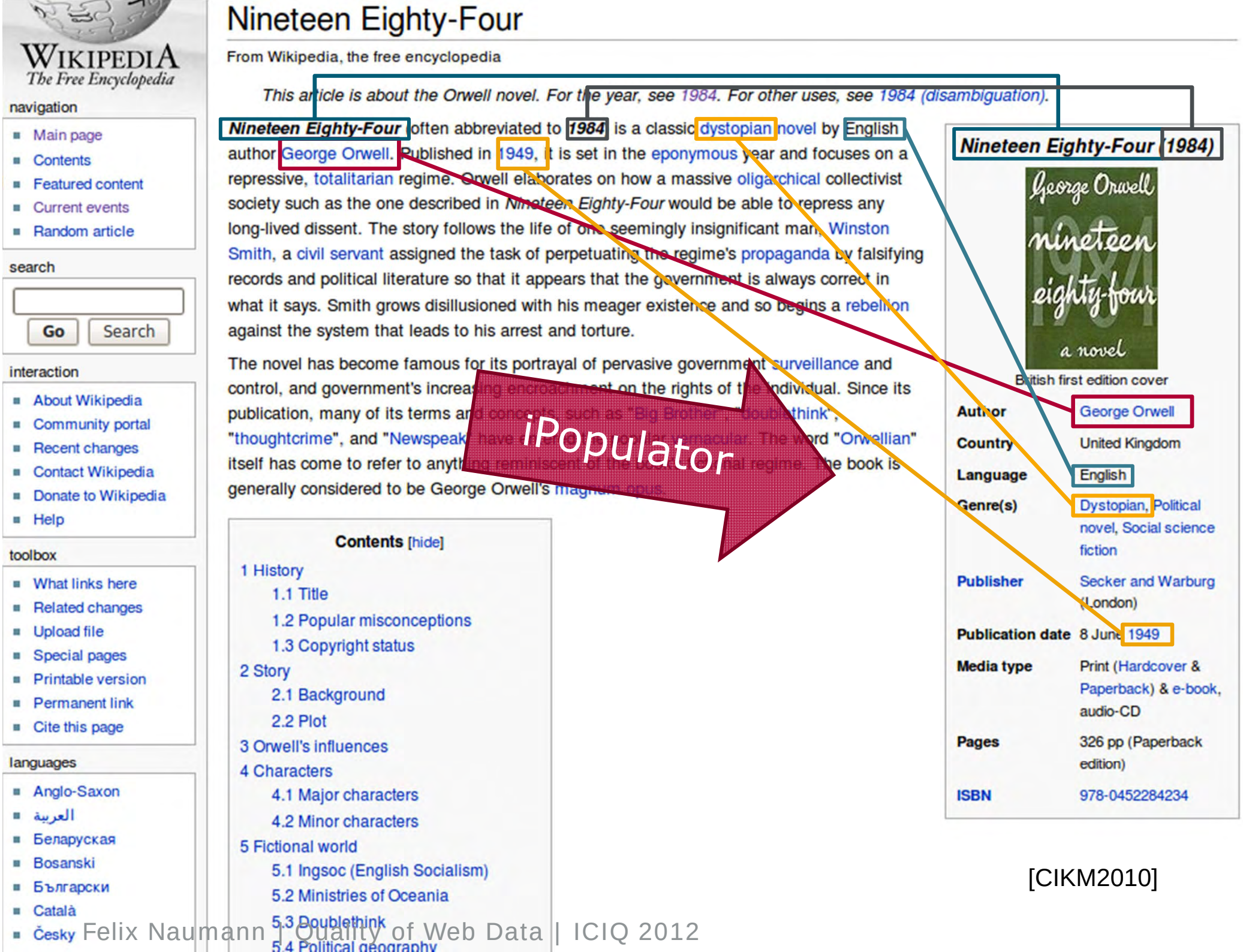


Killer app?

Overview

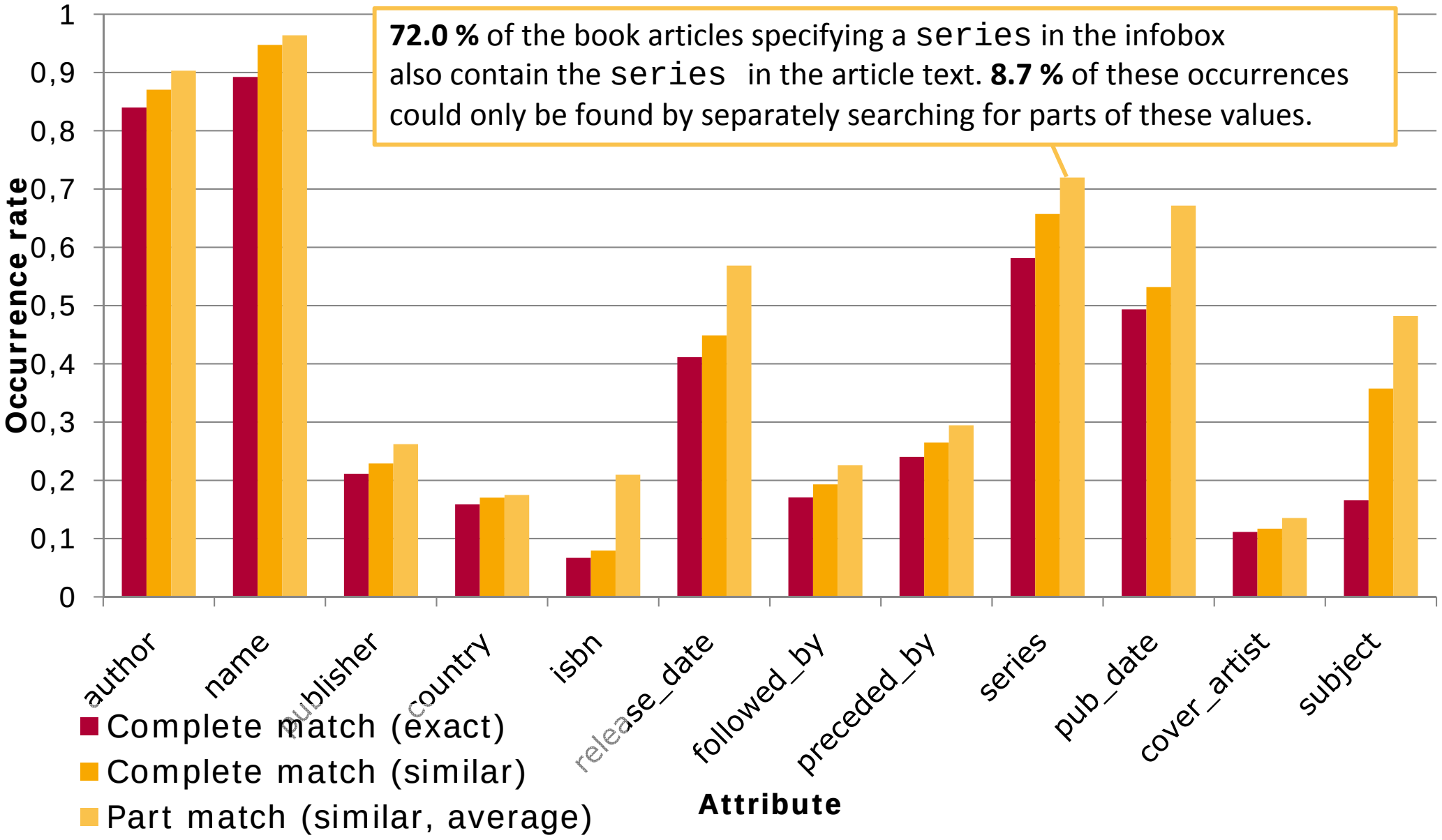
- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - …of mops and brooms
  - Cross-language integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience



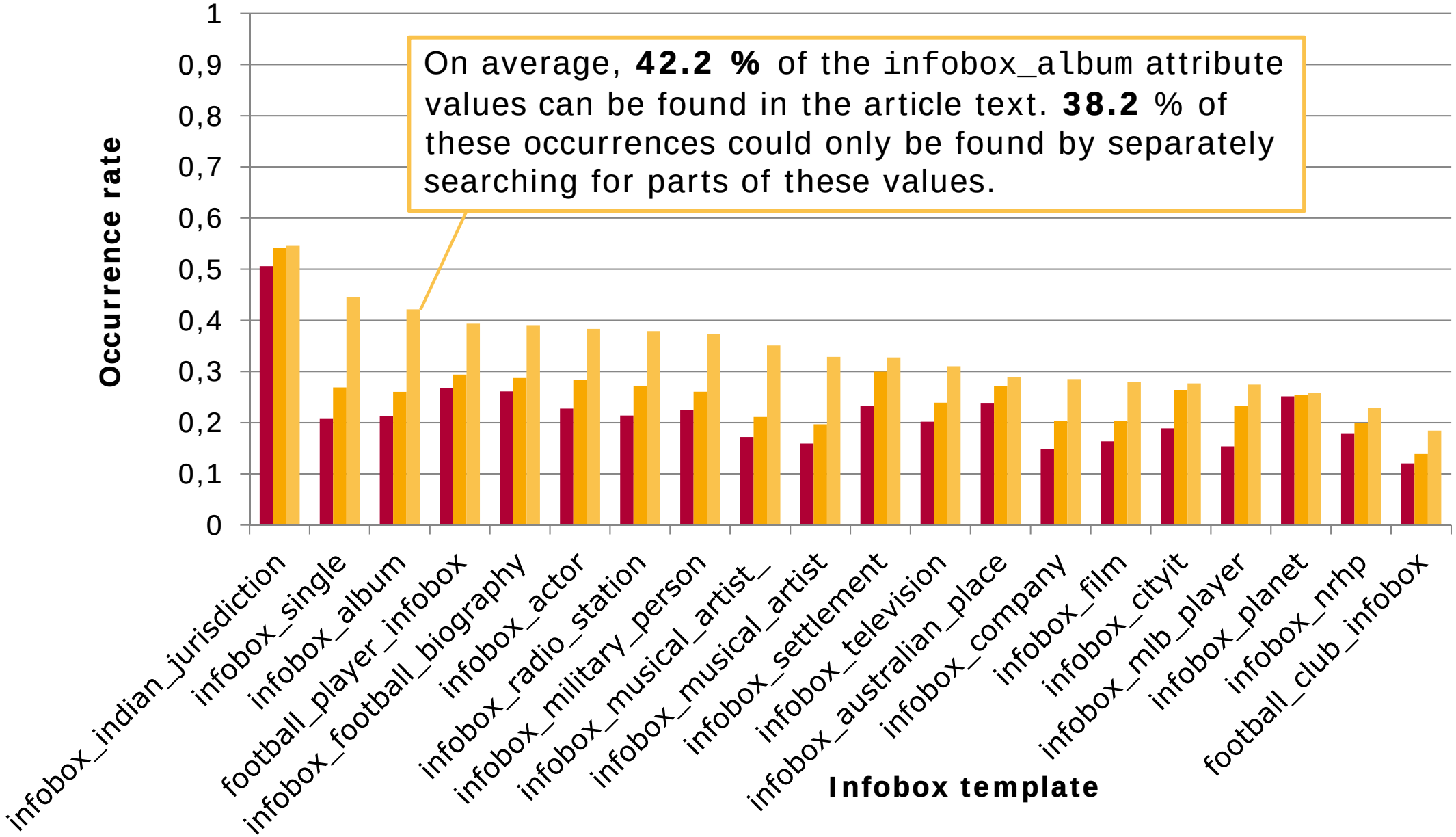

WIKIPEDIA
The Free Encyclopedia
navigation
Main page
Contents
Featured content
Current events
Random article
search
Go
Search
interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help
toolbox
What links here
Related changes
Upload file
Special pages
Printable 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. Since its publication, many of its terms and concepts, such as "Big Brother", "doublethink", "thoughtcrime", and "Newspeak" have entered into everyday vernacular. The word "Orwellian" itself has come to refer to anything reminiscent of the totalitarian regime. The book is generally considered to be George Orwell's magnum opus.
Contents [hide]
1 History
1.1 Title
1.2 Popular misconceptions
1.3 Copyright status
2 Story
2.1 Background
2.2 Plot
3 Orwell's influences
4 Characters
4.1 Major characters
4.2 Minor characters
5 Fictional world
5.1 Ingsoc (English Socialism)
5.2 Ministries of Oceania
5.3 Doublethink
5.4 Political geography
iPopulator
Nineteen Eighty-Four (1984)
George Orwell
nineteen eighty-four
a novel
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

[CIKM2010]

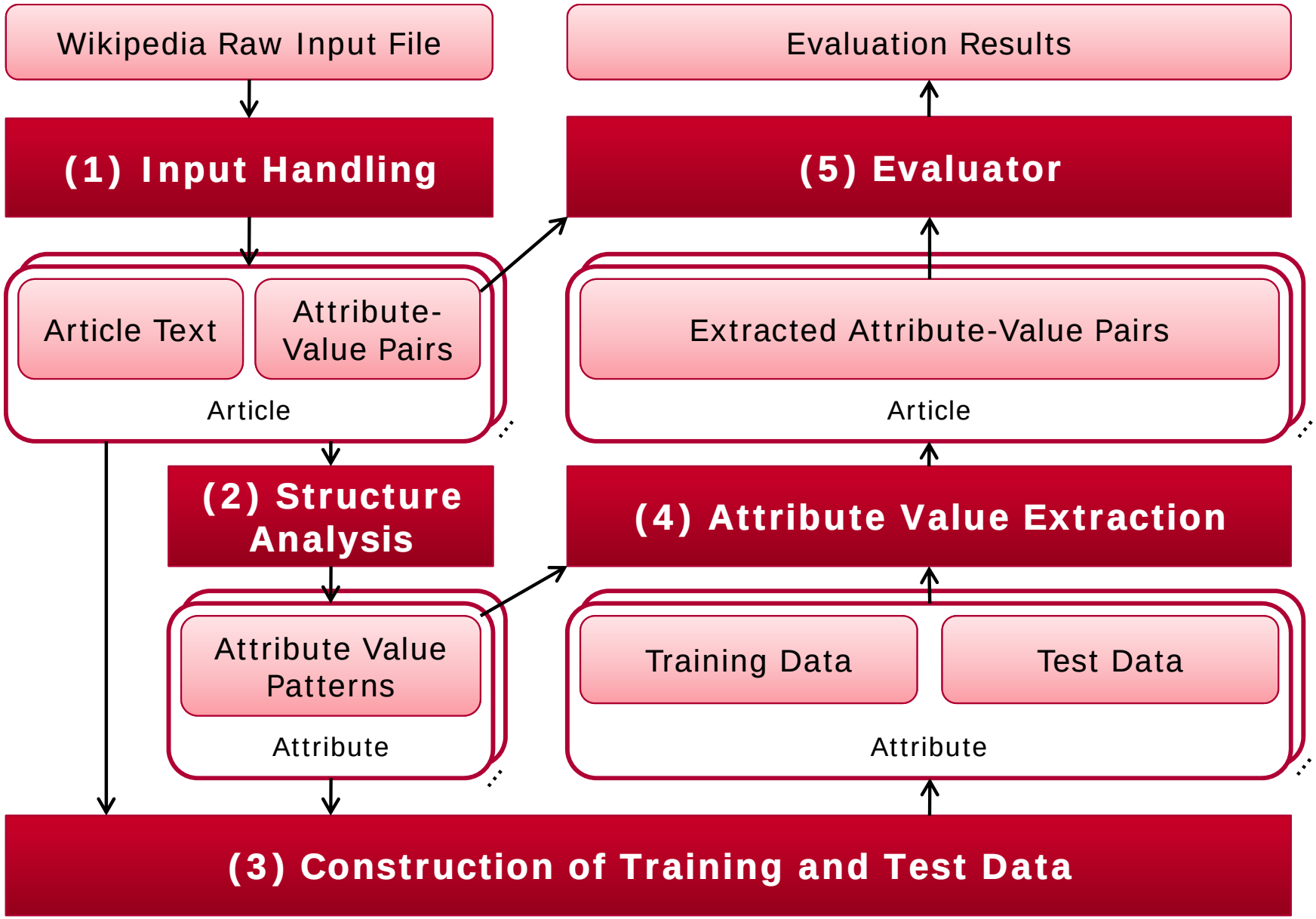
Occurrence of values in article text: 12 most frequent attributes in infobox_book



