Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Dr. Crowdsource

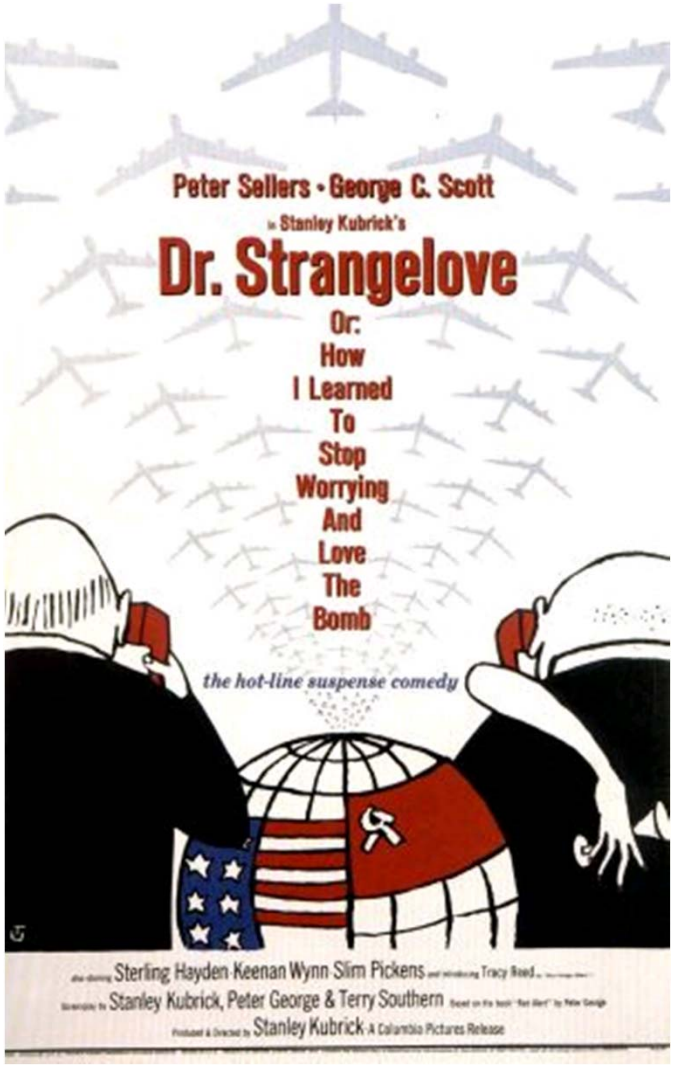
or: How I Learned to Stop Worrying and Love Web Data

Invited Talk @ BEWEB 2011

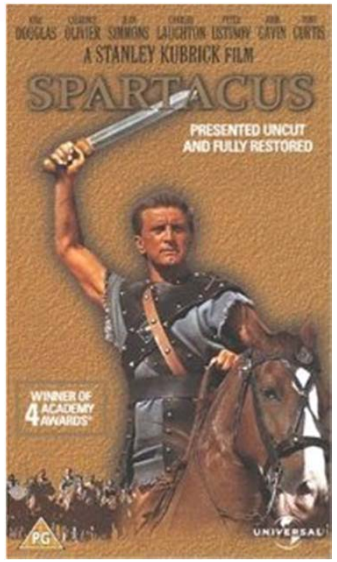
25.3.2011

Felix Naumann

Dr. Strangelove – for the small frys…

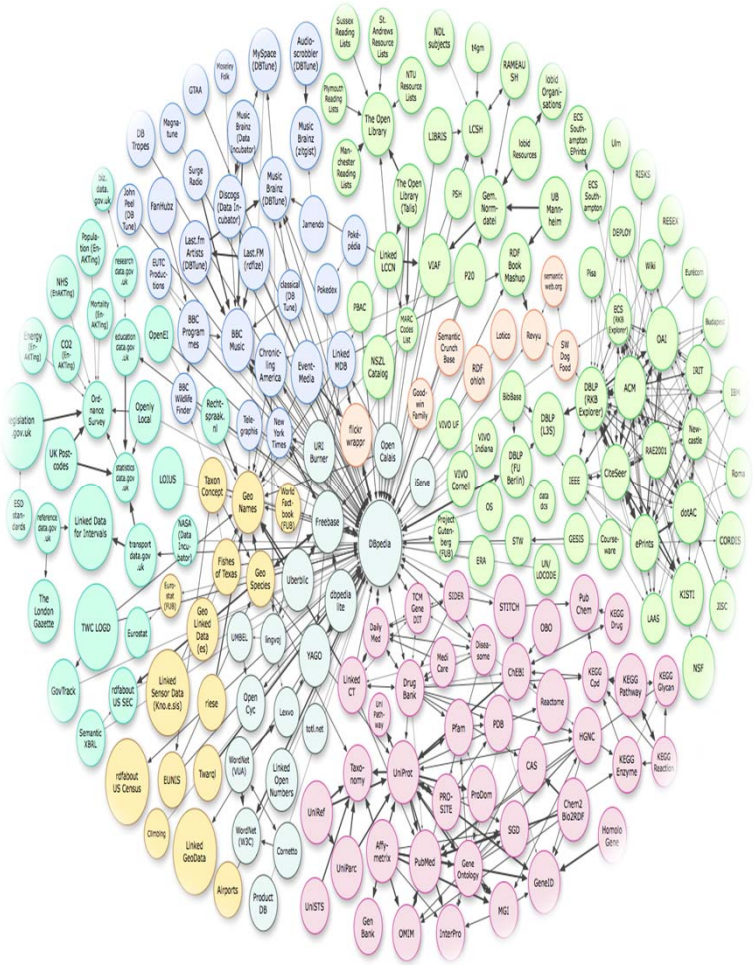


Stanley Kubrick's
FULL METAL JACKET

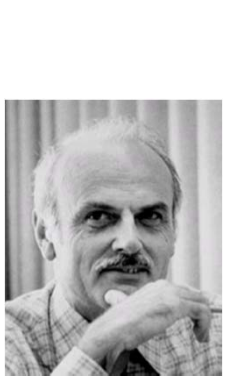


Overview

- 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



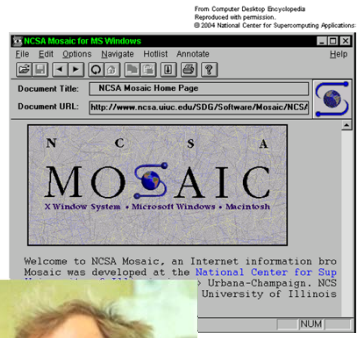
DBMS

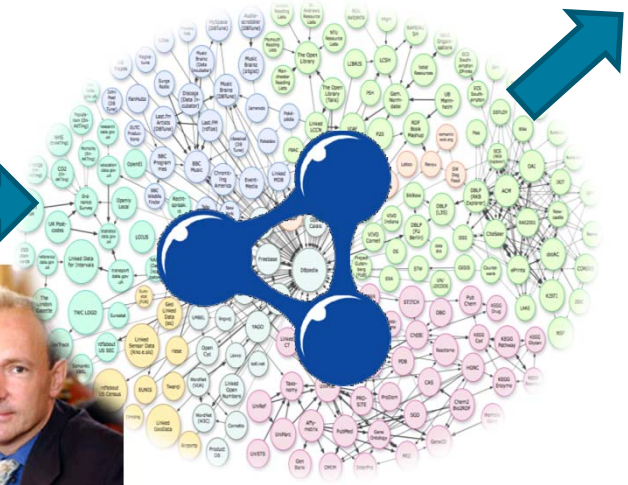
DBMS

DBMS DBMS DBMS DBMS DBMS DBMS DBMS

Web Apps

Extraction





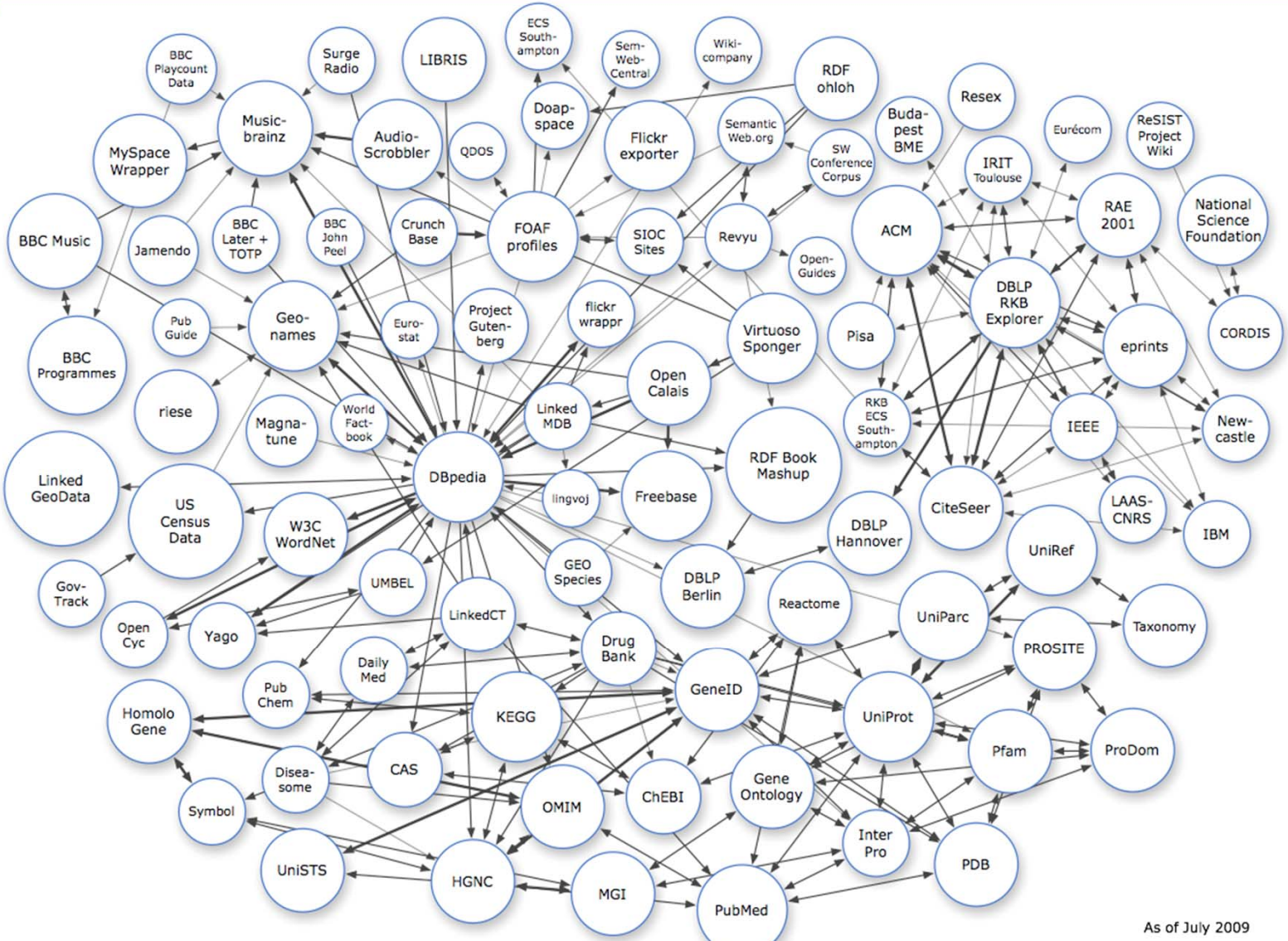
Linked Data & Data Spaces – a database guy's PoV

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!