20 most frequent templates

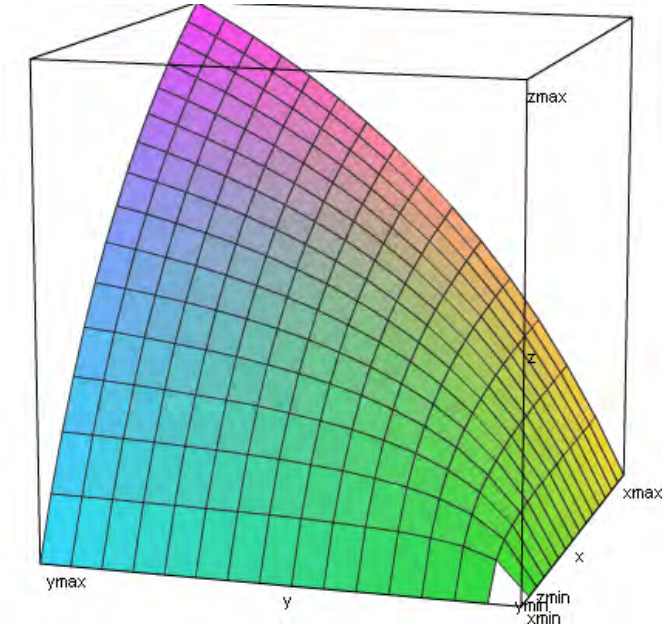


Architecture of iPopulator



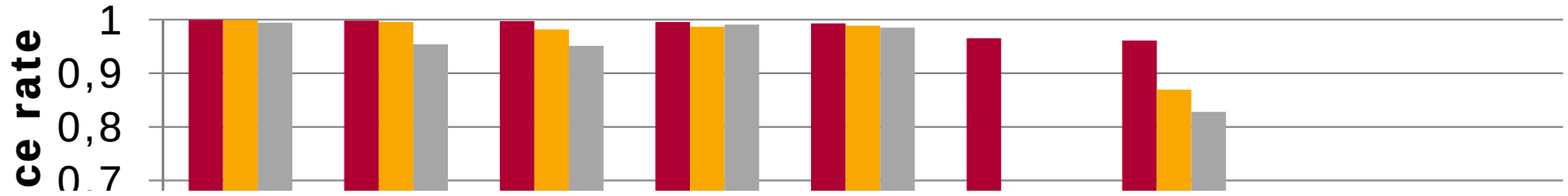
Learning infobox values with iPopulator

- Wikipedia Raw Input File
- Structure analysis
  - Attribute value patterns: **Employees** 433,362 (2012)[2]
- Construct training data
  - Hide some infobox data
- Machine learning
  - One extractor per attribute
- Extract values for previously hidden parts
  - Evaluate precision, recall, and F-measure



$z = 2\,(x \cdot y) / (x + y)$

Evaluation: infobox_planet

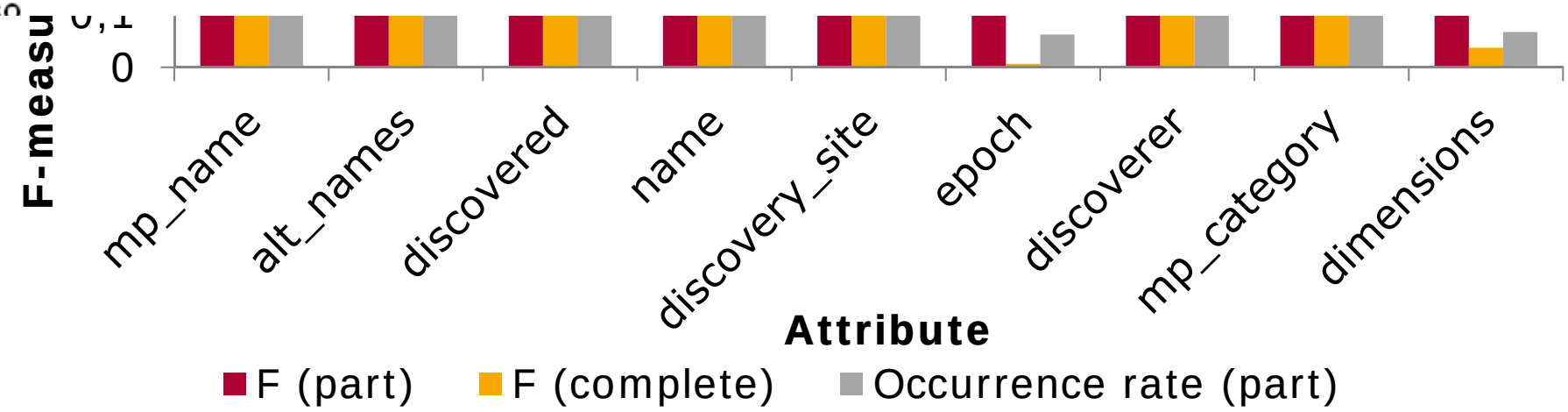


22032 Mikekoop

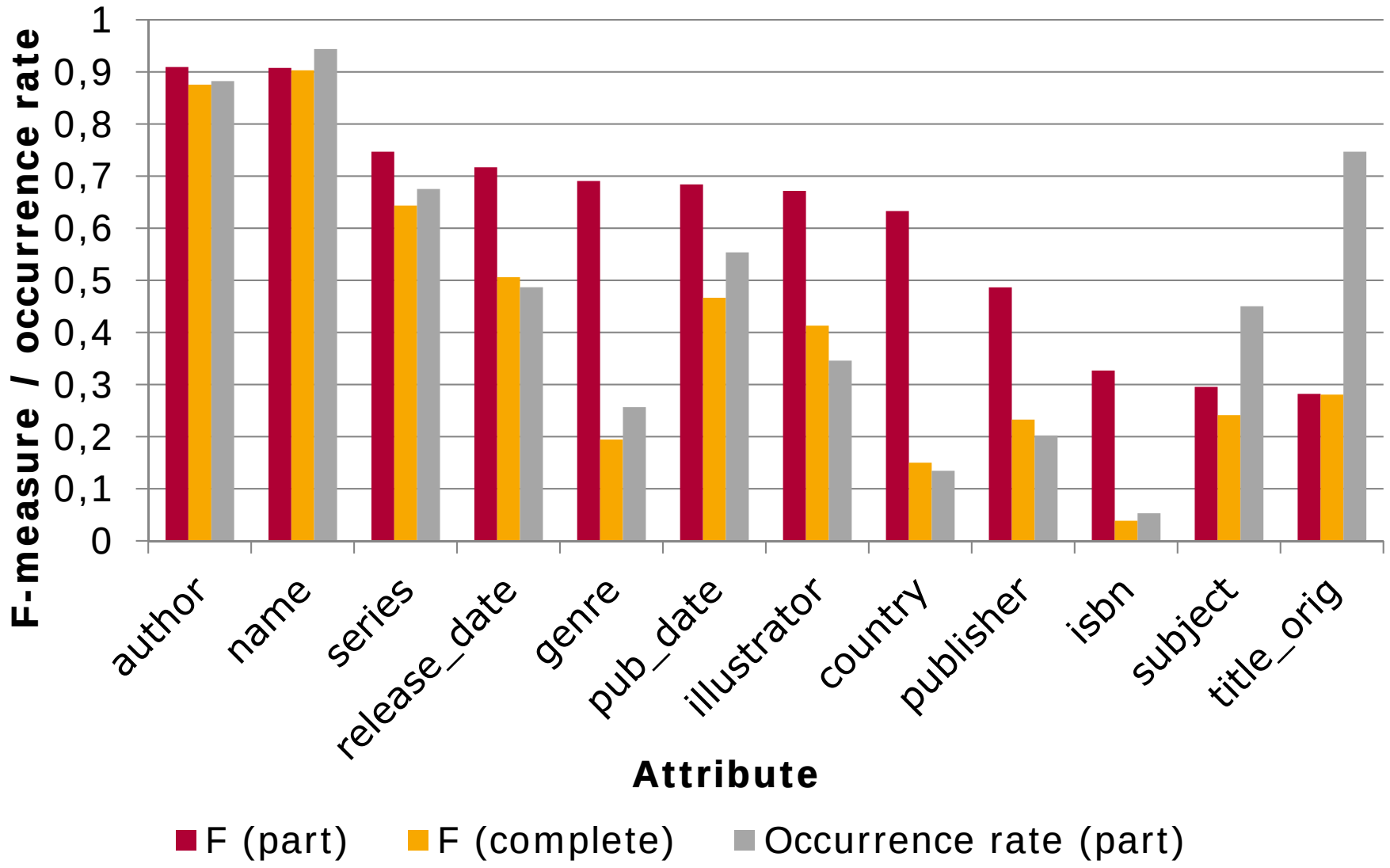
From Wikipedia, the free encyclopedia

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.

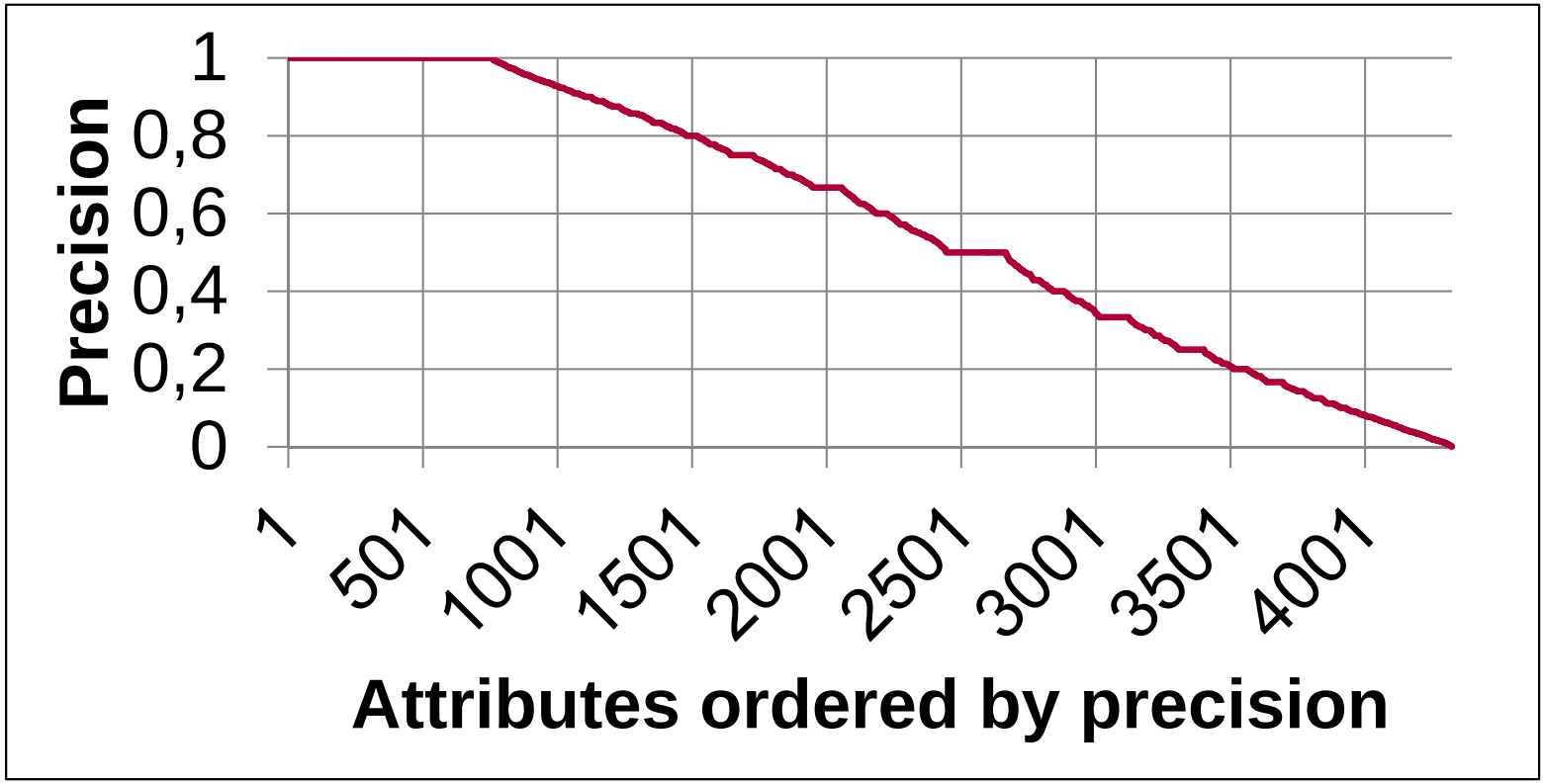
See also [edit]



Evaluation: infobox_book



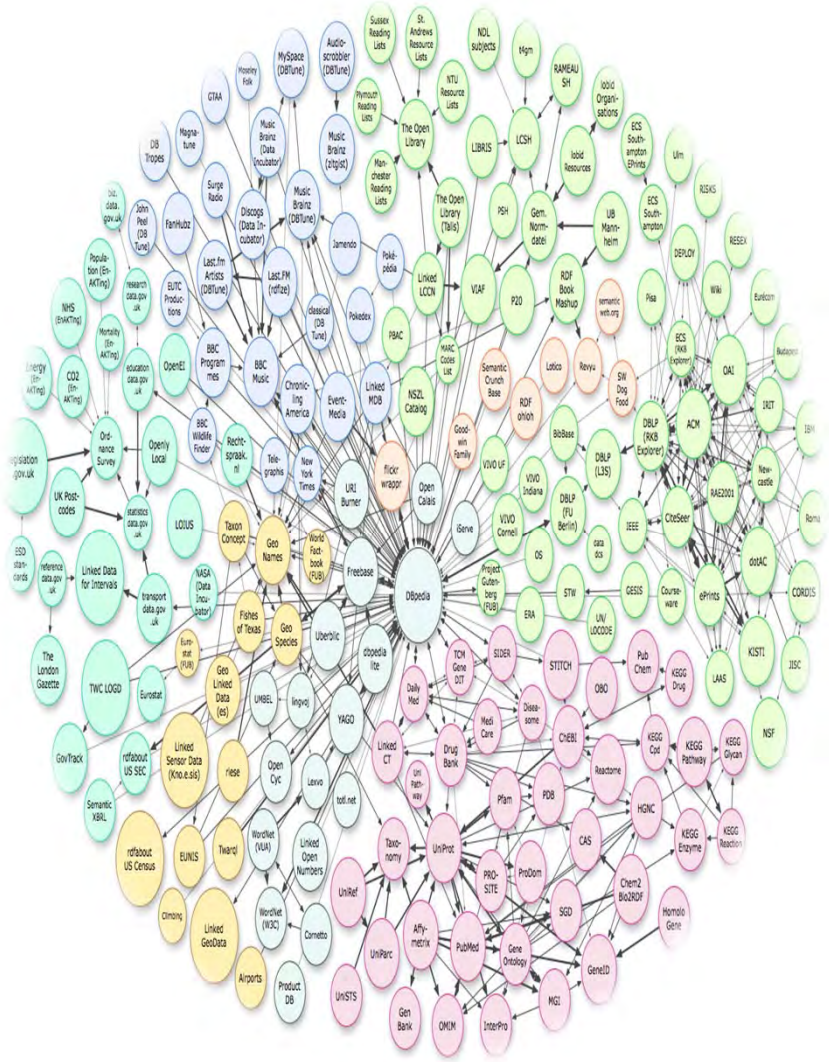
Evaluation on all attributes (>4000) of all infobox templates (>800)