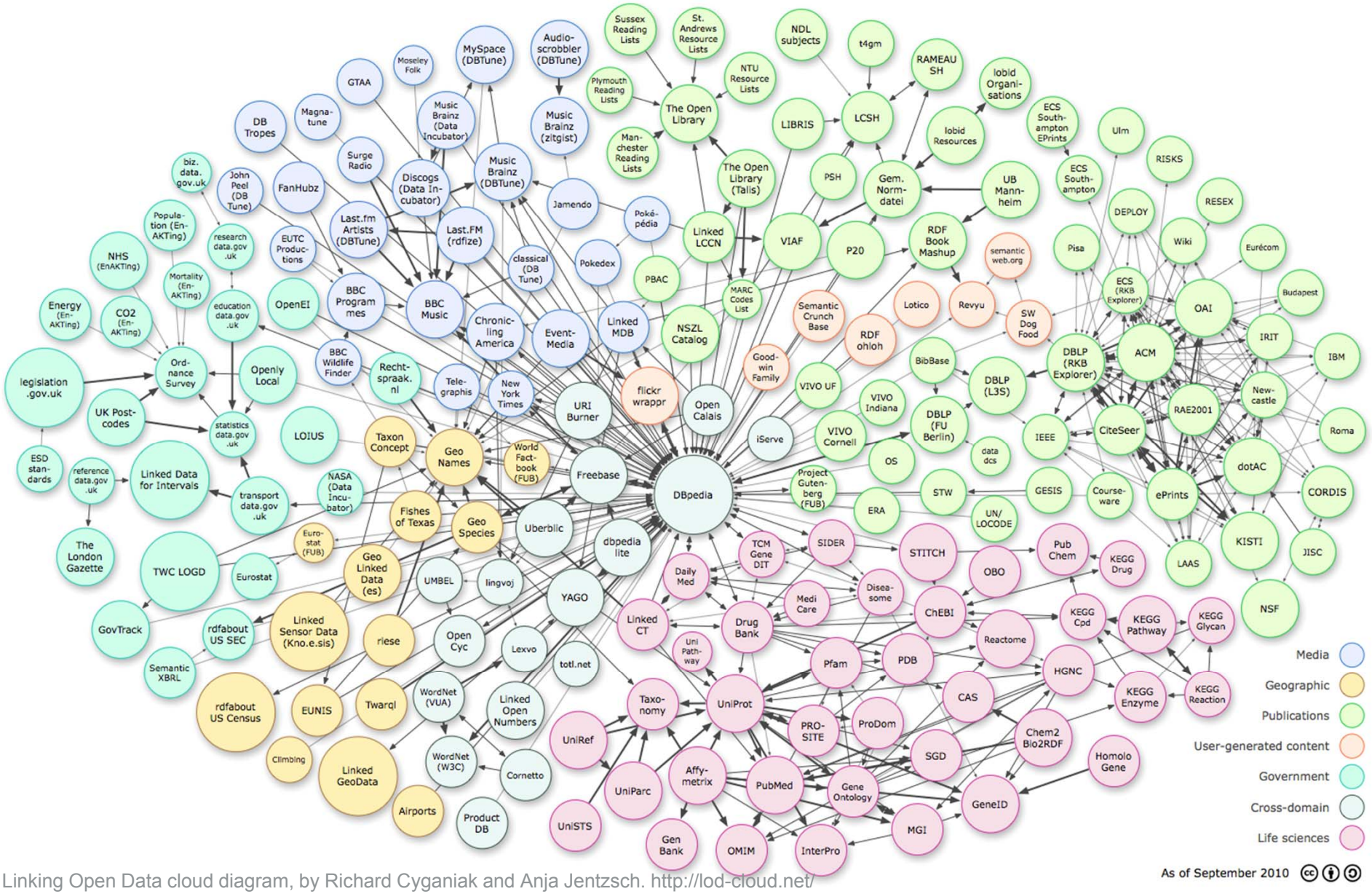
- The Good
  - Comes as triples
    S: http://.../Uppsala
    P: location
    O: http://.../Sweden
  - Often user generated
  - Nice domains
  - Free
- The Bad
  - Voluminous
  - Heterogeneous
- The Ugly
  - Dirty, inconsistent, sparse

Linked Data Graph



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

Linked Data Graph



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 -	--	--	--	--	--	--	--	--	--
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	--	--	--	--	--	--	--	--	--	--	--	--	--

- 672 million triples
  - 286 million English
- From 97 languages of Wikipedia
- 3.5 million things
  - 364,000 persons
  - 462,000 places
  - 99,000 music albums
  - 54,000 films
  - 16,500 video games
- http://wiki.dbpedia.org/Datasets

And more sources

- Government data
  - www.data.gov
  - 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, …
- …

„Raw data now!"

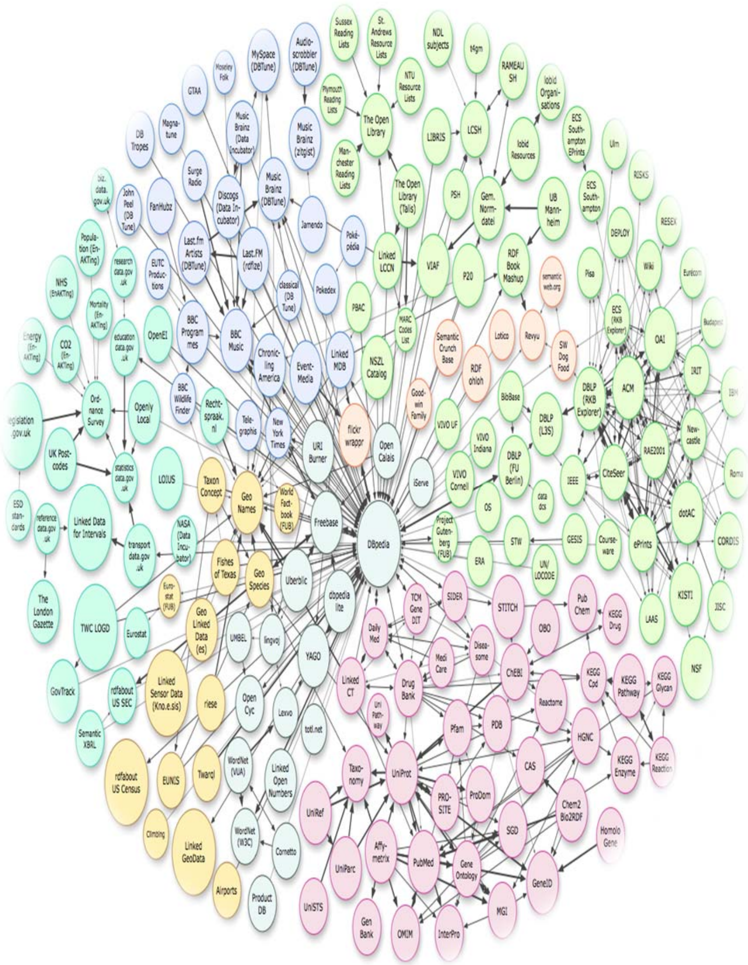
Use cases

- General purpose integration: Create rich knowledge bases
  - Semantic Web
  - Improved search / question answering
  - Link creation and data enrichment
  - Cleansing: data correction and validation
- Domain specific integration
  - Creation of high quality data sets: Complete & accurate
  - Enhancement of organization-internal data
  - Create reference data sets
  - Mashups

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



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).

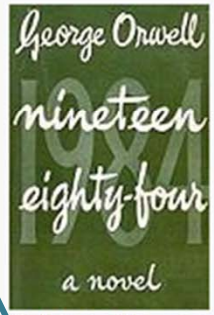
Nineteen Eighty-Four (often abbreviated to ***1984***) is a classic dystopian novel by English author George Orwell. Published in 1949, it is set in the eponymous year and focuses on a repressive, totalitarian regime. Orwell elaborates on how a massive oligarchical collectivist society such as the one 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 th itself has come to refer to anything reminiscent generally considered to be George Orwell's mag

Contents [hide]

iPopulator

Nineteen Eighty-Four (1984)

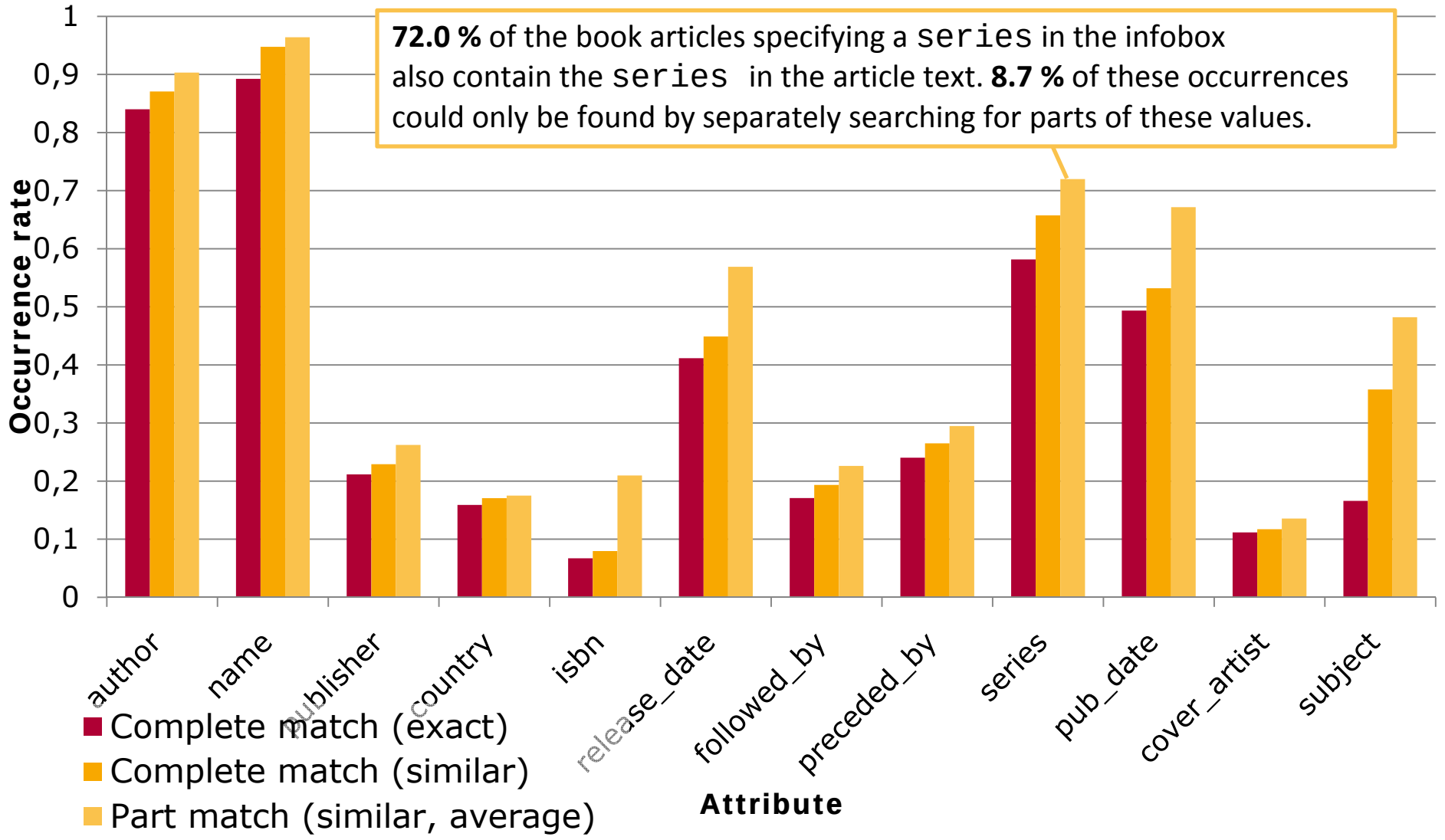


British first edition cover

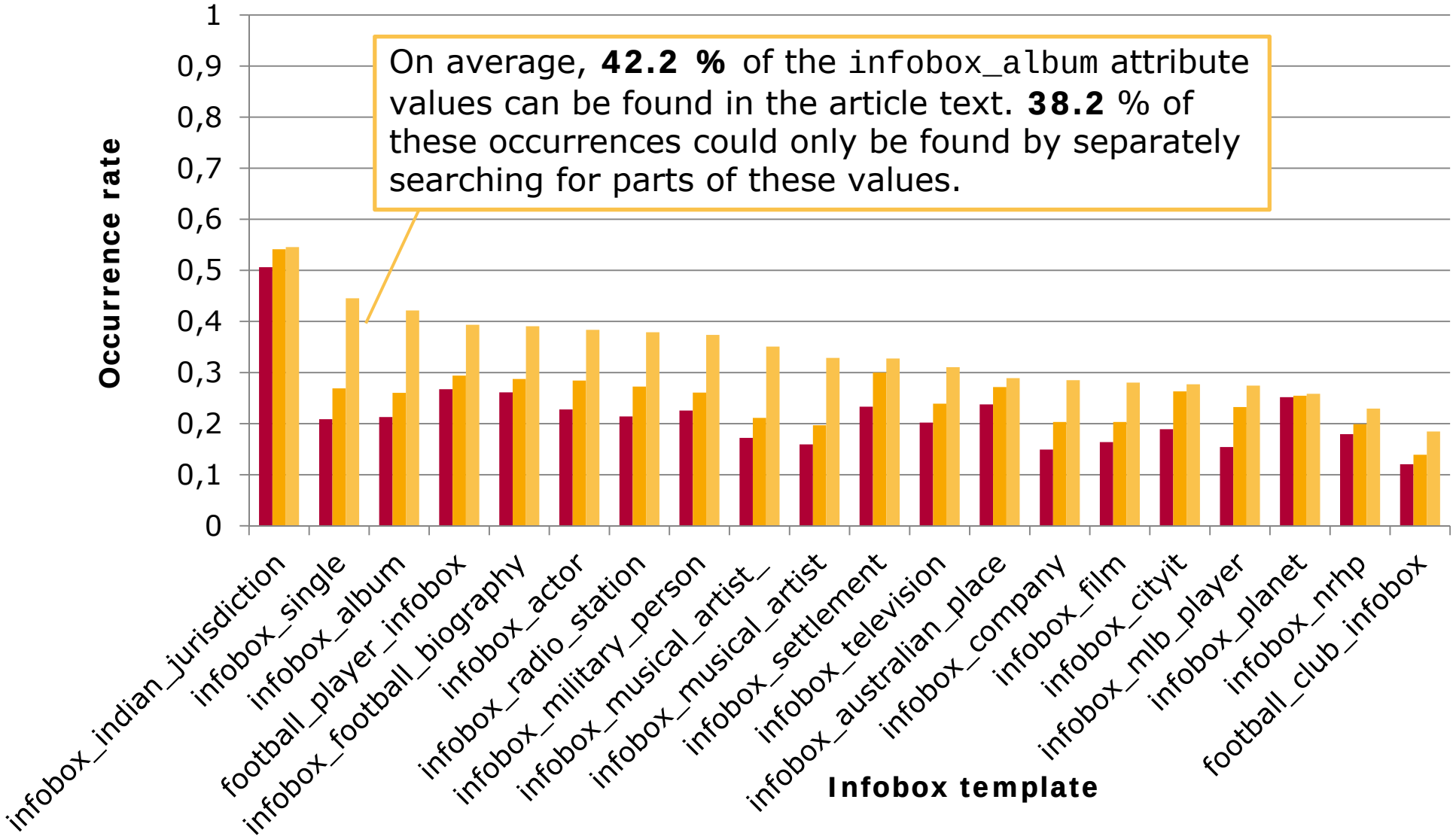
Author	George Orwell
Country	United Kingdom
Language	English
Genre(s)	Dystopian, Political novel, Social science fiction
Publisher	Secker and Warburg (London)
Publication date	8 June 1949
Media type	Print (Hardcover & Paperback) & e-book, audio-CD
Pages	326 pp (Paperback edition)
ISBN	97

Master thesis by Dustin Lange
Now PhD student at HPI
Topic: similarity search

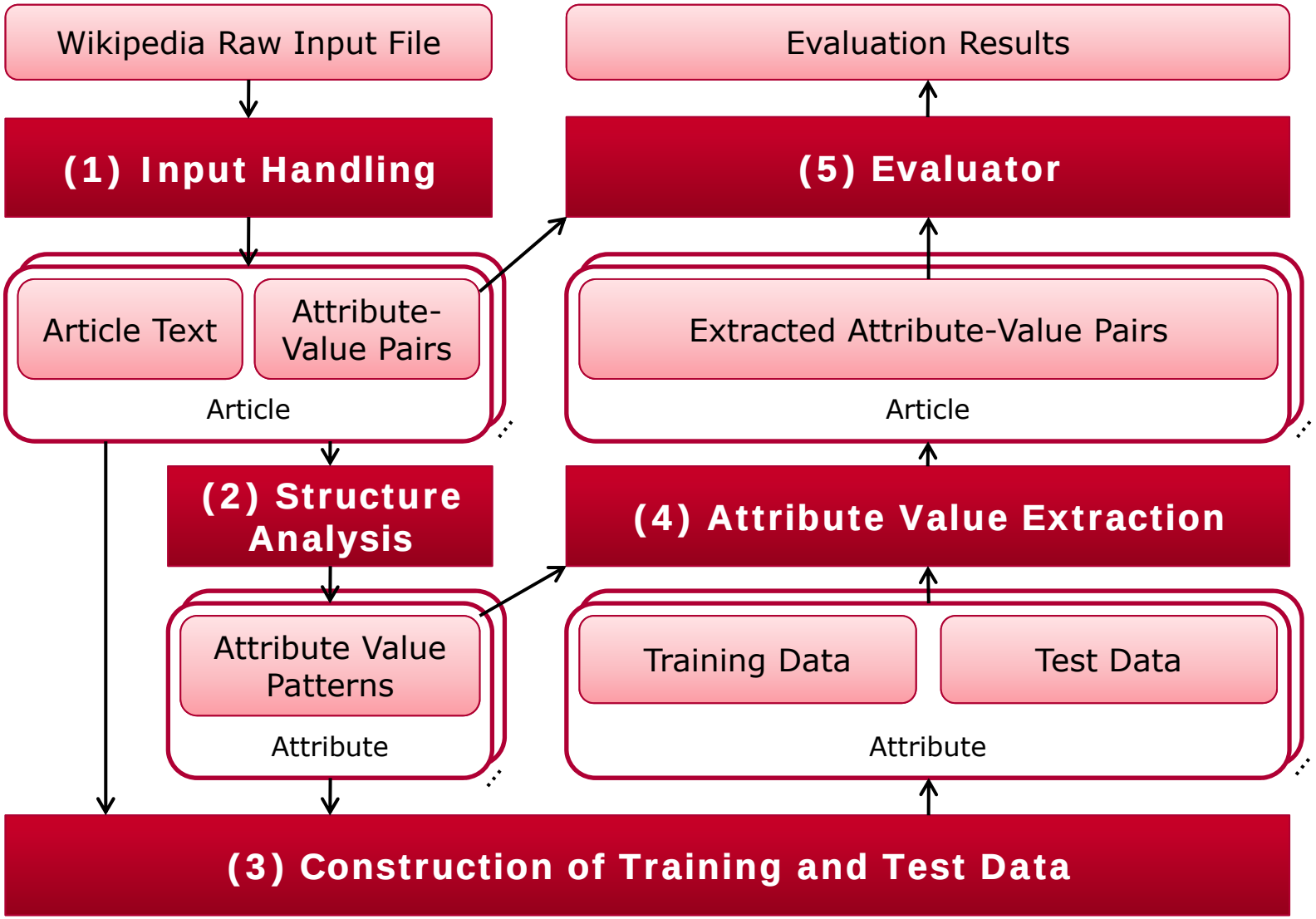
Occurrence of values in article text: 12 most frequent attributes in infobox_book



20 most frequent templates

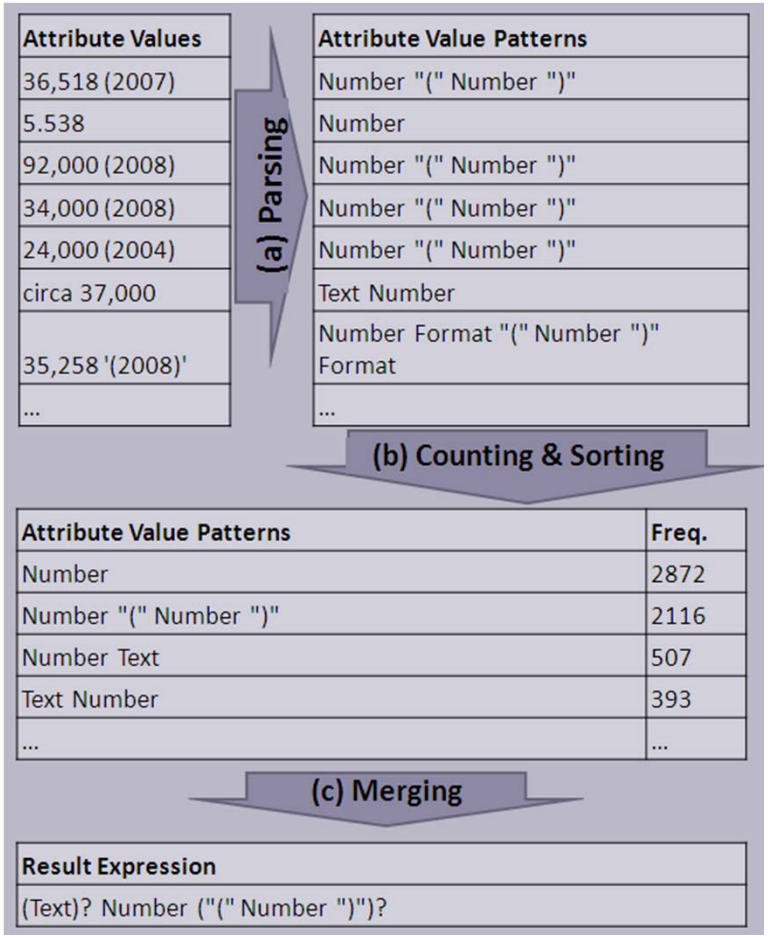


Architecture of iPopulator



Structure Analysis

- Values of an attribute often share similar structure.
  - Extract value parts
  - Constructing homogeneous values from parts
- Determine common structure for each infobox template attribute
- Example: number_of_employees from infobox_company



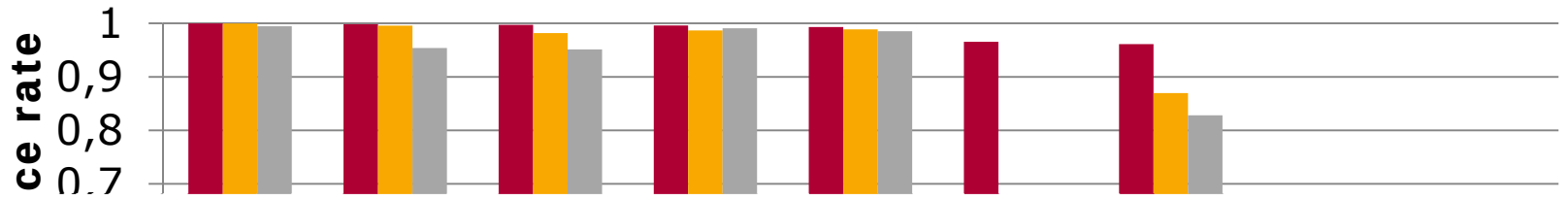
Training Data and Extraction

- Exploit existing infoboxes as training data
- Mark occurrences of infobox attribute values as training examples
  - Similarity measure to label fuzzy occurrences
- Automatic extraction method learns to recognize these occurrences by analyzing token (word-level) features
- Create extractors for thousands of infobox template attributes
- Extract parts of attribute values from different article text positions
- Arrange extracted value parts

Example: IBM employed 399,409 people in 2009.