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

The problem – a format mess

Commitment position key: SI2.514875.1

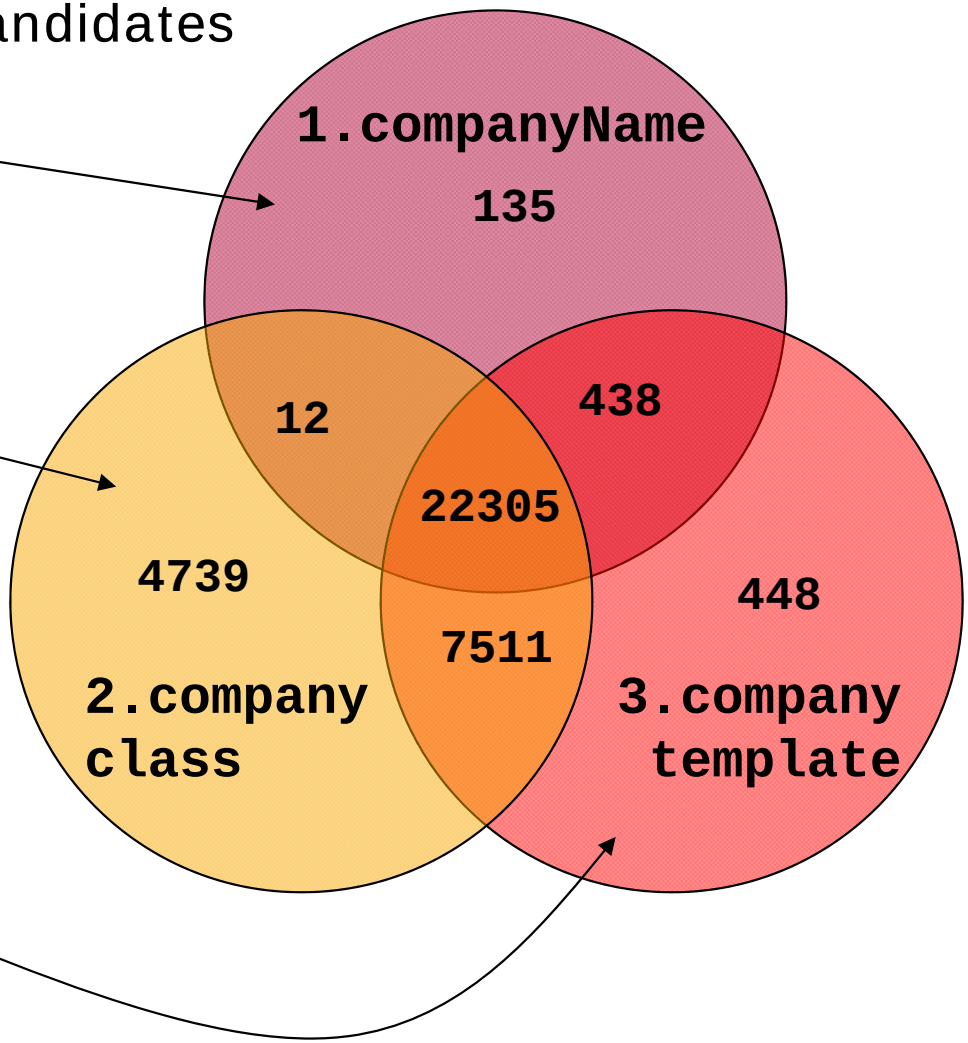
Year:	2008	Amount €:	**99.965.021,40**
Subject of grant or contract:	2007-EU-50010-P EasyWay " - K(2008) 8479		
Responsible Department:	Trans-European Transport Network Executive Agency	Budget line name and number:	Financial support for projects of common interest in the trans-European transport network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

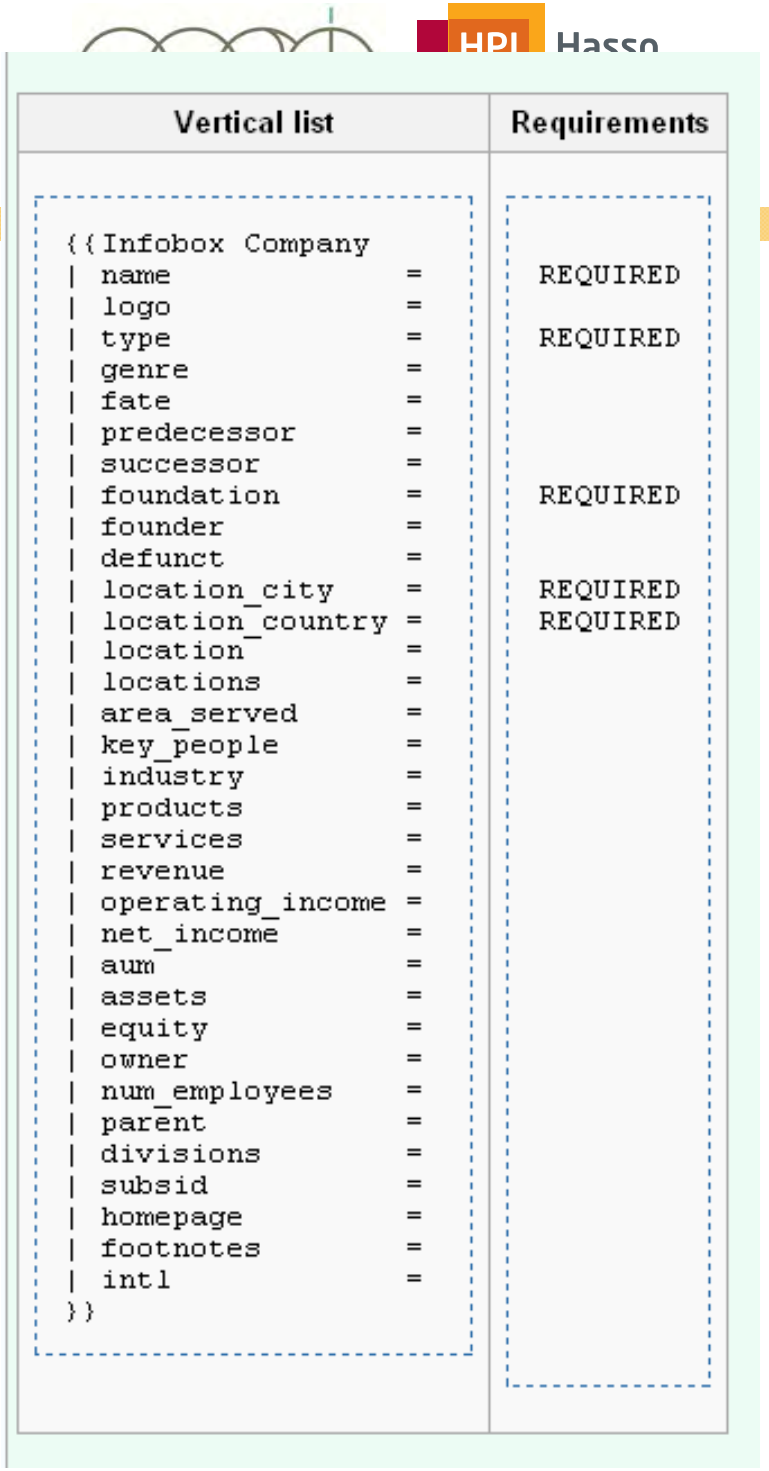
The problem – a domain mess

- 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 =	
}}	

http://en.wikipedia.org/wiki/Template:Infobox_company offers some explanation

The problem – a schema mess

- Triples and ill-defined templates invite disaster.
- Authors can invent new attributes
- Schema chaos: Many attribute synonyms
  - Hundreds of different attributes
- Schema misuse: Many attribute homonyms

- **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

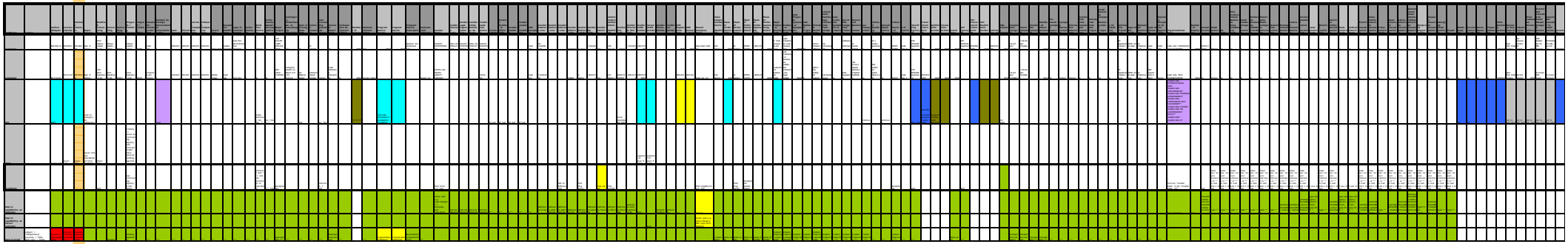
- 34567 companies with 455821 triples
- 1729 different attributes
  - 894 appear only once
- After cleansing by DBpedia
  - 34711 companies with 368185 triples
  - Only 50 different attributes

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

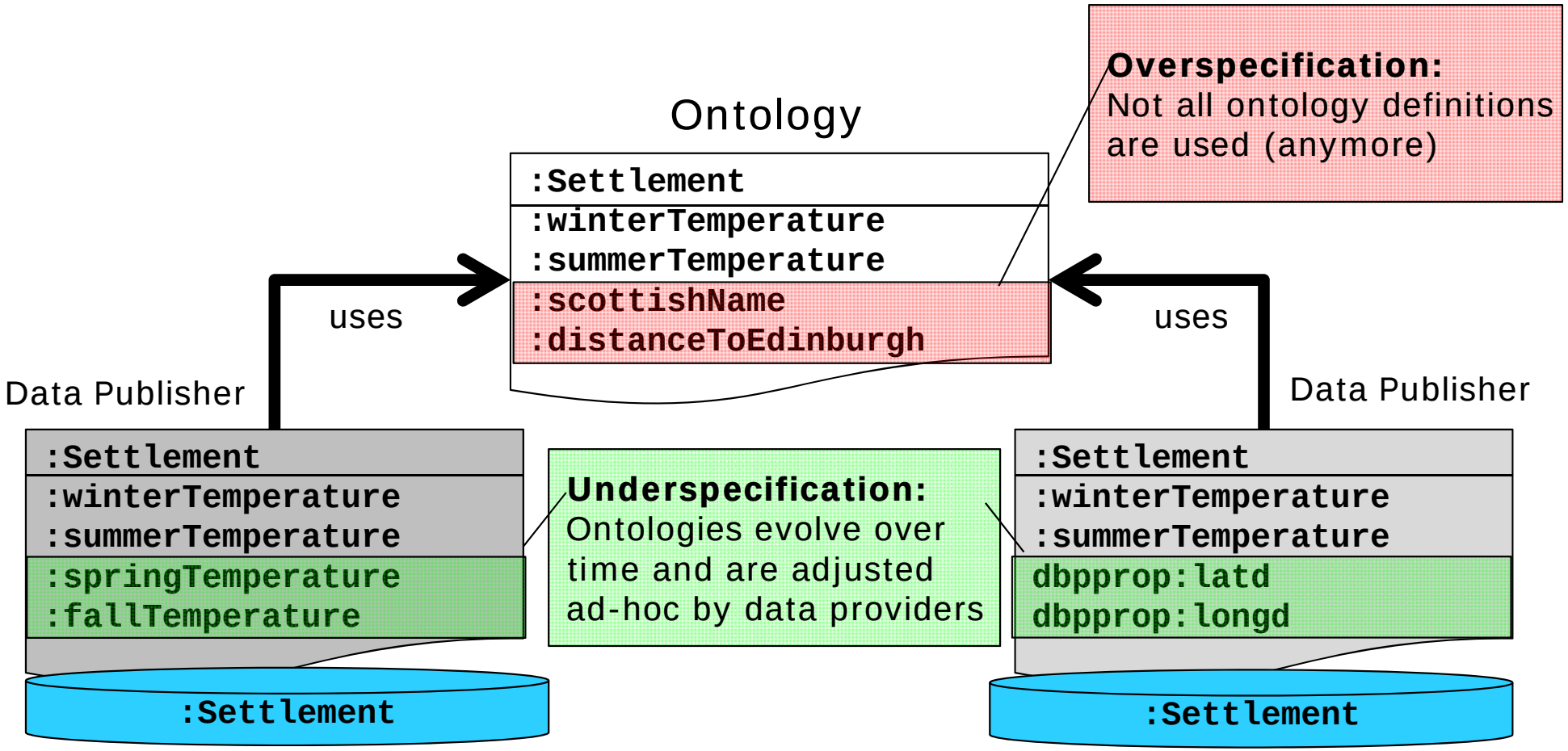
fieldName	<info>	Dollars Obligated	Current Contract Value	Ultimate Contract Value	Major Agency	Modified Contracting Agency	Contracting Agency	Contracting Office	Program / Funding Agency	Program / Funding Office	Reason For Purchas For DoD
example1		$220,989,132	$220,989,132	$220,989,132	Dept. of Defense	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			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							

Schema mapping and data transformation

1pt font!



Ontologies to the rescue?



Mismatch examples

[CIKM2012a

More examples for underspecification in DBpedia 3.7

Property	Class	Support
:numberOfEpisodes	:TelevisionShow	64.362%
:numberOfSeasons	:TelevisionShow	46.267%
:genre	:Artist	56.624%
:philosophicalSchool	:Philosopher	76.225%
:countySeat	:AdministrativeRegion	9.470%
:anthem	:Country	18.730%
:depth	:Lake	33.698%
:numberOfGraduateStudents	:College	93.590%

More examples for overspecification in DBpedia 3.7

Property	Class	Support
:scottishName	:Settlement	0.000%
:distanceToEdinburgh	:Settlement	0.021%
:waistSize	:Person	0.013%
:philosophicalSchool	:Person	0.202%
:countySeat	:PopulatedPlace	0.831%
:anthem	:PopulatedPlace	0.147%
:depth	:Place	0.723%
:numberOfGraduateStudents	:EducationalInstitution	0.300%

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:

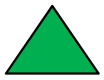
Wal-Mart

Undisclosed

?

Assets exceed £4 billion GBP

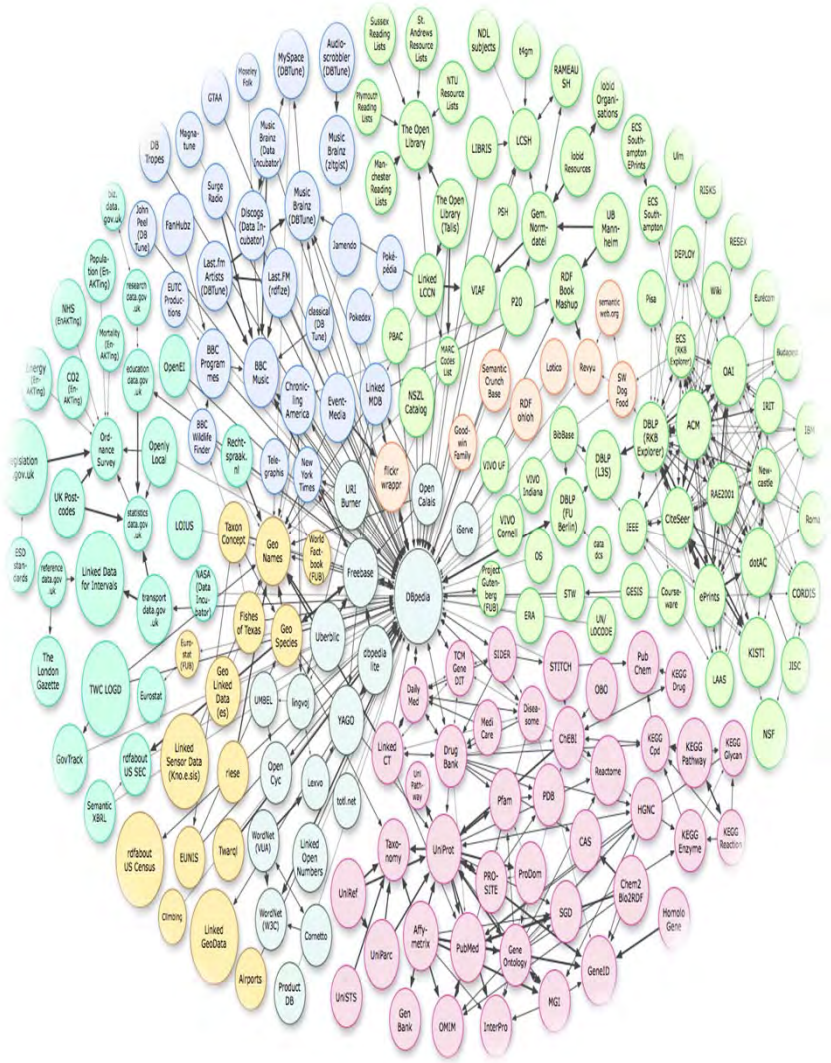
http://www.credit-suisse.com/investors/en/reports/2007_results_q4.jsp

Image:green_up.png → 

€ bn (as of 2004)

Overview

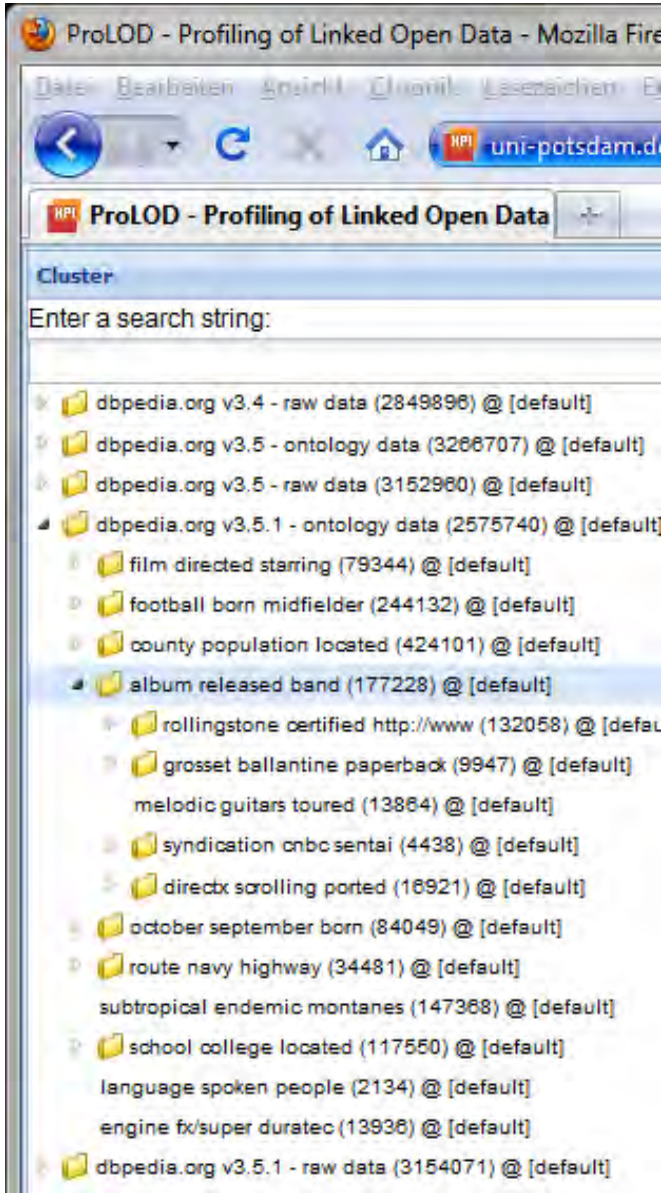
- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - …of mops and brooms
  - Cross-language integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience



ProLOD profiling tasks

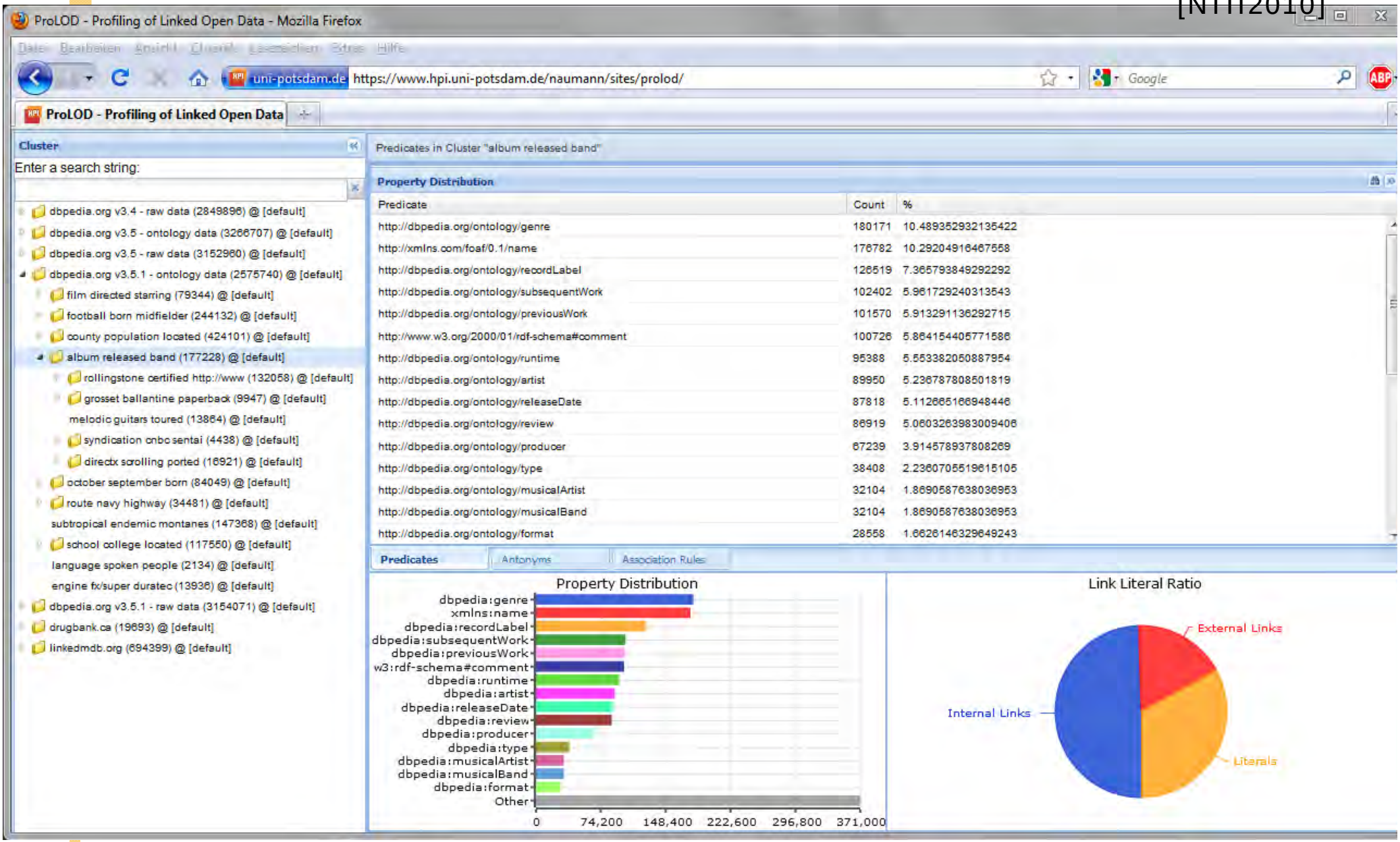
[NTII2010]

- 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 Linked Open Data

[NTII2010]



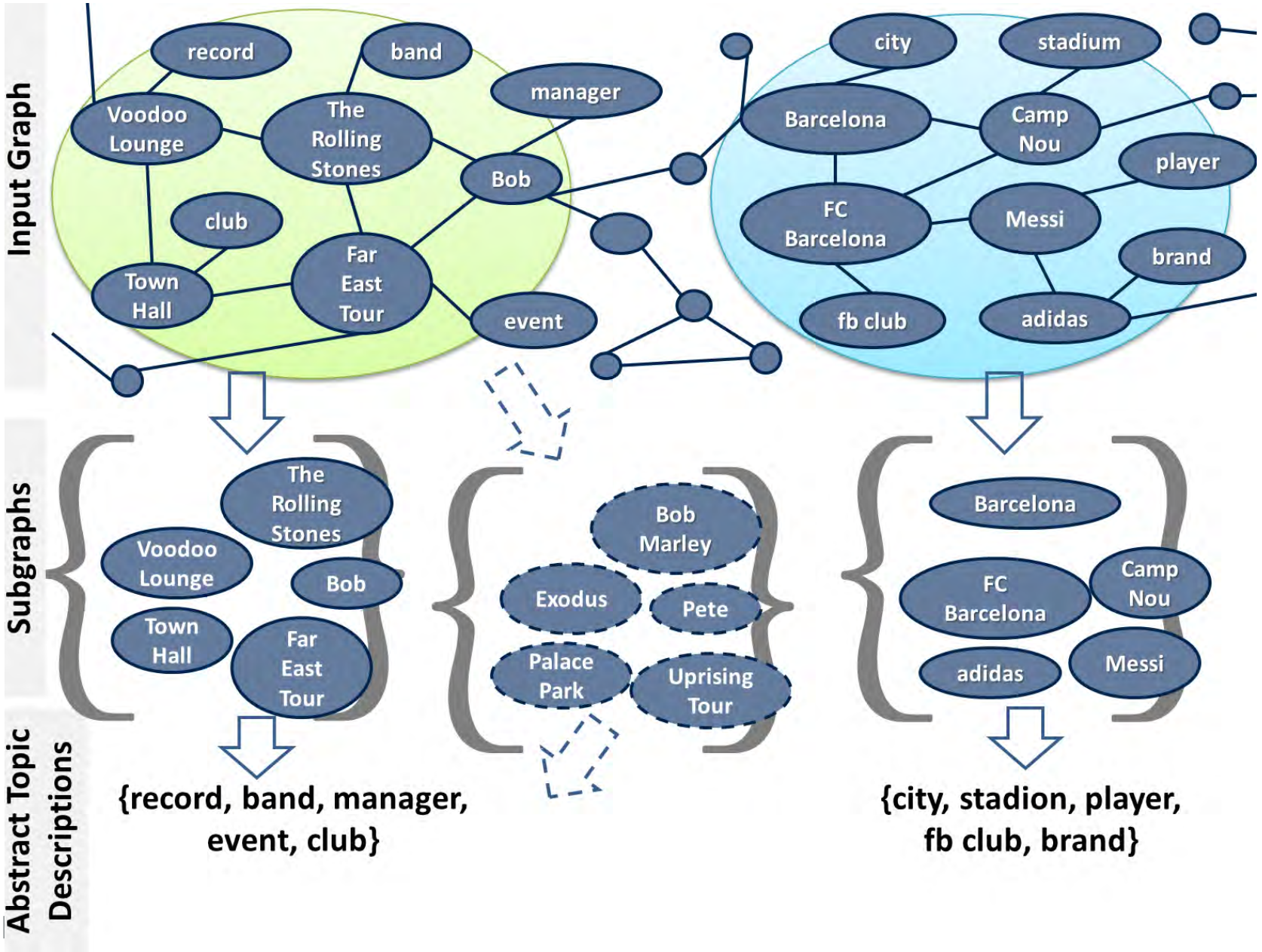
Beyond traditional profiling: Topic discovery on Webdata

[CIKM2012b]

- Large amounts of heterogeneous graph data emerge from different domains.
- How to select appropriate data for the task at hand?
- There is no means to describe dataset topics in a structured and abstract way!

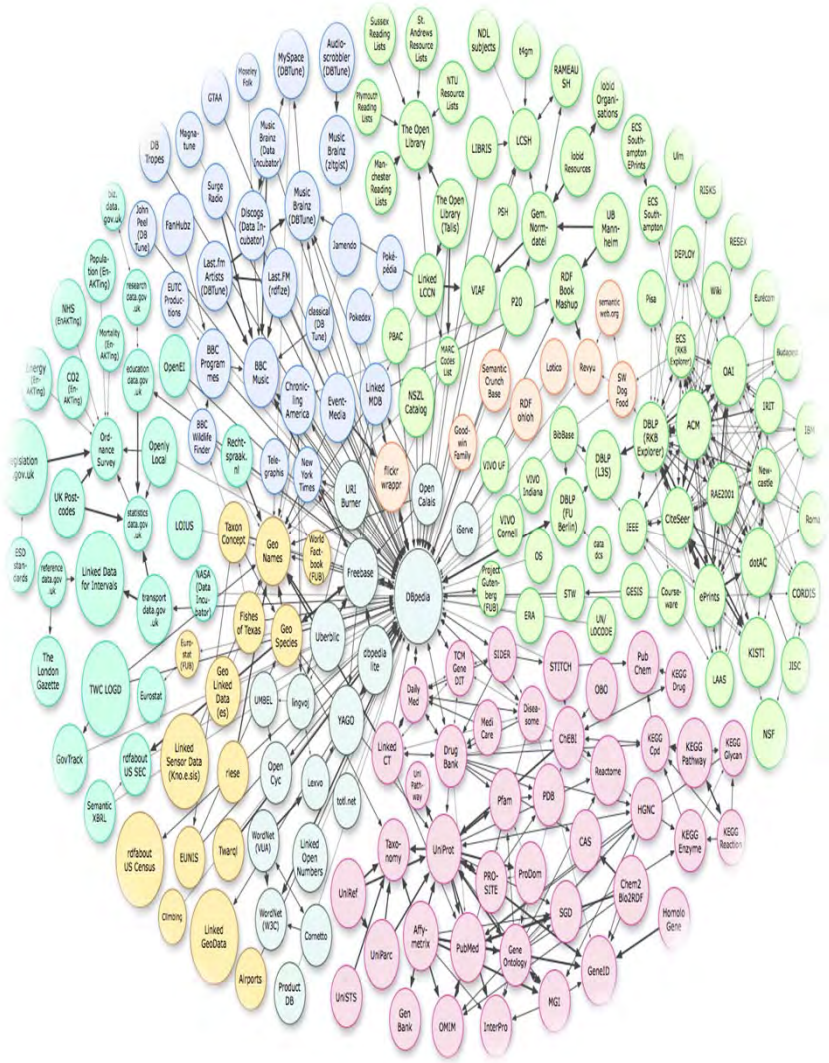
- Perform an overlapping graph partitioning that relies on structure and coherence of classes in the graph only.
  - Our partitioning does not rely on textual labels
  - Each partition covers a topic.