Structure	Number	“(“	Number	“)”
Extracted value parts	399,409		2009	
Result value	399,409 (2009)			

Evaluation: infobox_planet

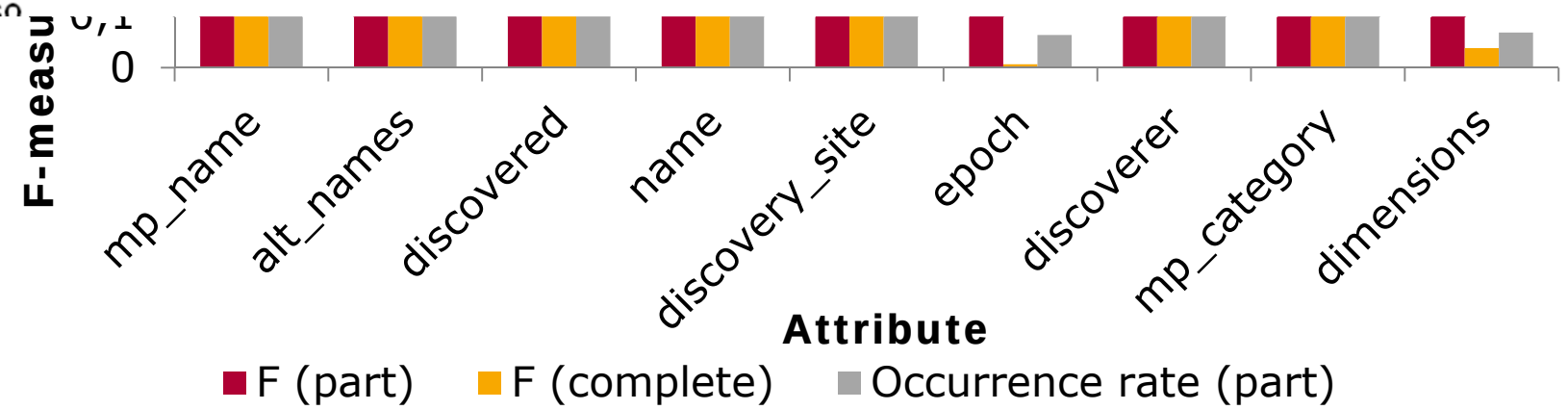


22032 Mikekoop

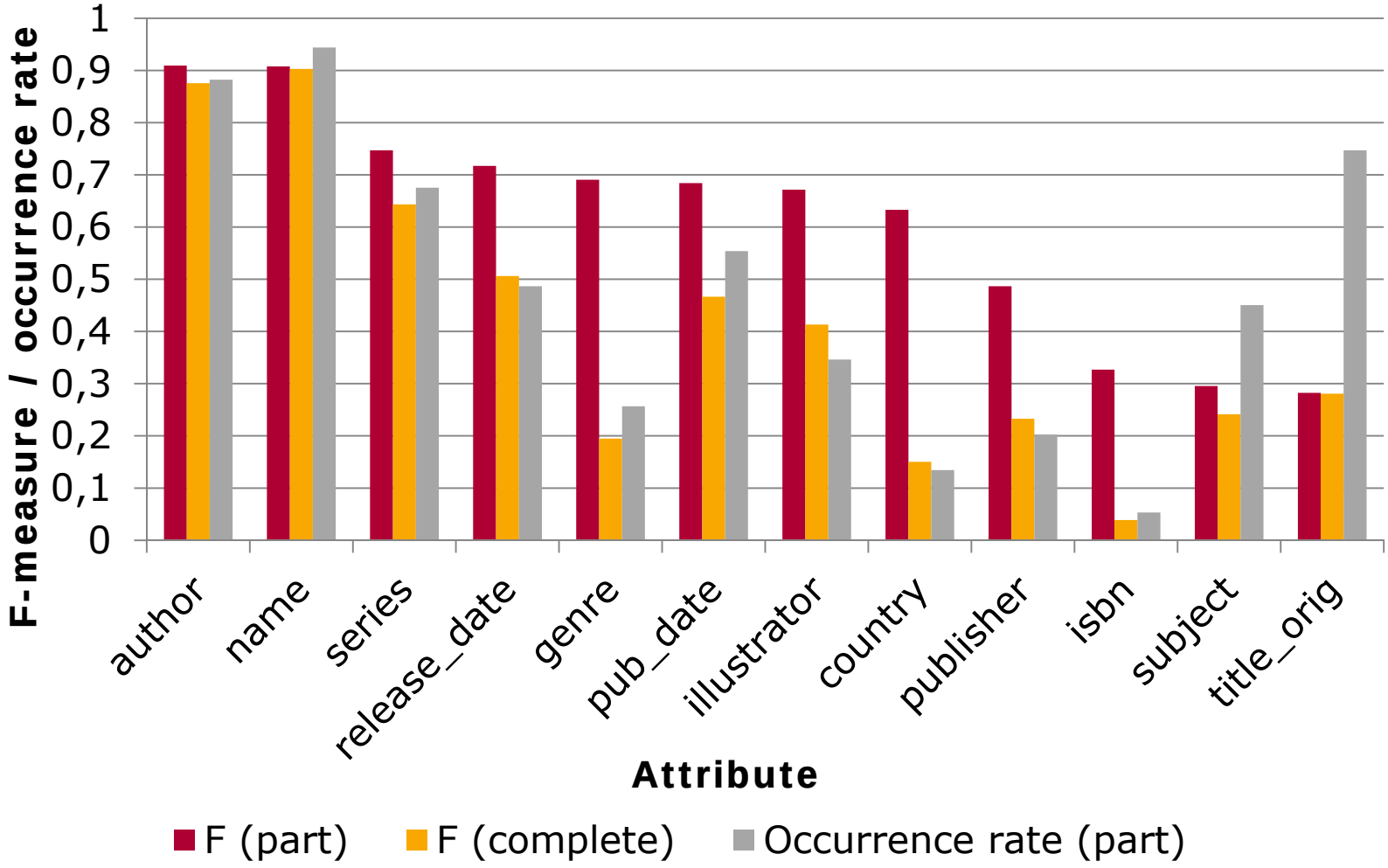
From Wikipedia, the free encyclopedia

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.

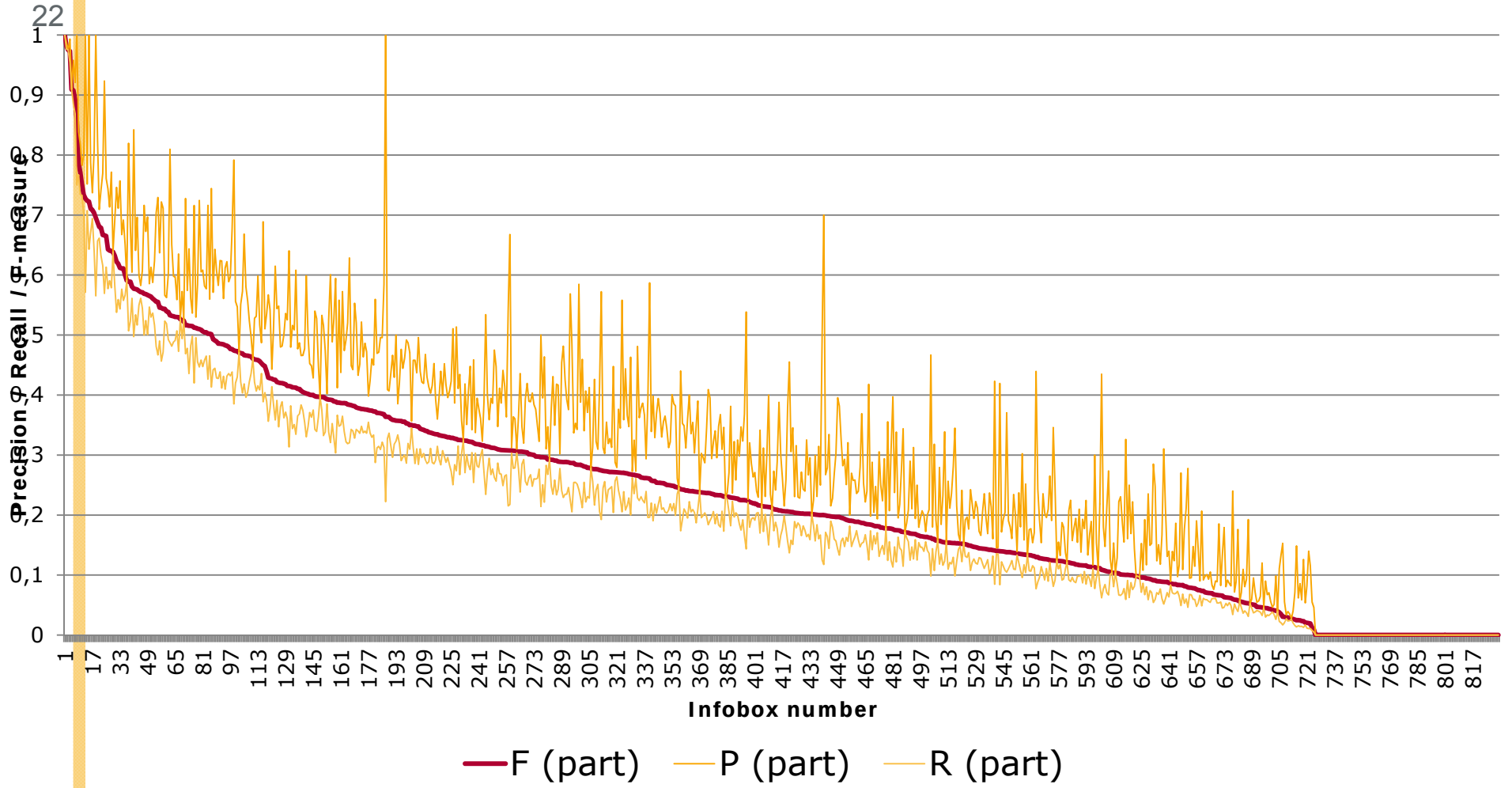
See also [edit]