Topics from graph-patterns



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

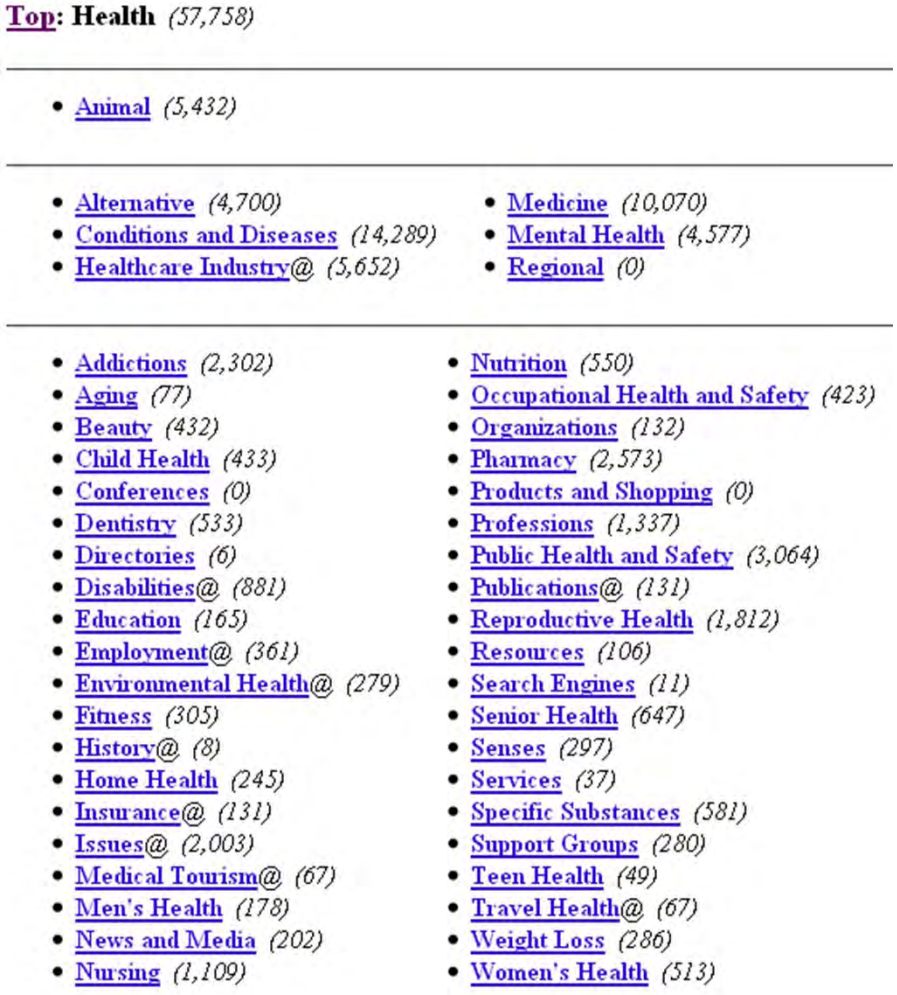
Data Extraction & scrubbing

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

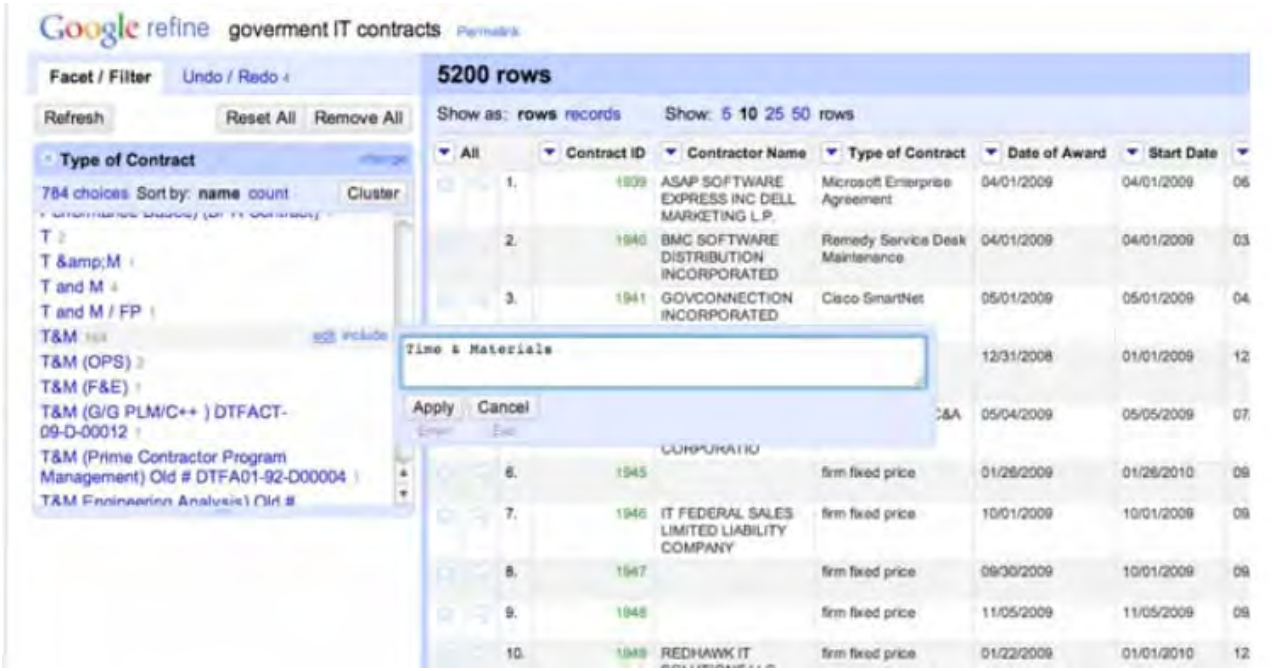
Step 2: Schema matching and mapping

- 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/employee /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	subsidiary, subsidiaries		/business/company/subsidiary

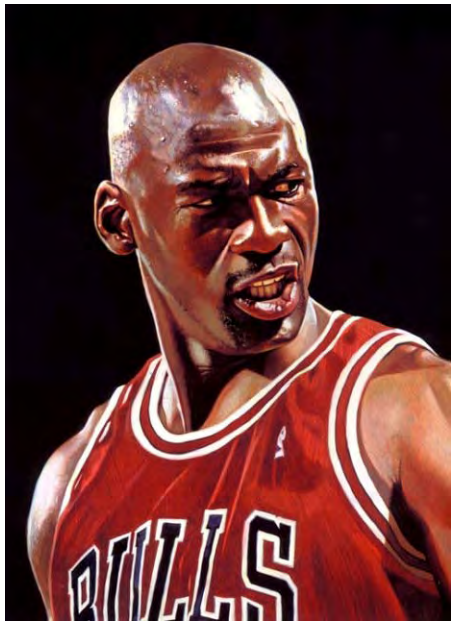
Step 3: Data extraction & scrubbing

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



Step 4: Entity matching

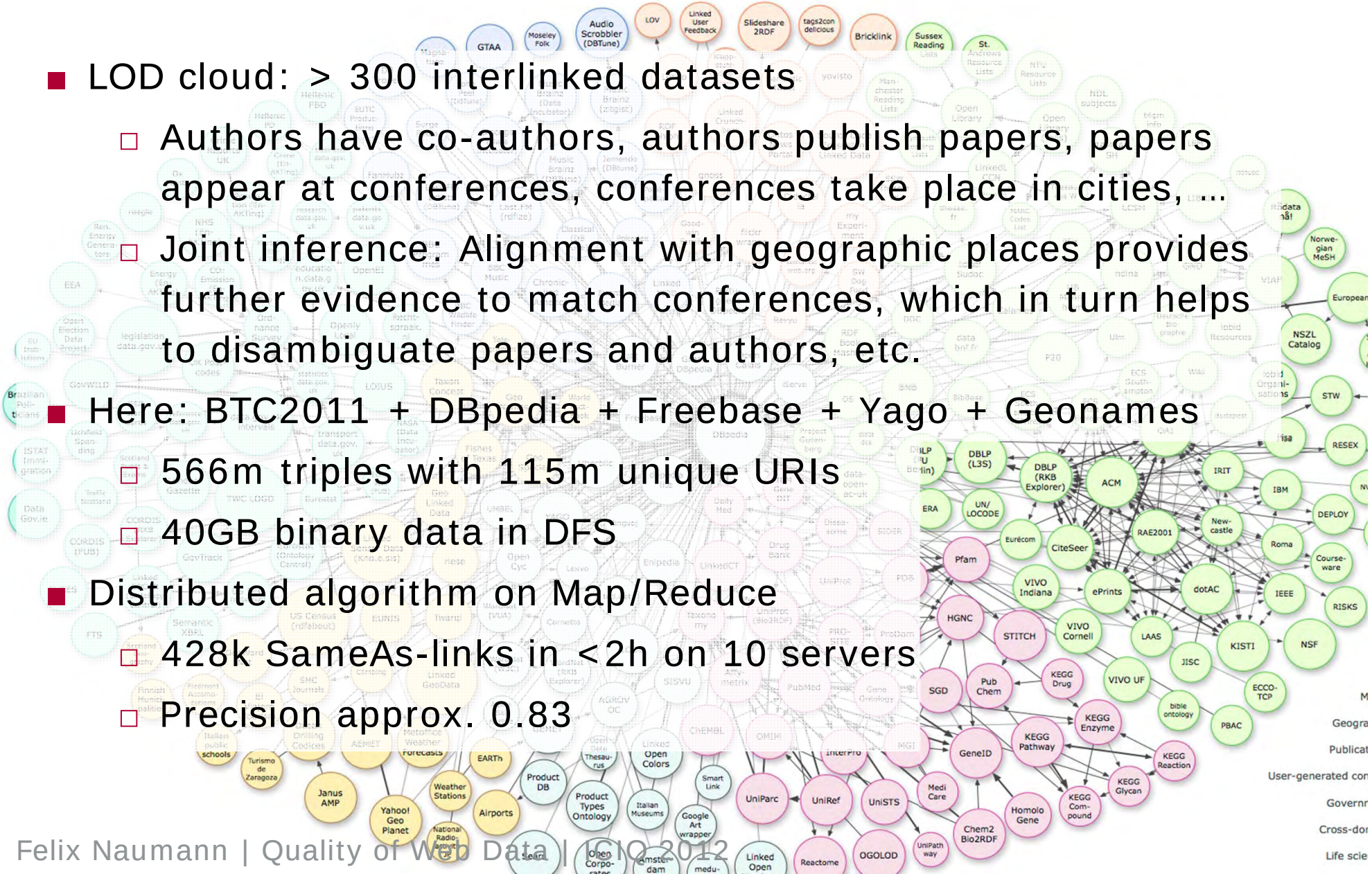
- Duplicate entities
- Linking between entities
- Challenges
  - Fuzzy matching: Similarity measures
  - Data volume: Partitioning algorithms
  - Sparse data
    - **`Michael Jordan born_in Miami`**



Web-scale entity matching

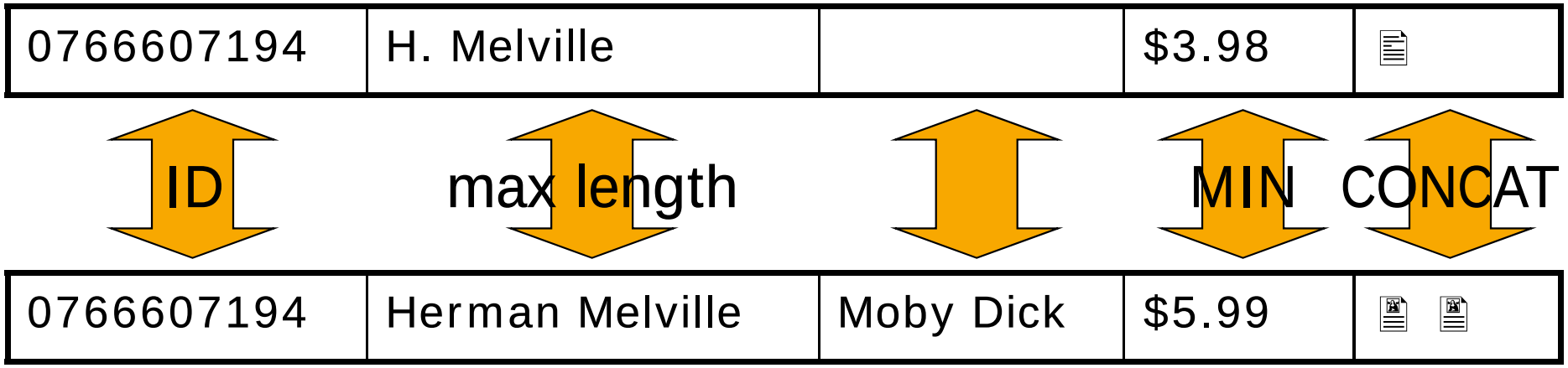
[CIKM2012c]

- LOD cloud: > 300 interlinked datasets
  - Authors have co-authors, authors publish papers, papers appear at conferences, conferences take place in cities, …
  - Joint inference: Alignment with geographic places provides further evidence to match conferences, which in turn helps to disambiguate papers and authors, etc.
- Here: BTC2011 + DBpedia + Freebase + Yago + Geonames
  - 566m triples with 115m unique URIs
  - 40GB binary data in DFS
- Distributed algorithm on Map/Reduce
  - 428k SameAs-links in <2h on 10 servers
  - Precision approx. 0.83



Step 5: Data fusion

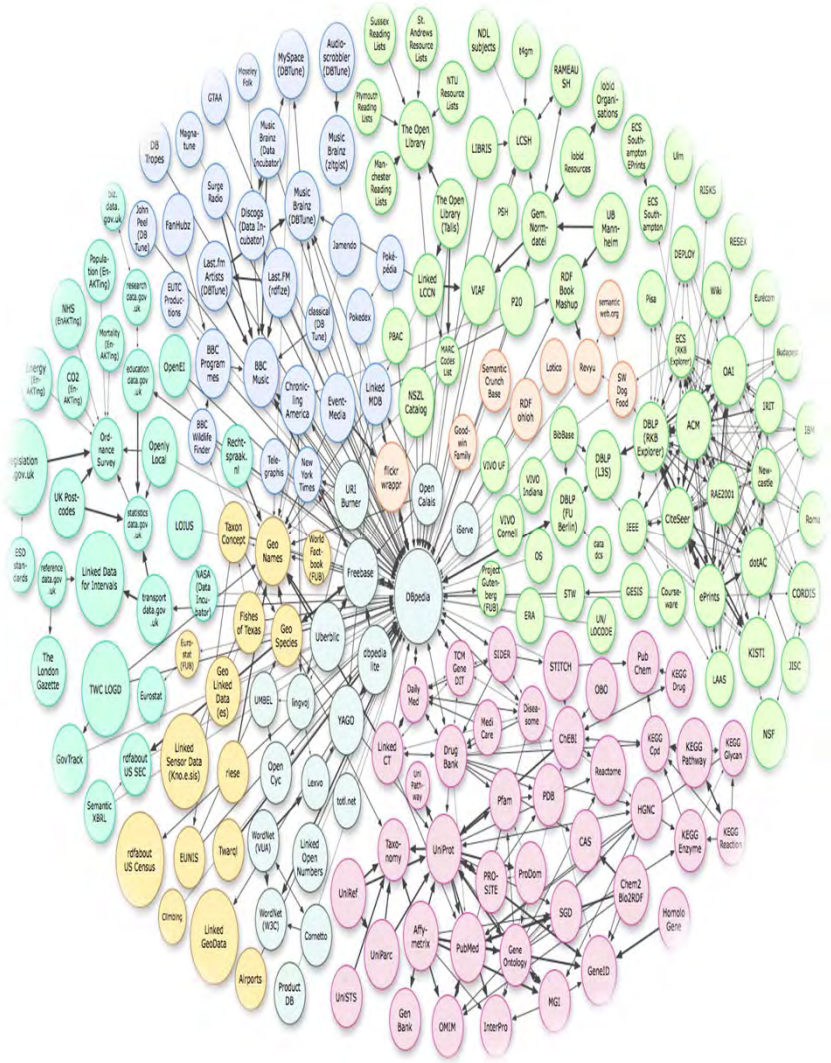
- Combine multiple representations of real-world entities
  - Survivorship, consolidation, etc.
- Resolve data conflicts
  - Conflict resolution functions
  - Reputation / accuracy / freshness -> “truth discovery”