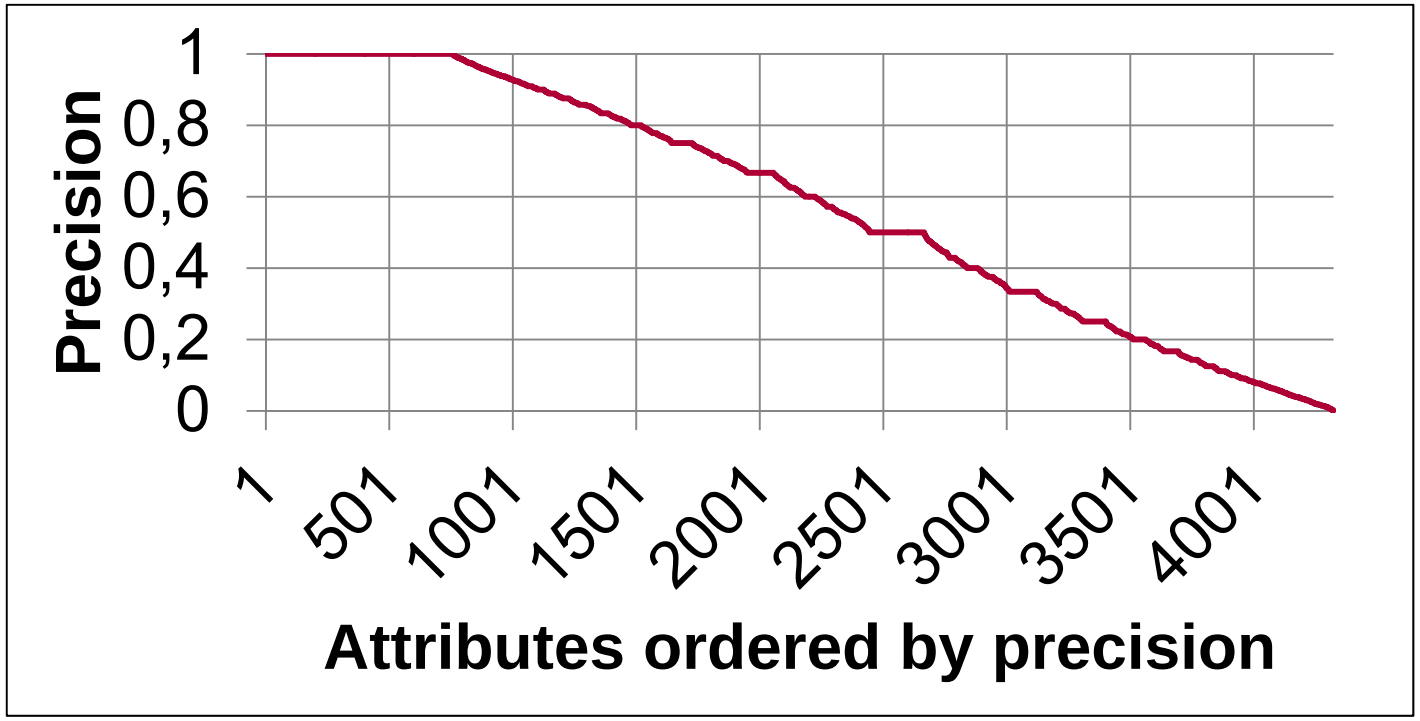
Evaluation: infobox_book



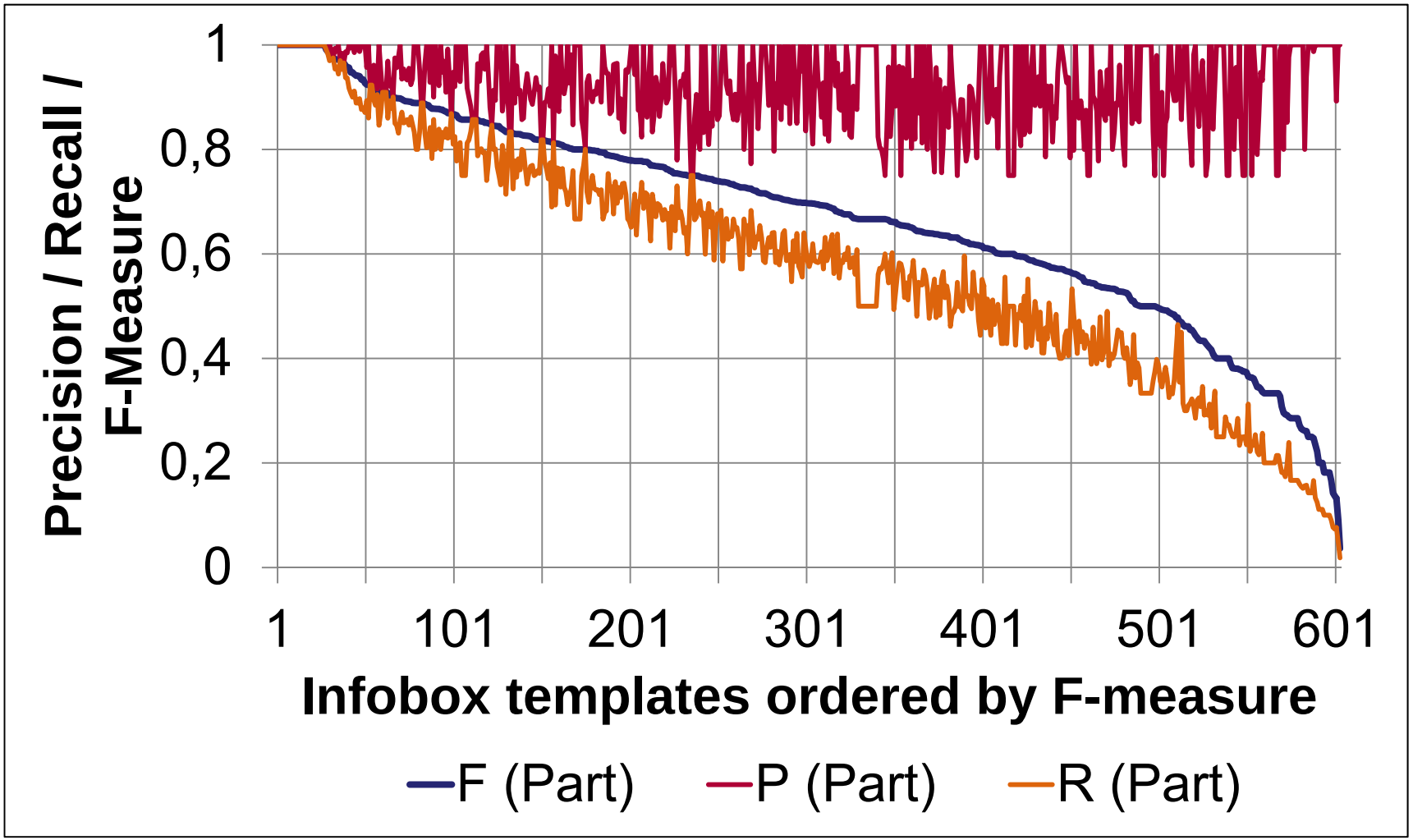
Evaluation: All Infobox-Templates



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

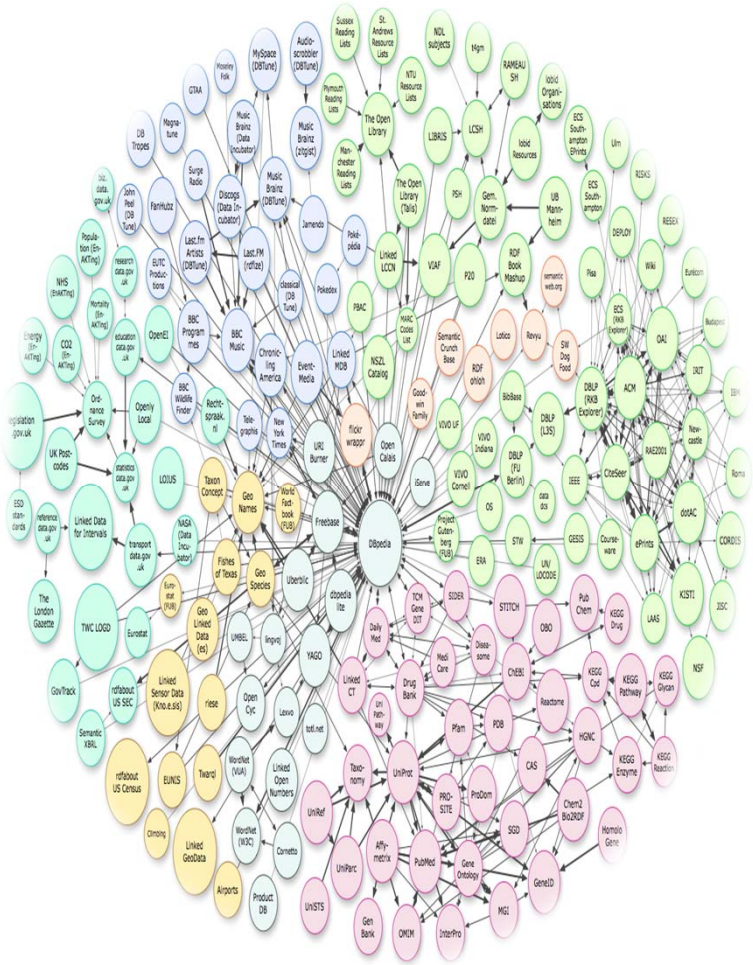


Evaluation on all attributes with Precision > 0.75



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, rules
  - Bandwidth ⟷ Patience
- Schema
  - Structure ⟷ Schema Mapping
  - Semantics ⟷ Domain knowledge
- Data
  - Formatting ⟷ Scrubbing
  - Duplicates ⟷ Entity Matching

Now: Examples for each

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 %

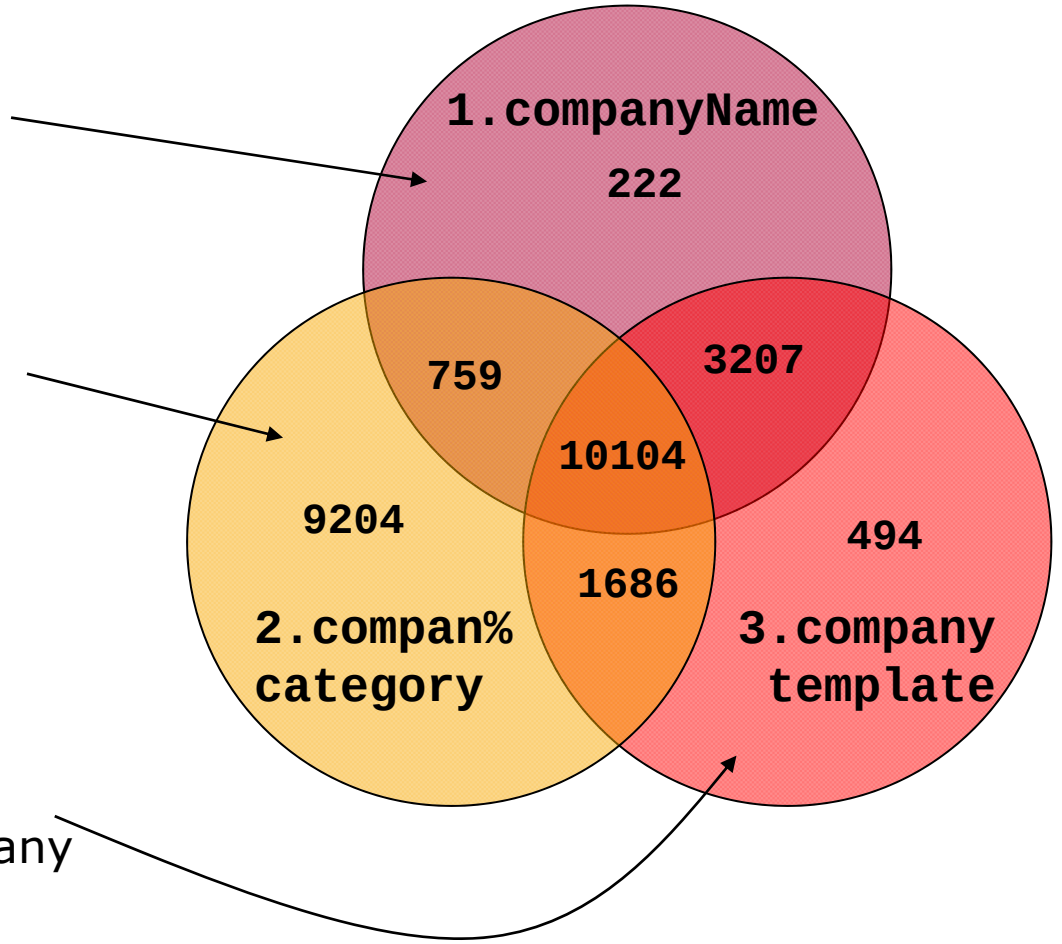
Beneficiary

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*REPUBLIC		
Address:			

```
{ "_id" : "euFinance#28994",
  "year" : 2008,
  "nameOfBeneficiary" : "ROBERT BOSCH GMBH*",
  "coordinator" : false,
  "countryTerritory" : "Germany 70049 STUTTGART",
  "coFinancingRate" : "67,51 %",
  "amount" : 3199959.00,
  "commitmentPositionKey" : "F13.A22622.1",
  "subjectOfGrantOrContract" : "MULTISPECTAL TERAHERTZ, INFRARED ...
  "responsibleDepartment" : "Information Society and Media",
  "budgetLineNameAndNumber" : "Support for research ...
}
```