- Retain data lineage

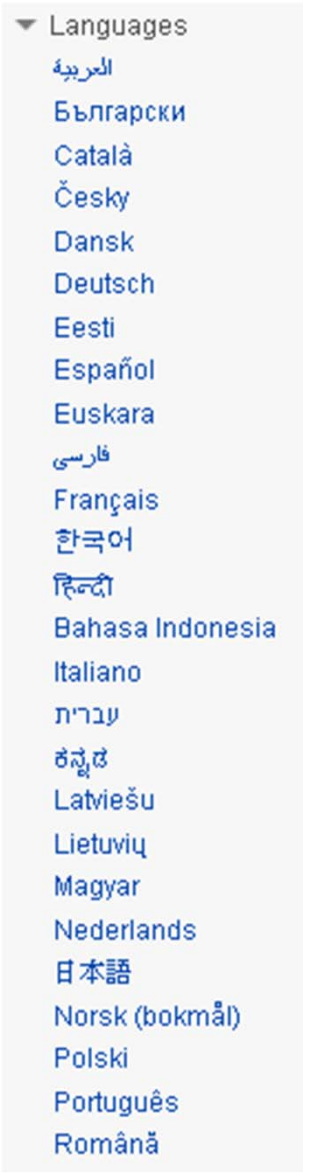
Overview

- 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



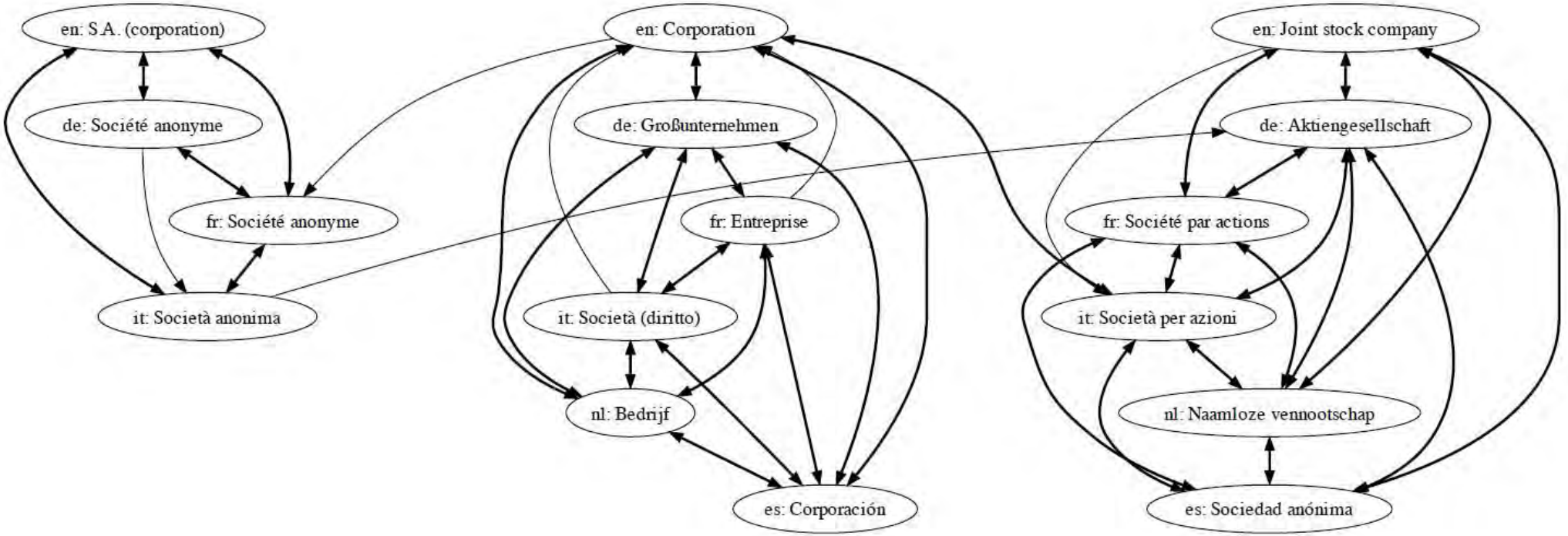
Multi-Lingual Wikipedia

- 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

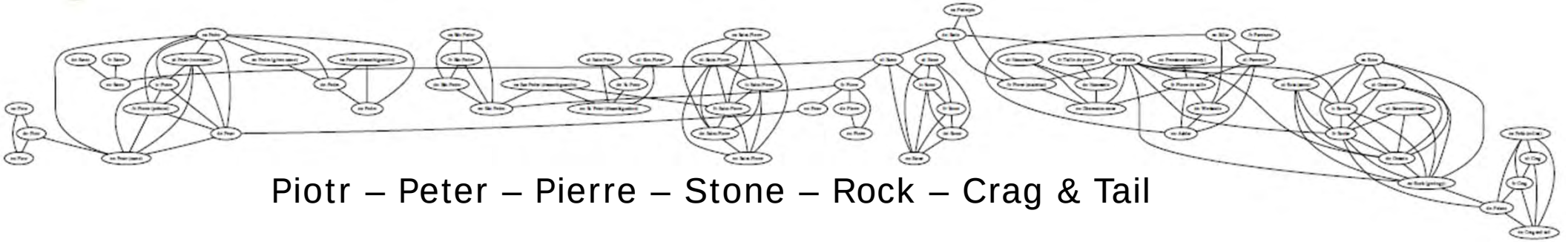


Interlanguage links (ILLs)

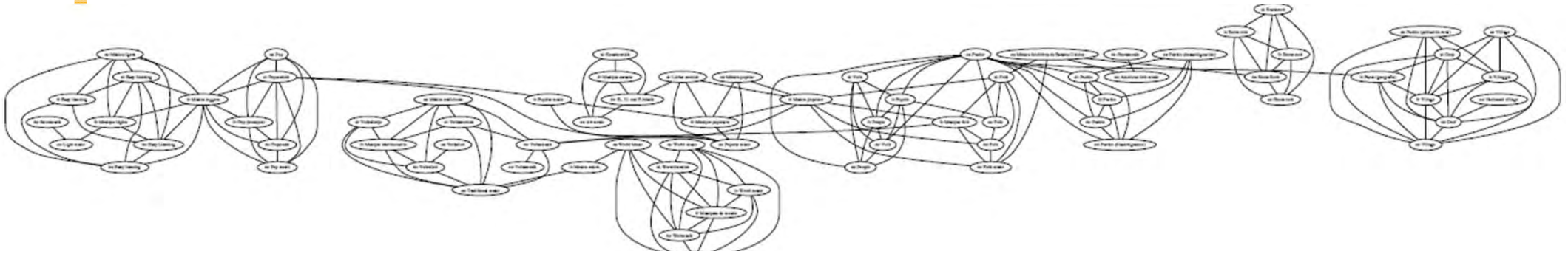
- First, evaluate quality of ILLs and build duplicate clusters
  - Build connected components using cross-language links (on the six largest languages)
- But, largest weakly connected component has 108 articles
  - 26 English, 26 German, 21 French, 13 Italian, 13 Dutch, and 9 Spanish articles



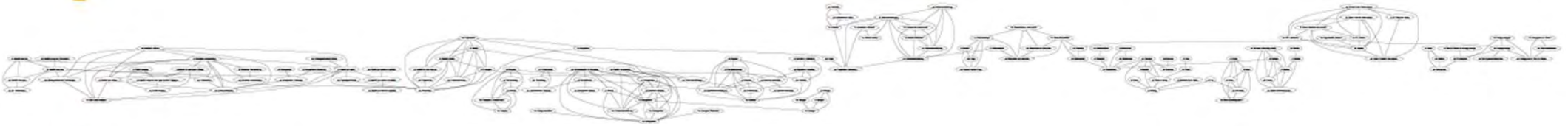
Other large components



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

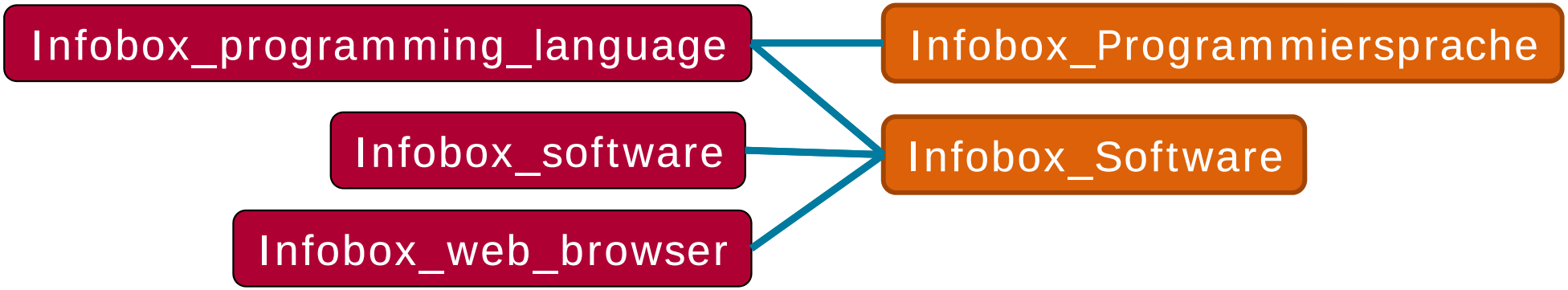
BCC
- Bidirectionally connected components (BCC)
- Undirected graph of bidirectional components is connected
- 4,241 BCCs of which 2,980 are incoherent

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 when 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 km² (344.3 sq mi)
Elevation	34 - 115 m (-343 ft)
Population (31 March 2010)[1]	
- City	3,440,441
- Density	3,857.6/km² (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

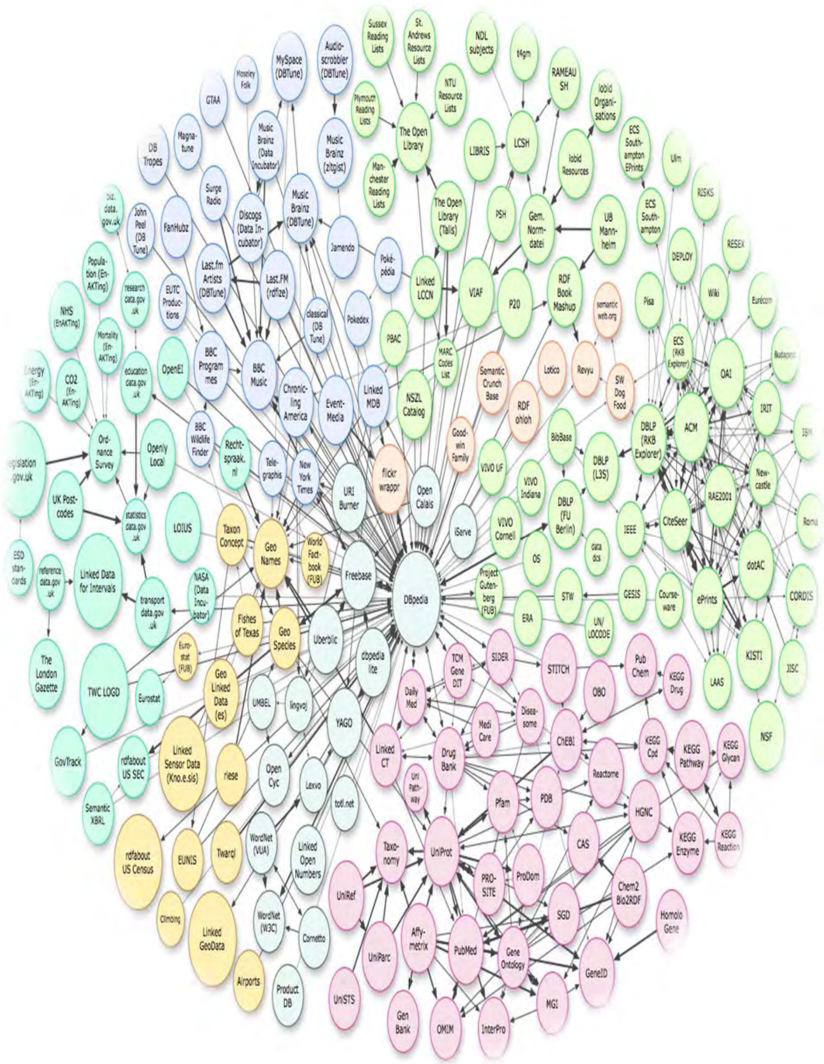
Next step by community: Wikidata

- 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

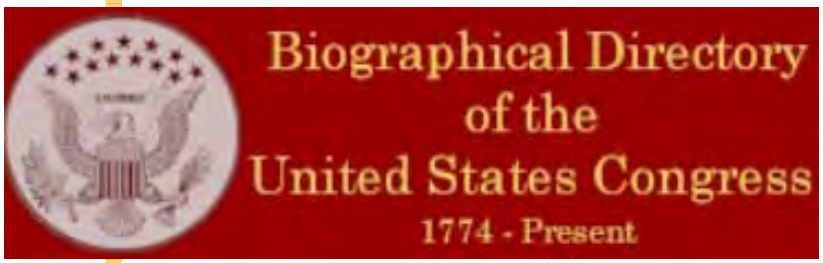
- Begin: April 2012 – much to do
- http://meta.wikimedia.org/wiki/Wikidata/de

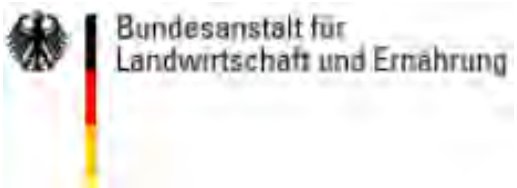
Overview

- 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

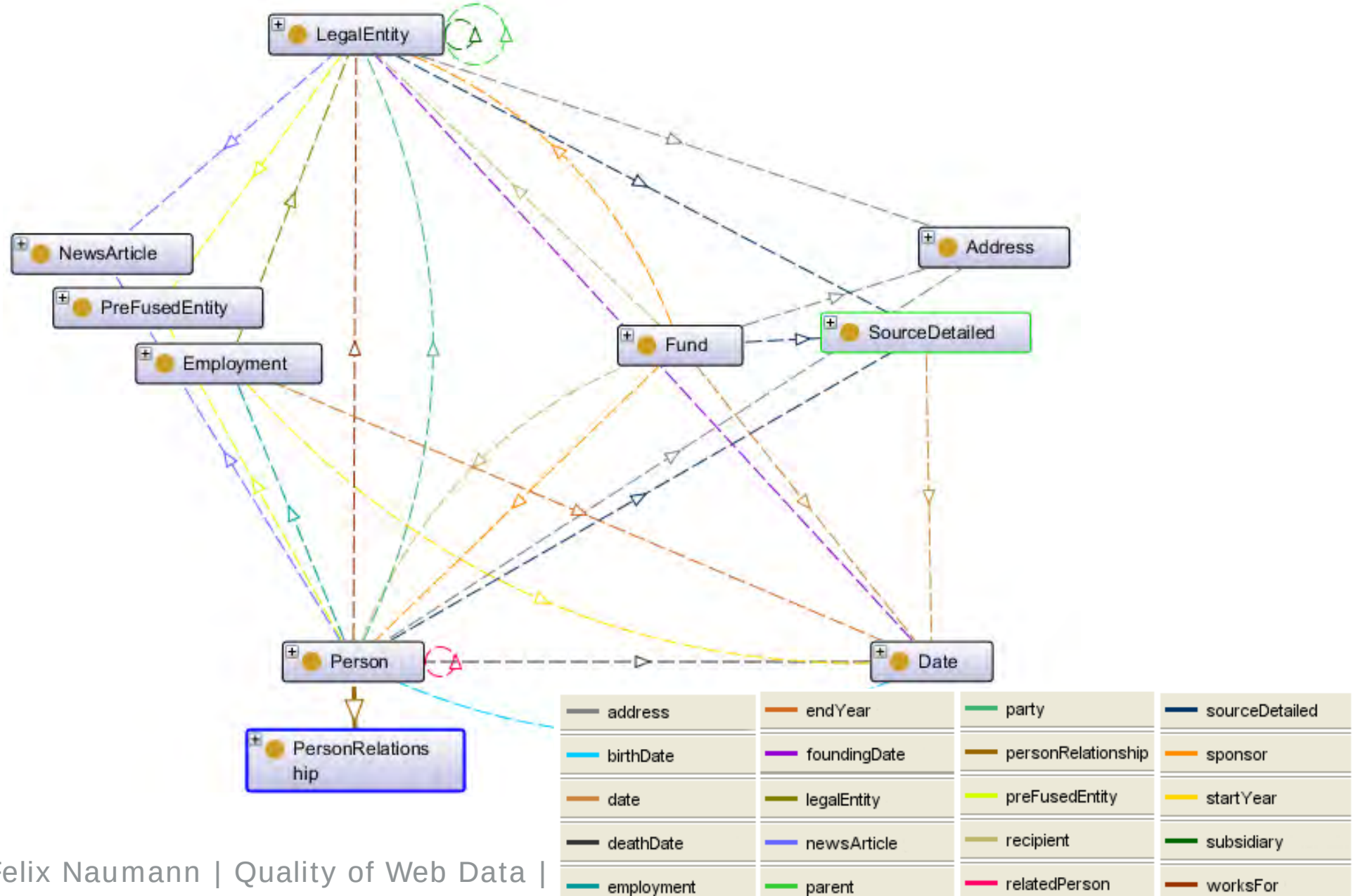


Motivation – Wealth of Open Gov Data



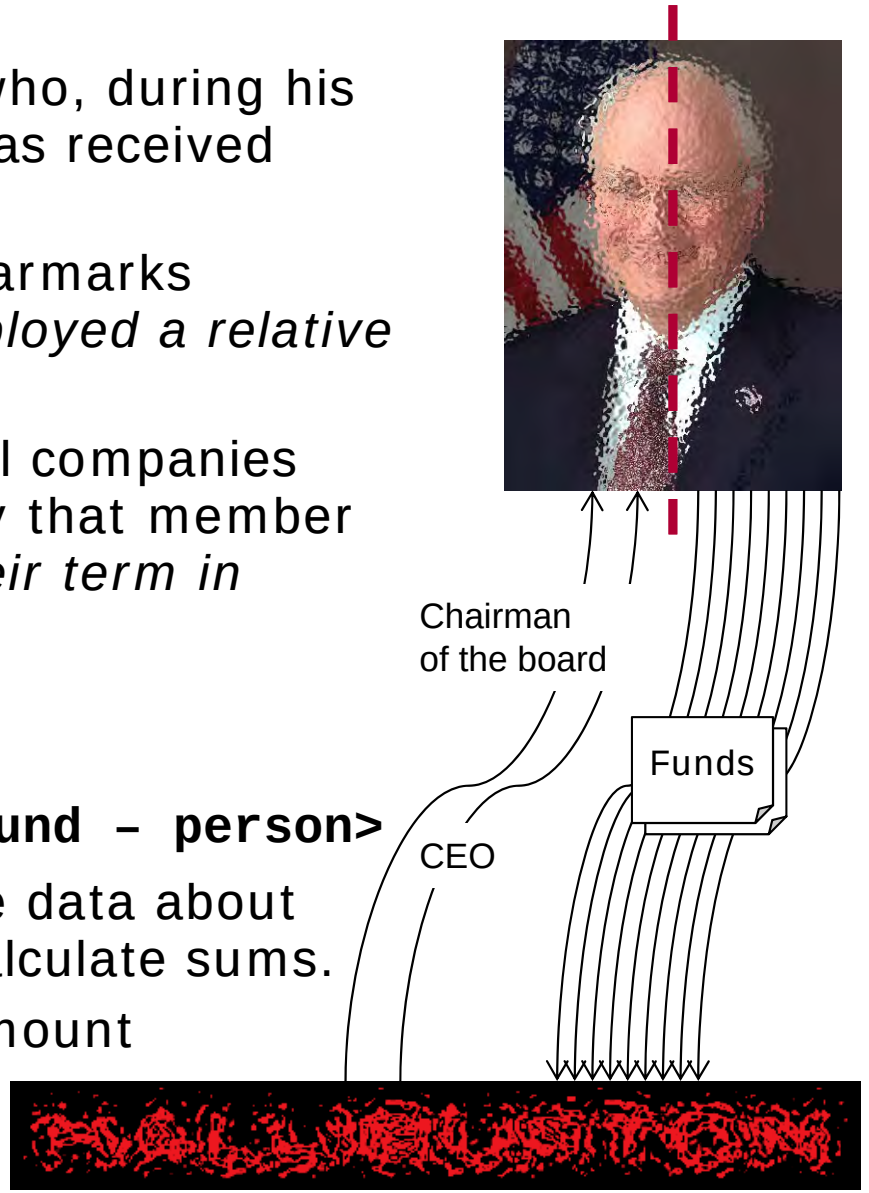


Companies, Companies, and People



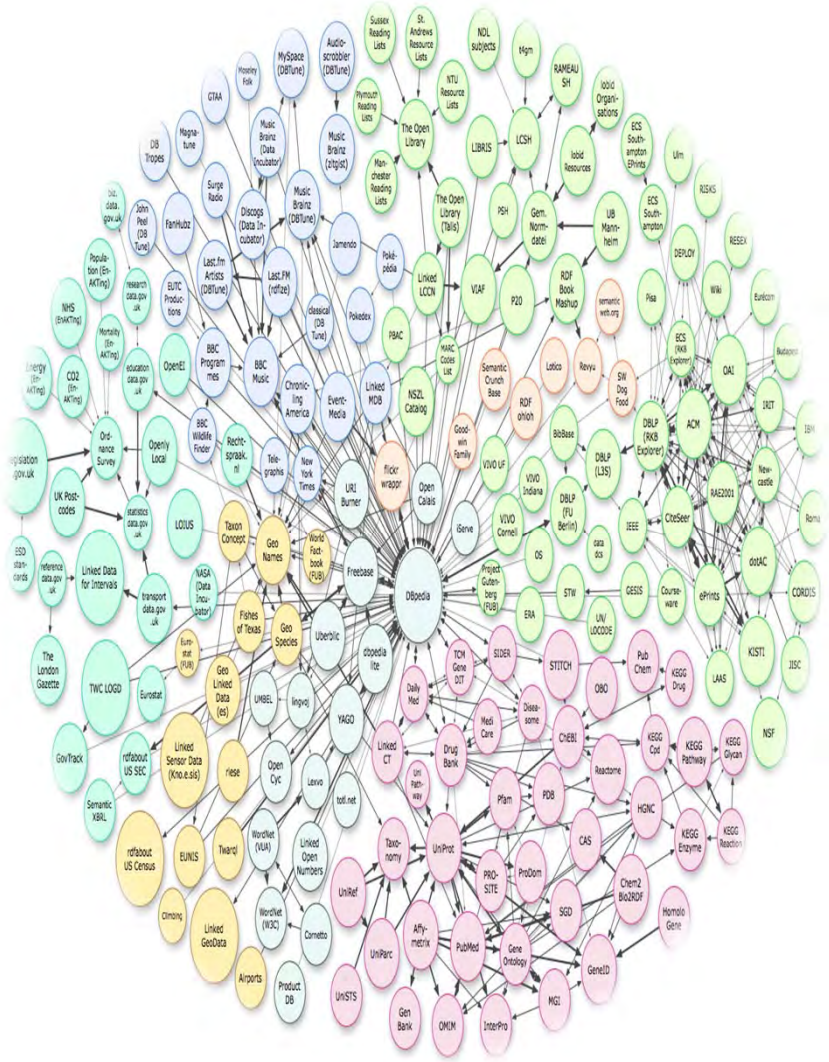
Interesting queries

- 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 …”



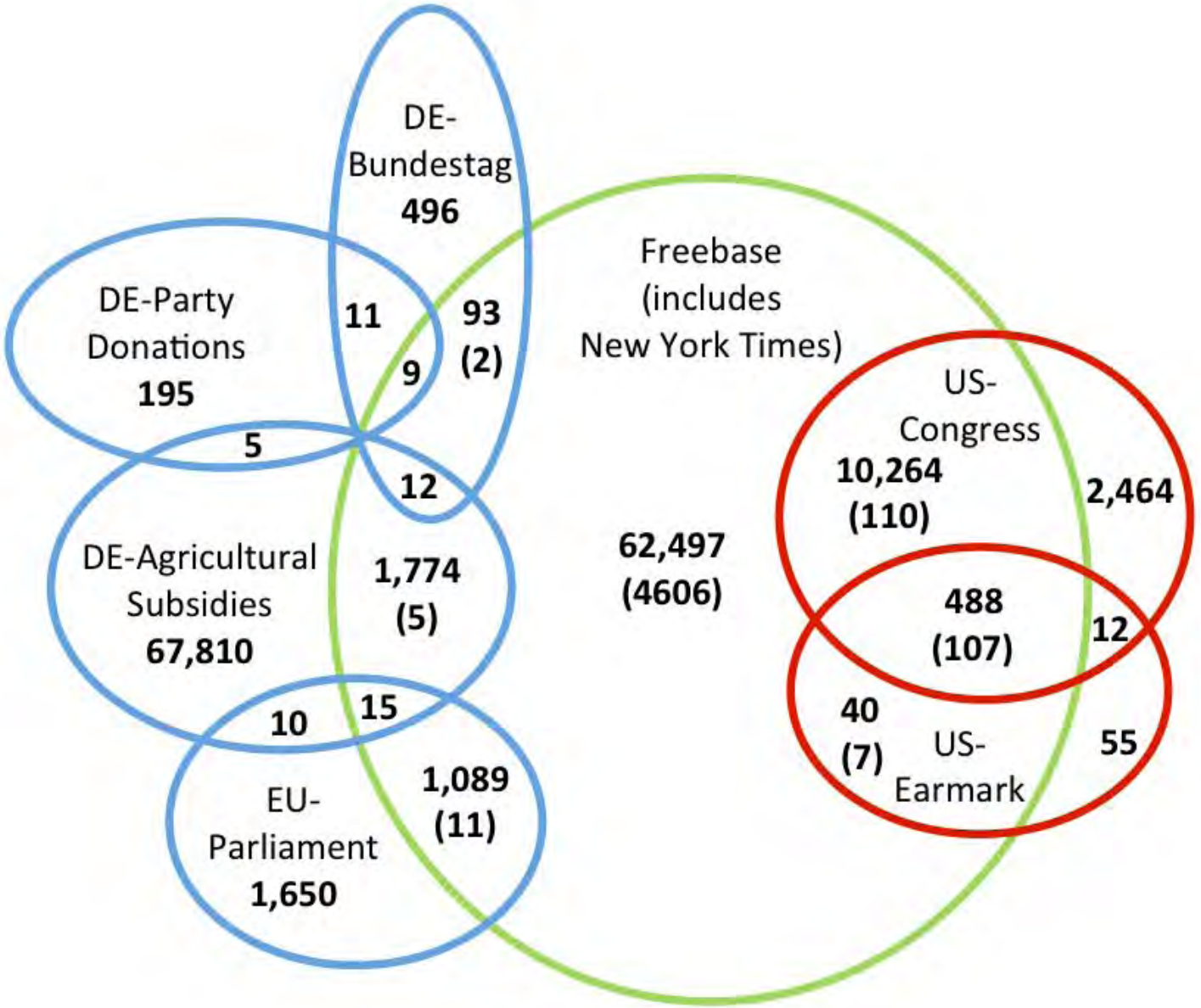
Overview

- 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

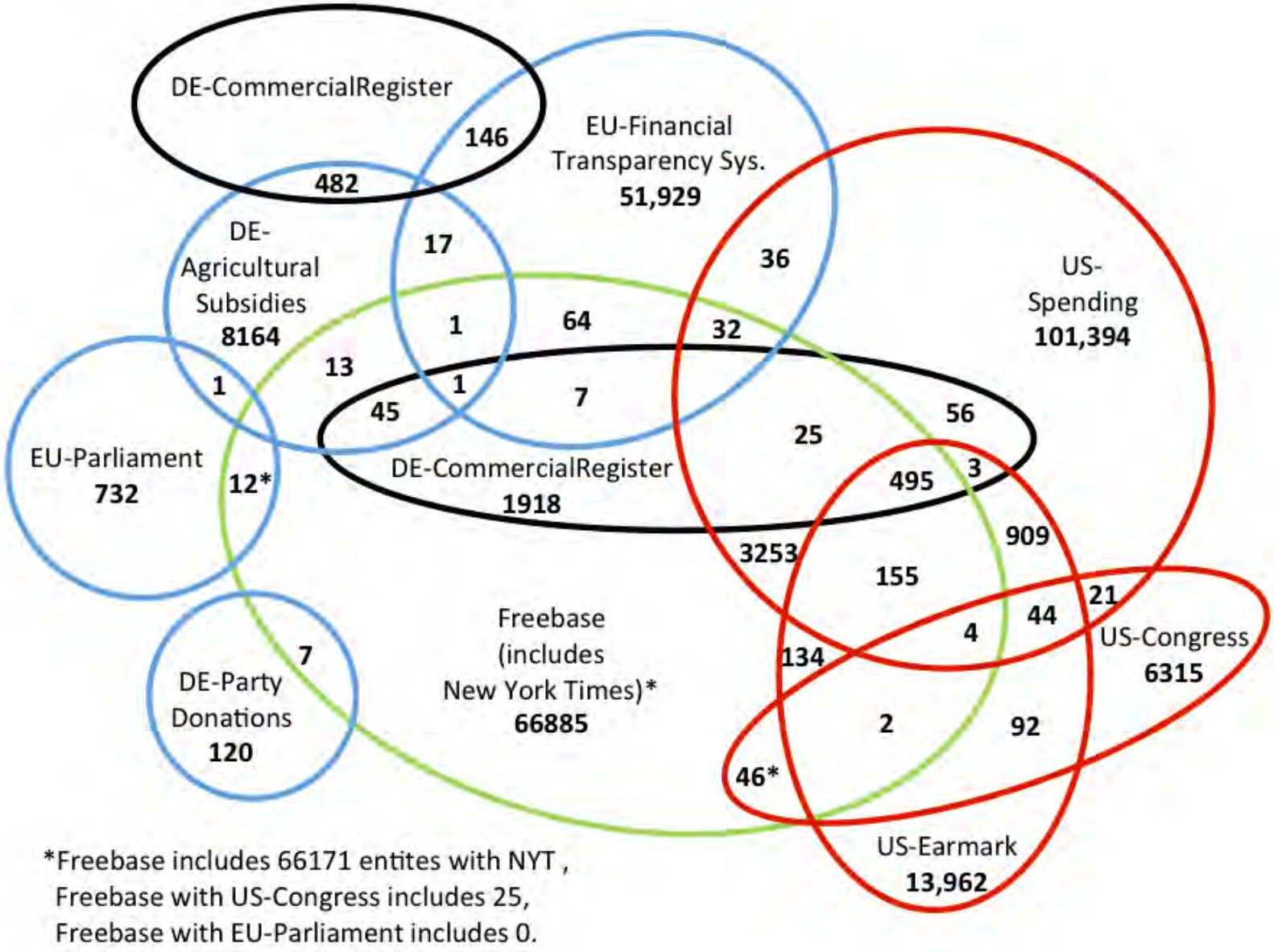


Step	Time		Input size on master node and element count			Details
	Jaql 0.4	Jaql 0.5	LegalEntity.json	Person.json	Fund.jon	
Scrubbing						
Scrub and map 15 files	1h 15min	1h 15min	Start with 11 GB size			- map and normalize attributes - set references within a source (includes many joins) - group entities / 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	-||-			- compute transitive closure of IDs (transform and combine with UDF)
			Join 5 402 fused IDs with 7530 Leg.E.			- 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	-||-	-||-	-||-	- 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
5
12
DE-Agricultural
Subsidies
67,810
1,774
(5)
10
15
EU-
Parliament
1,650
1,089
(11)
Freebase
(includes
New York Times)
62,497
(4606)
US-
Congress
10,264
(110)
2,464
488
(107)
12
40
(7)
US-
Earmark
55