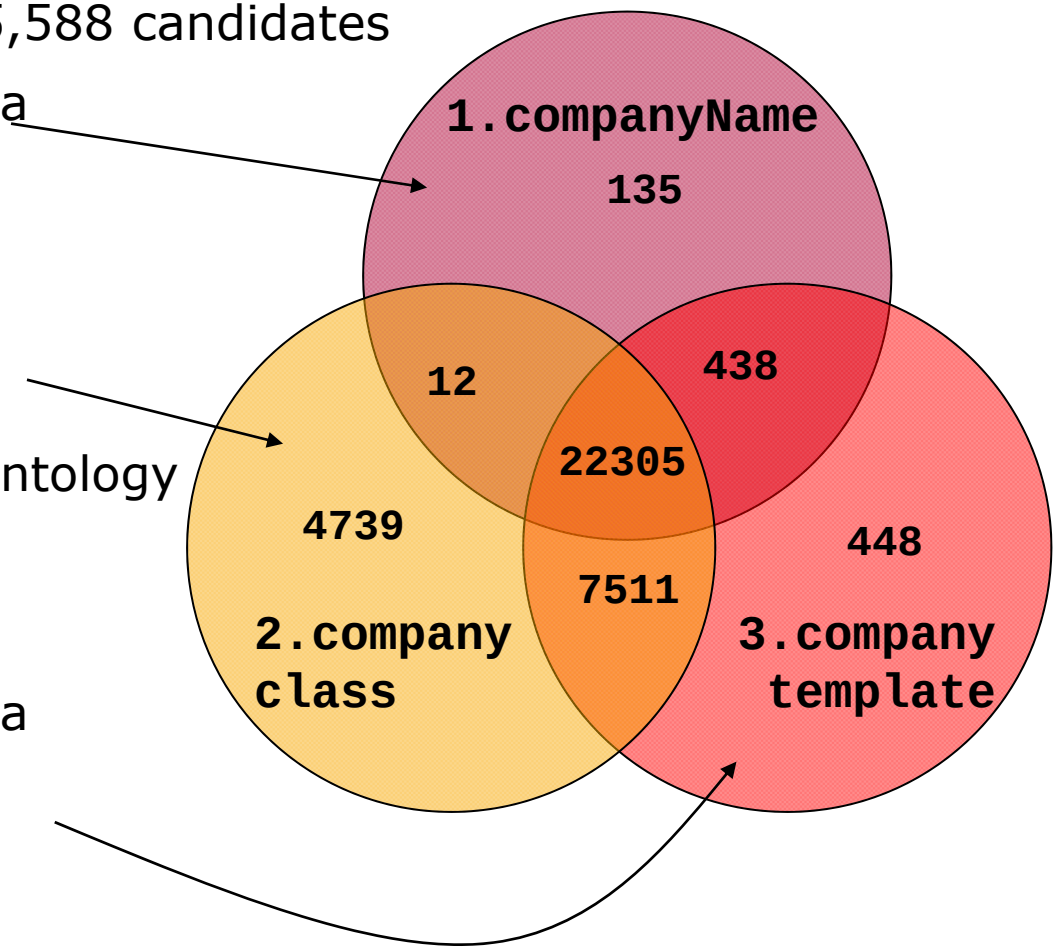
The problem – a domain mess 2008

- What is a company?
- Def. 1: Entities having a companyName
  - 14292 companies
- Def. 2: Entities in a category that starts with 'compan%`
  - 21753
- Def. 3: Entities having a wikiPageUsesTemplate with value Template:infobox_company
  - 15491



The problem – a domain mess 2011

- 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

- Wikipedia/DBpedia: 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**

Infoboxes with CompanyTemplate

- 1083 different attributes
  - 499 appear only once
- Of the 1083 attributes, 39 distinct ones contain ‘name’ as substring
- 273 companies without any name attribute

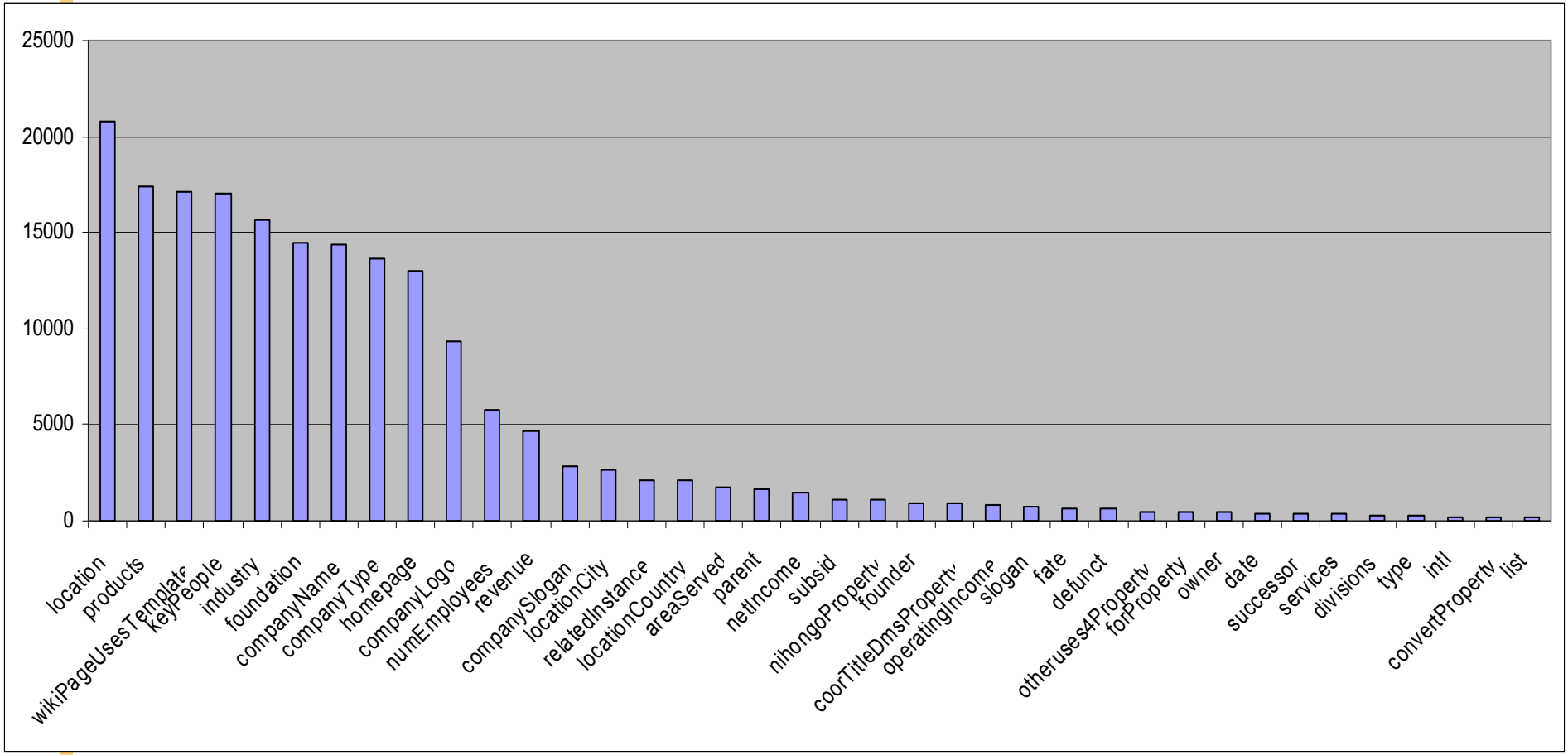
location	20617	companyName	13355
products	18176	name	2036
wikiPageUsesTemplate	18048	surname	25
keyPeople	17836	railroadName	8
industry	16822	companyNickname	4
foundation	15826	pastNames	4
homepage	14476	absNameProperty	3
companyType	13433	dnvNameProperty	3
companyName	13355	labelName	3
companyLogo	9006	logoFilename	3
numEmployees	6207	dvdEuroCompanyName	2
revenue	5030	filename	2
locationCity	4098	longName	2
locationCountry	3212	websitename	2
companySlogan	2815	alternativeNames	1
areaServed	2557	birthname	1
relatedInstance	2284	brandName	1
type	2152	bTcgvuvCompanyName	1
parent	2054	companyNameLocal	1
name	2036	companyNamesBigBum	1
netIncome	1663	europeanTradeAssociationCompanyName	1
founder	1597	familyCorporationCompanyName	1
subsid	1232	formerNames	1
nihongoProperty	1141	fukCompanyName	1
slogan	1087	golfFacilityName	1
coorTitleDmsProperty	960	hangulName	1
logo	925	iceCreamCompanyName	1
services	904	nativeName	1
operatingIncome	896	nickname	1
owner	680	officialName	1
otheruses4Property	510	oldName	1
intl	503	organisationName	1
forProperty	467	publicCompanyName	1
divisions	429	renamed	1
date	422	shortName	1
locations	419	wineryName	1

Infoboxes in Company class 2011

- 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

Profiling Companies



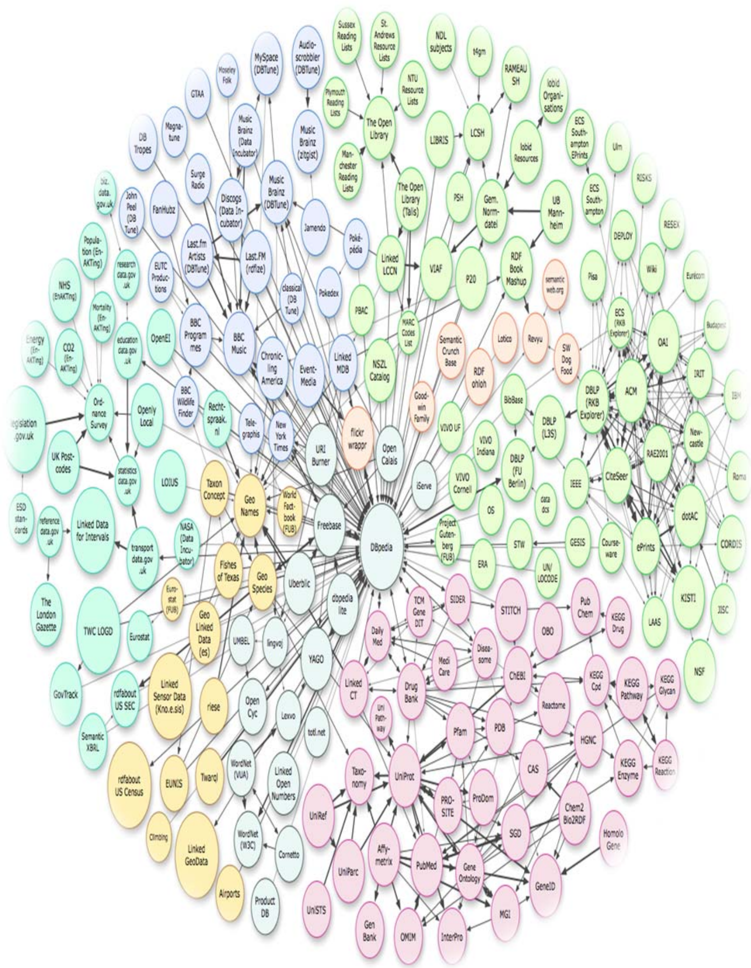
fieldName	<info>	Dollars Obligated	Current Contract Value	Ultimate Contract Value	Major Agency	Modified Contracting Agency	Contracting Agency	Contracting Office	Program / Funding Agency	Program / Funding Office	Reason For Purchas For DoD
example1		$220,989,132	$220,989,132	$220,989,132	Dept. of Defense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