Organisations



http://govwild.org

[ISEM2010,WWW2012]

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

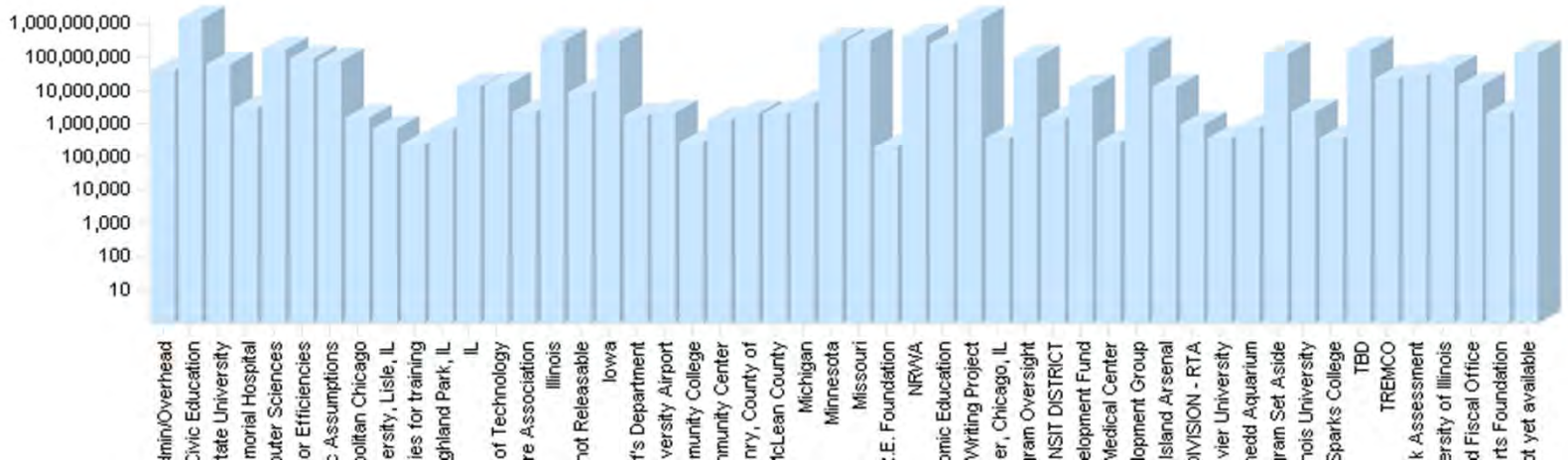
- Keyword Queries
- Linked Data Interface (dereference URIs)
- Exploration of entities mentioned in New York Times articles
- Data Download (RDF, SQL Dump, JSON files)

Barack Obama

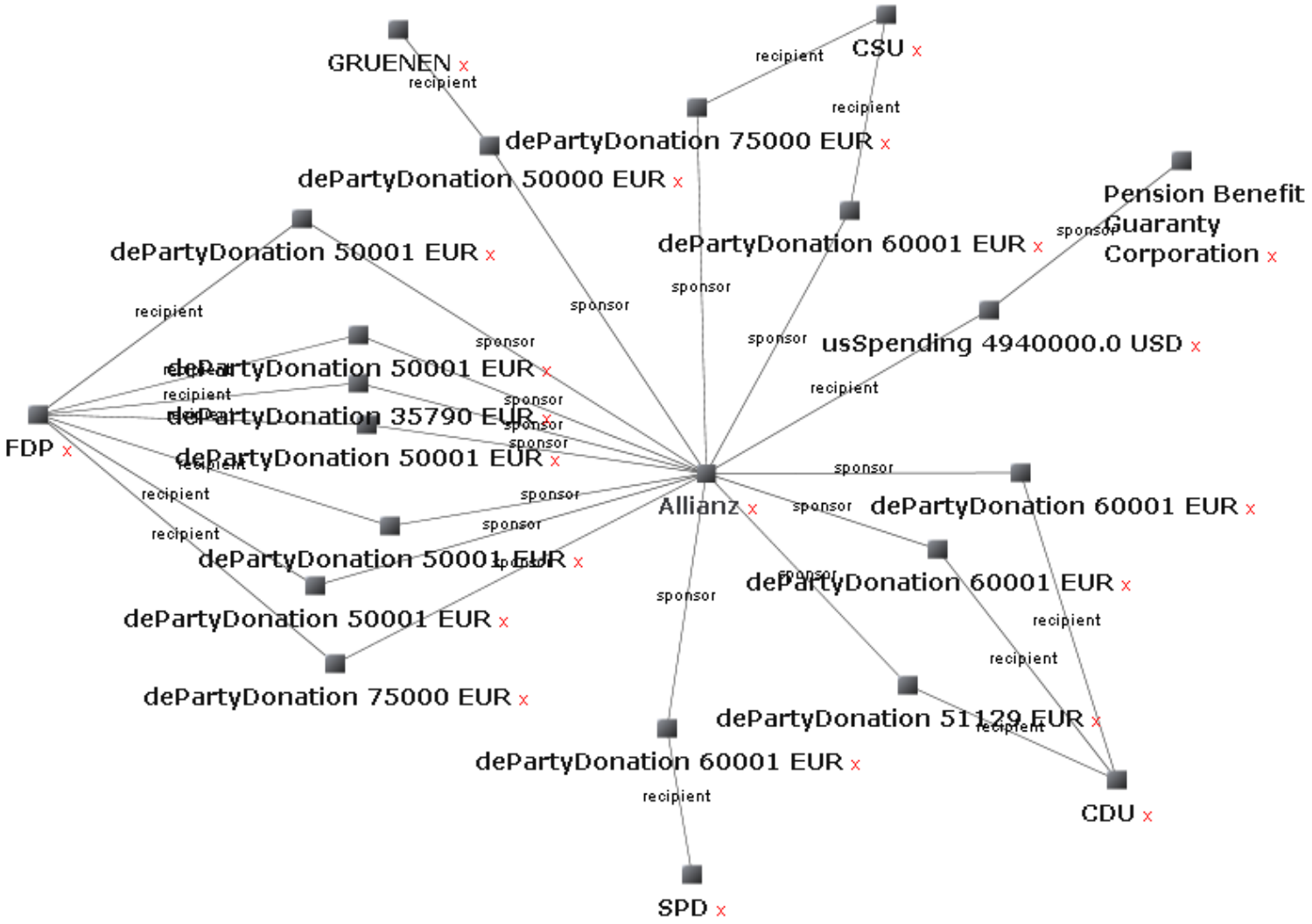
Barack Obama

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

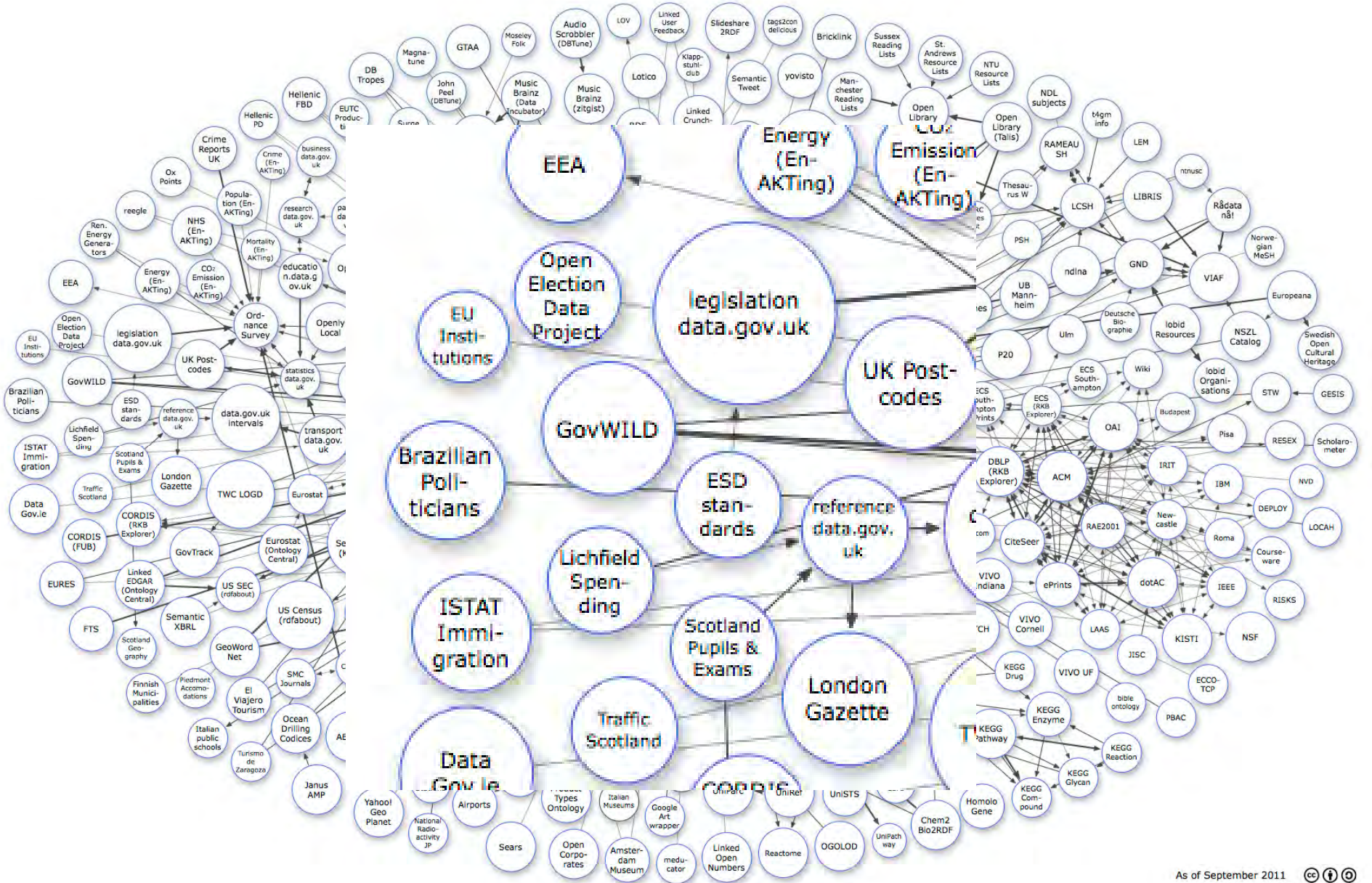
Earmarks



Allianz insurance



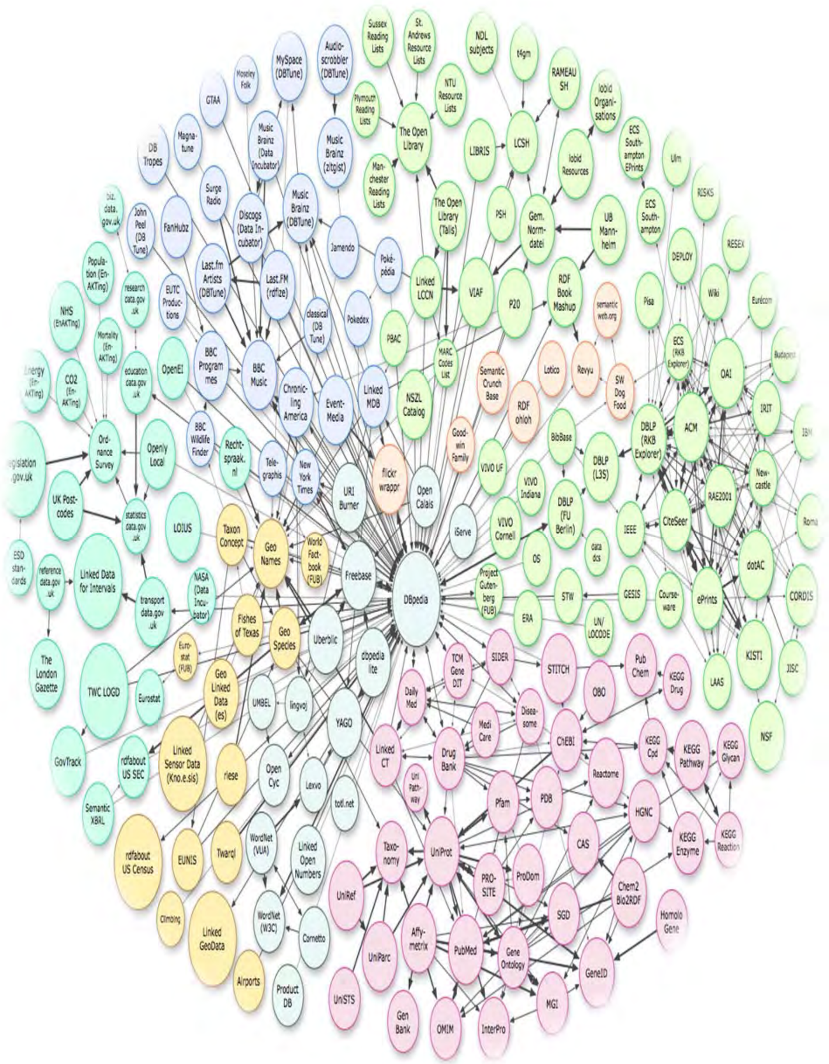
Linked Data Graph



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

- [CIKM2010] *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
- [NTII2010] *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
- [ISEM2010] *Linking Open Government Data: What Journalists Wish They Had Known*
  C. Böhm, F. Naumann, M. Freitag, S. George, N. Höfler, M. Köppelmann, C. Lehmann, A. Mascher, and T. Schmidt.
  Linked Data Triplification Challenge 2010 @ I-Semantics, Graz (honorable mention).
- [IS2012] *Cross-lingual Entity Matching and Infobox Alignment in Wikipedia*
  Daniel Rinser, Dustin Lange, Felix Naumann
  Information Systems (IS) (accepted), 2012
- [CIKM2012a] *Reconciling Ontologies and the Web of Data*
  Ziawasch Abedjan, Johannes Lorey, and Felix Naumann.
  In Proceedings of the International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, 2012.
- [CIKM2012b] *Latent Topics in Graph-Structured Data*
  Christoph Böhm and Gjergji Kasneci and Felix Naumann. In Proceedings of the Conference on Information and Knowledge Management (CIKM), 2012.
- [CIKM2012c] *LINDA: Distributed Web-of-Data-Scale Entity Matching (poster)*
  Christoph Böhm, Gerard de Melo, Felix Naumann, and Gerhard Weikum.
  In Proceedings of the International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, 2012.
- [WWW2012] *GovWILD: Integrating Open Government Data for Transparency (demo)*
  C. Böhm, M. Freitag, A. Heise, C. Lehmann, A. Mascher, F. Naumann, M. Hernandez, V. Ercegovac and P. Haase.
  In Proceedings of the International World Wide Web Conference (WWW), Lyon, France, 2012.