The problem – a data mess

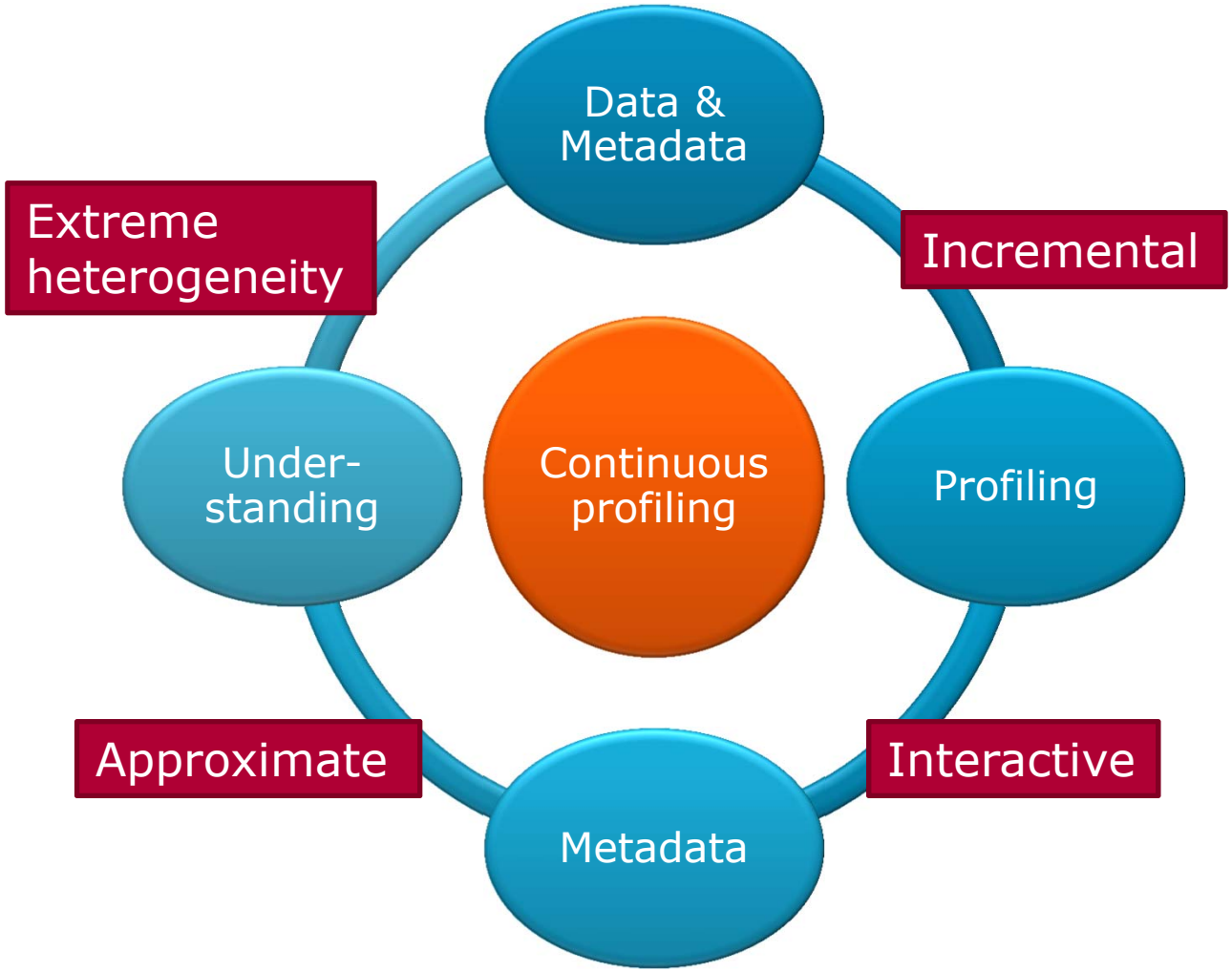
- Poor schemata: No types, no constraints
- Sloppy data entry:
  - Data value are neither standardized nor normalized
- Revenue attribute in DBpedia may contain different units, different currencies, and different number-formats.
  - 1.64 billion USD vs. $1640 m vs. 1,6 vs. more than one million Euro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

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

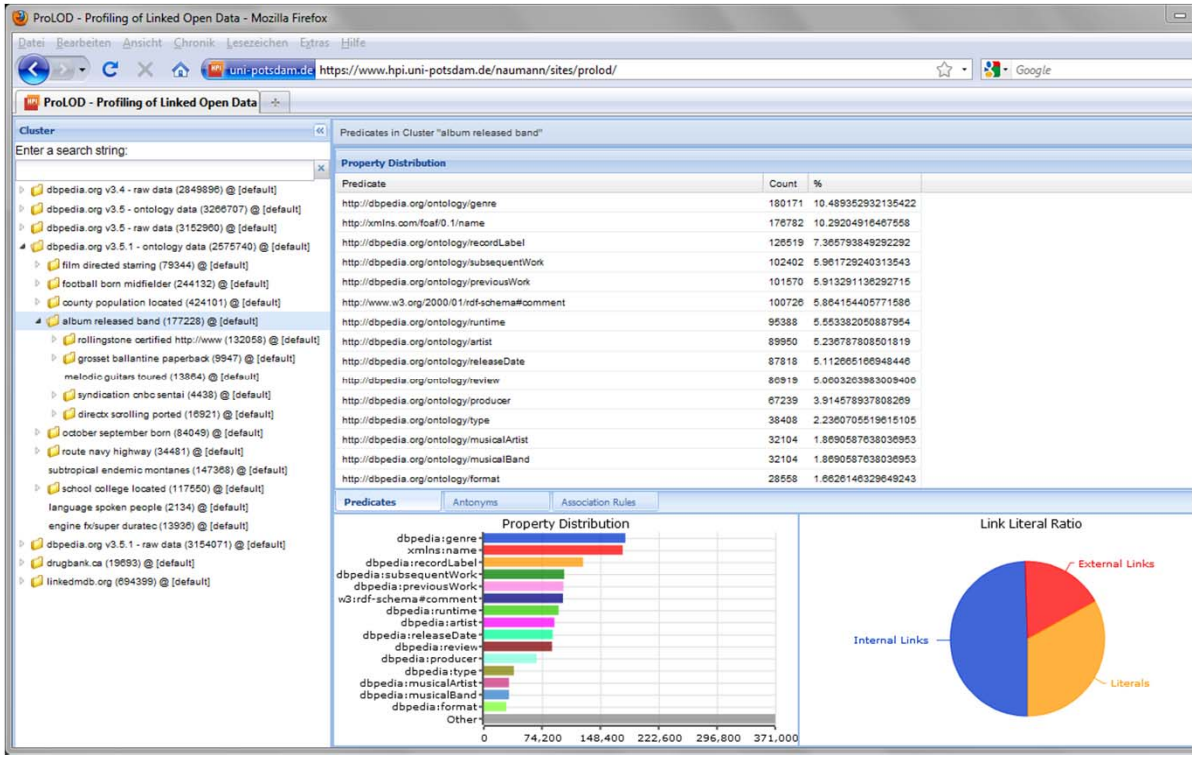


Data Profiling



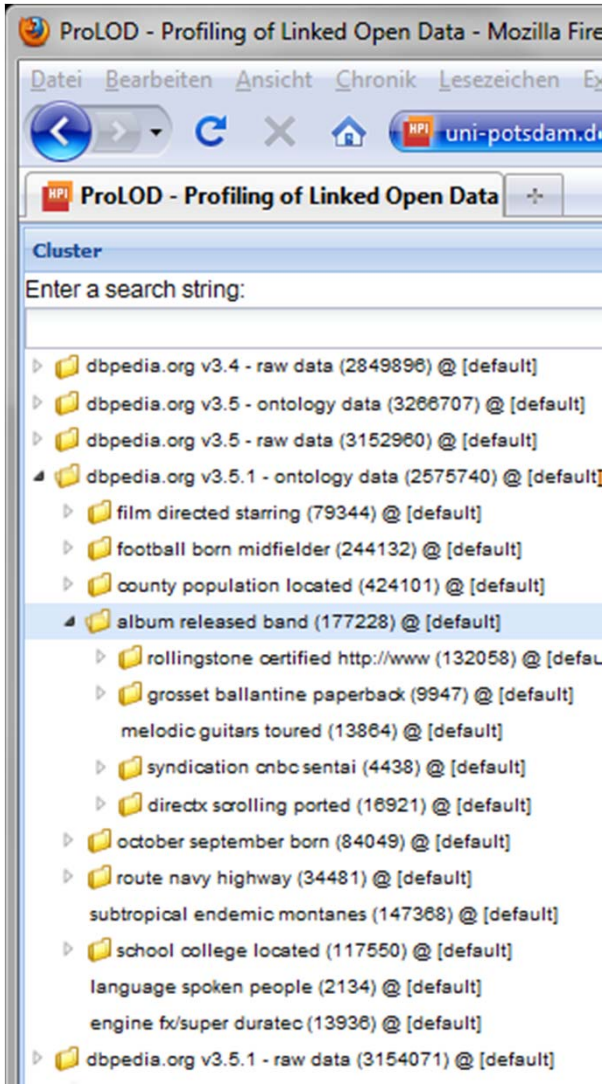
Prototype: ProLOD

- Platform for ongoing and future work
  - https://www.hpi.uni-potsdam.de/naumann/sites/prolod/
- Steps:
  - Data upload
  - Preprocessing
  - Visualization

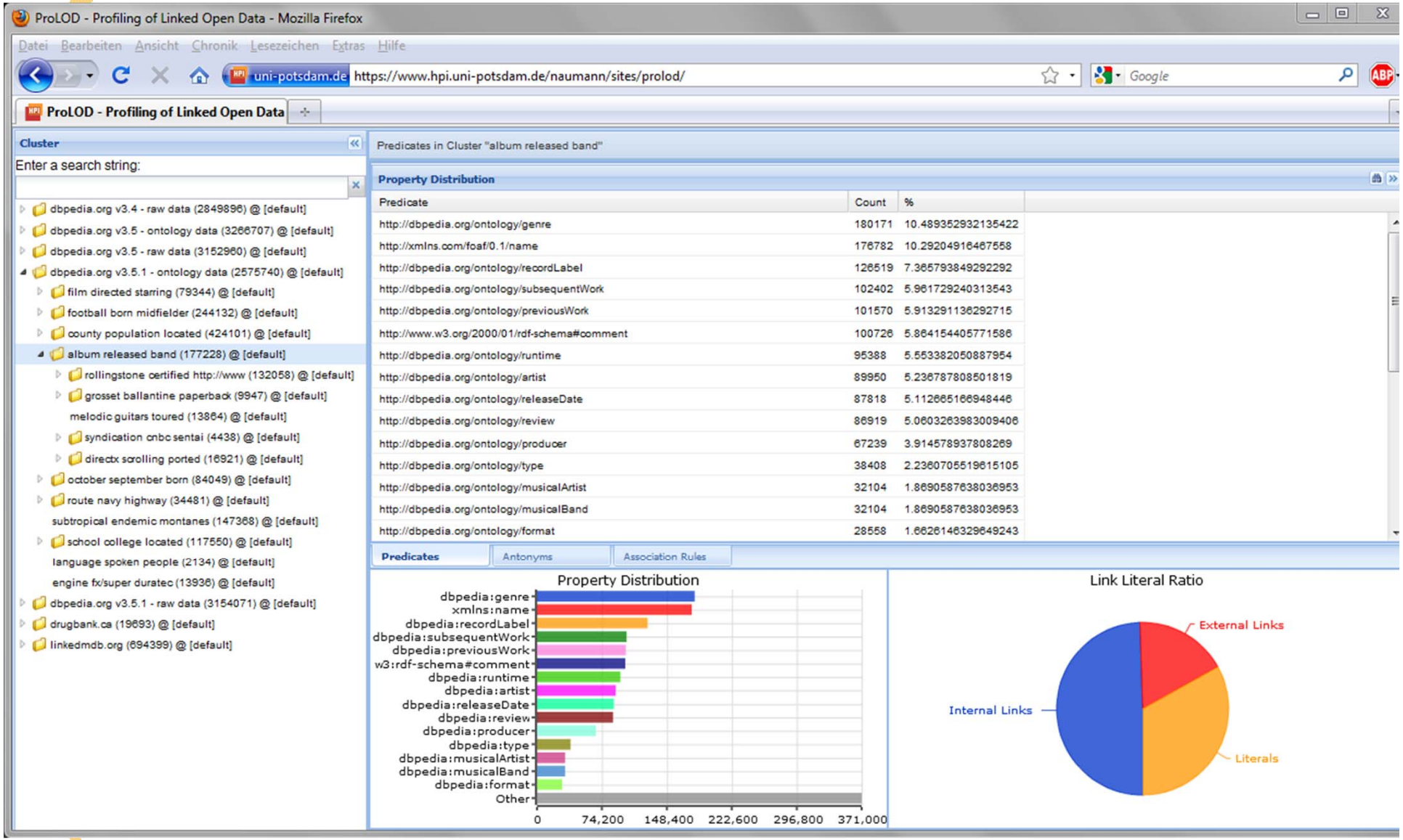


ProLOD profiling tasks

- 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

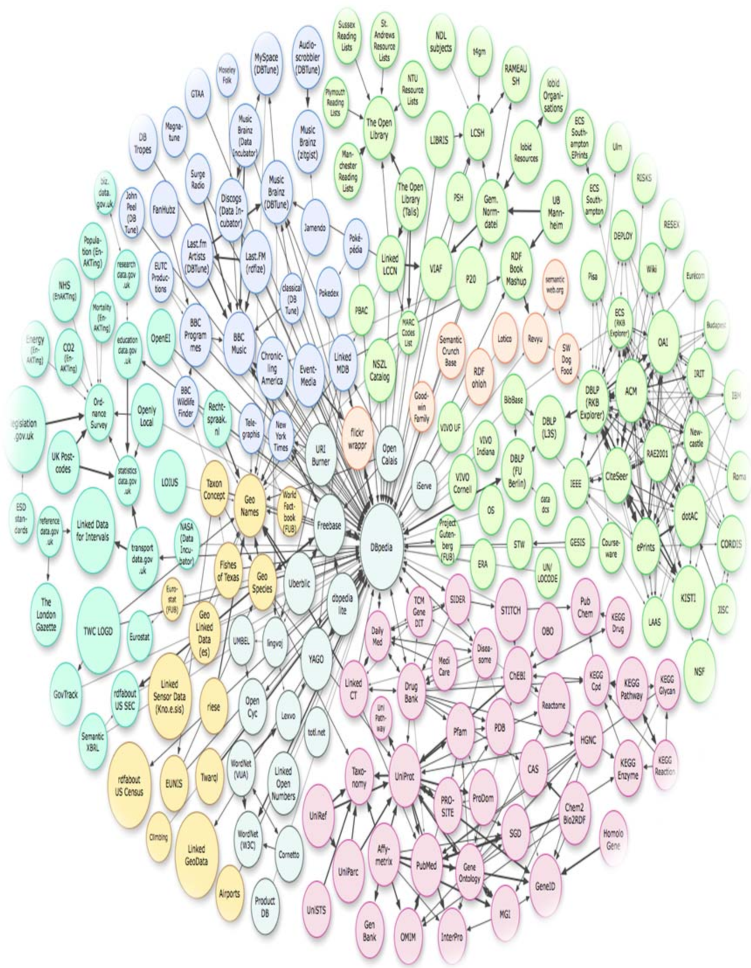


ProLOD – Profiling Linked Open Data



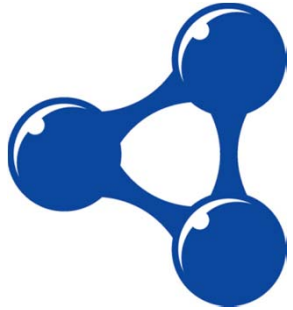
Overview

- 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



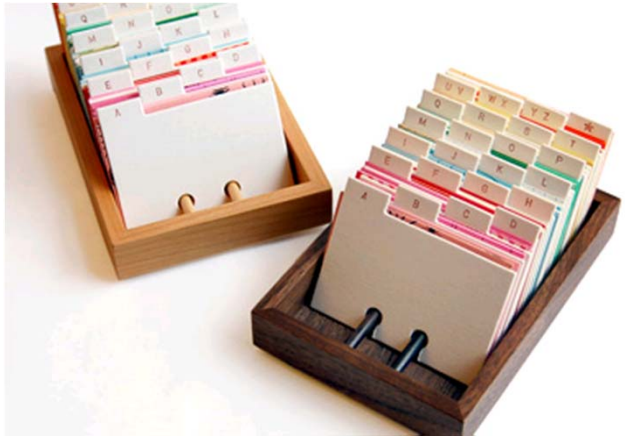
Midas – Integration project with IBM Almaden Research Center

- Linked Open Data (Midas,LOD)
  - Integrating DBpedia, Freebase, SEC and FDIC at the level company entities
- Regulatory sources (Midas.Finance)
  - Integrating unstructured/semi-structured data sources containing information about a wide range of entities (e.g., SEC and FDIC)
- Government (Midas.Gov)
  - Integrating structured data from government data sources like usaspending.gov, senate.gov, etc.
  - Persons, legal entities, funding



Five steps for integration

- Source Selection
- Schema Matching & Mapping
- Data Extraction & scrubbing
- Entity Matching
- Data Fusion



Five steps – 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

Five steps – Schema matching and schema mapping

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

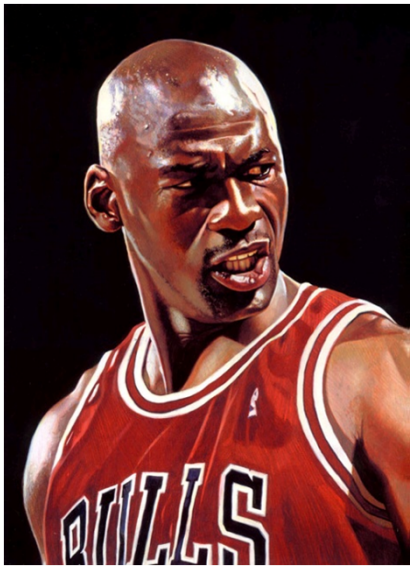
Final Schema	DBPedia	SEC	Freebase
dbpediaURI			/type/object/key
cik	secCik	CIK	
irsnumber			
companyName	companyName, name, nonProfitName	name	/type/object/name, /common/
address		BusinessAddress, MailingAddress	/location/mailing_address/stre /location/mailing_address/pos
locationCity	locationCity, location	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		

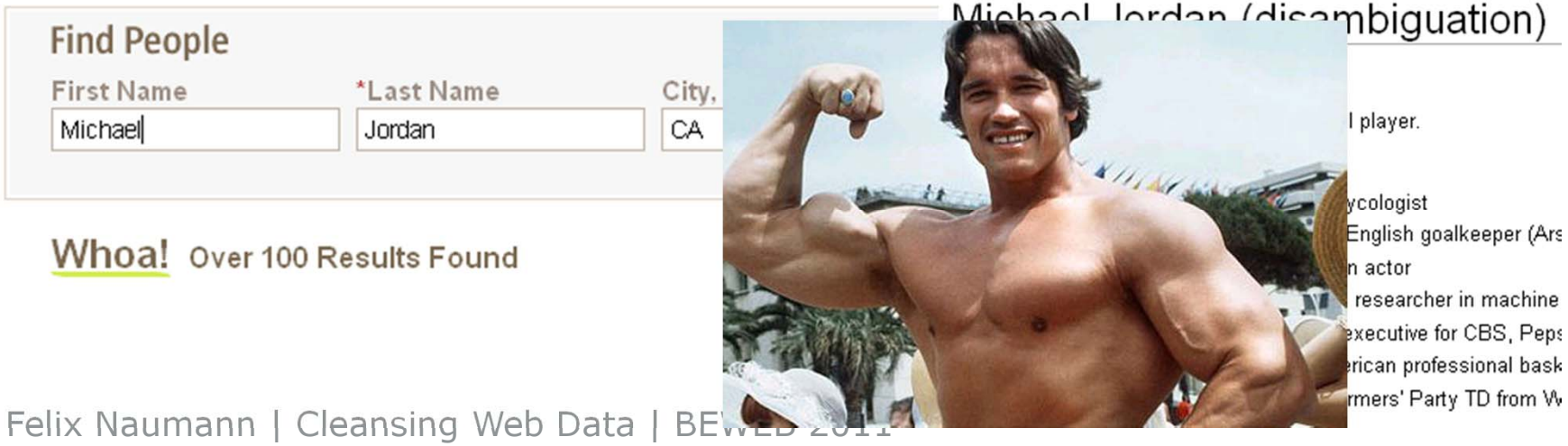
Five steps – Data extraction & scrubbing

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

Five steps – Entity matching

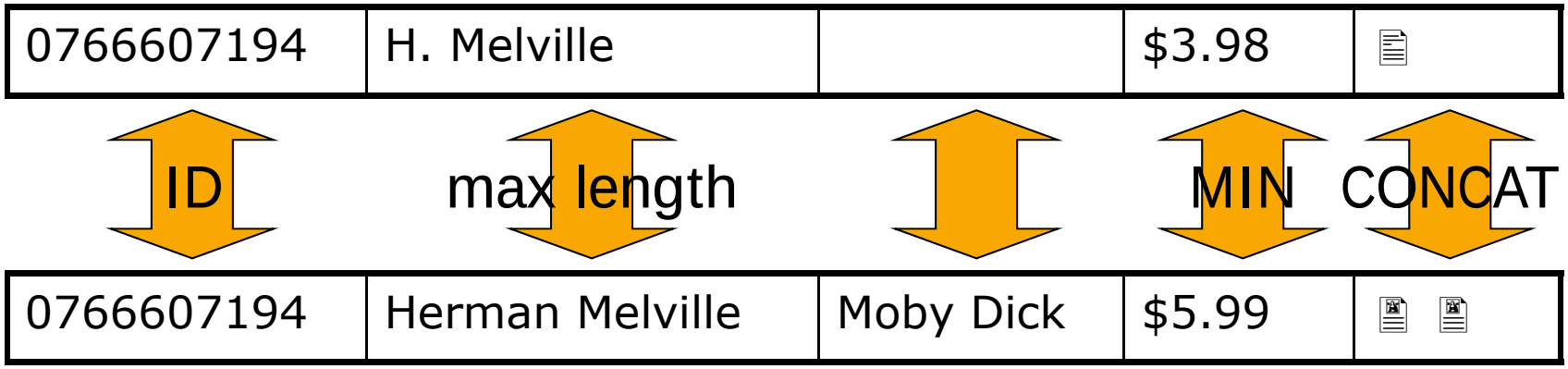
- Duplicate entries
- Linking between entries
- Challenges
  - Fuzzy matching: Similarity measures
  - Data volume: Partitioning algorithms
  - Sparse data
    - ◇ "Michael Jordan visited Indianapolis"





Five steps – Data fusion

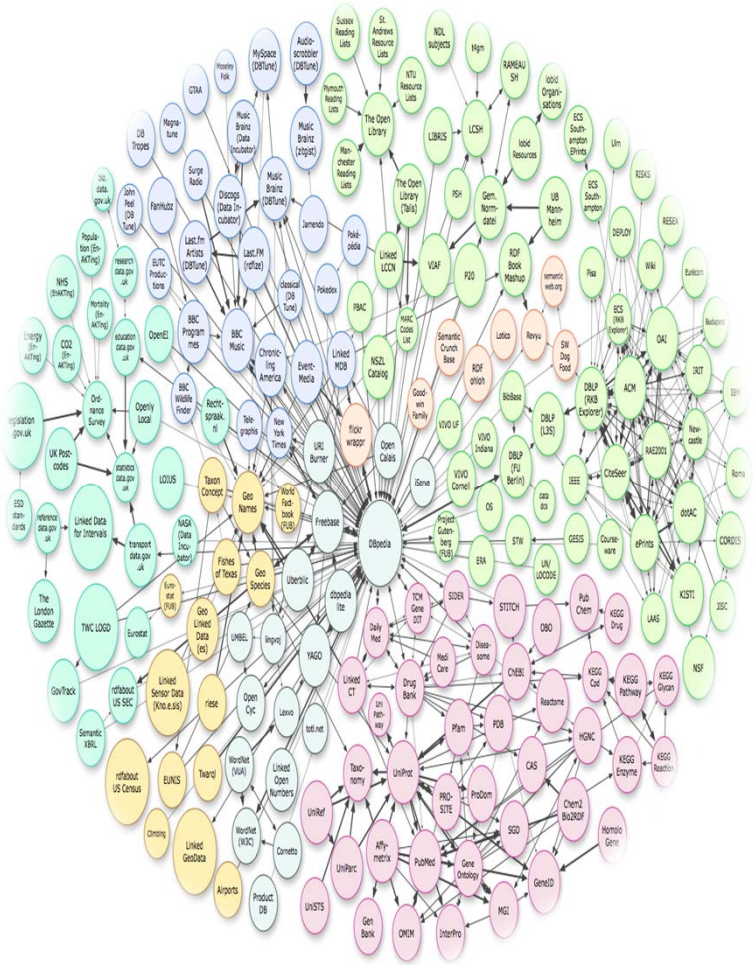
- Combine multiple representations of real-world entities
  - Survivorship, consolidation, etc.
- Resolve data conflicts
  - Conflict resolution functions
  - Reputation / accuracy / freshness -> “truth discovery”



- Retain data lineage

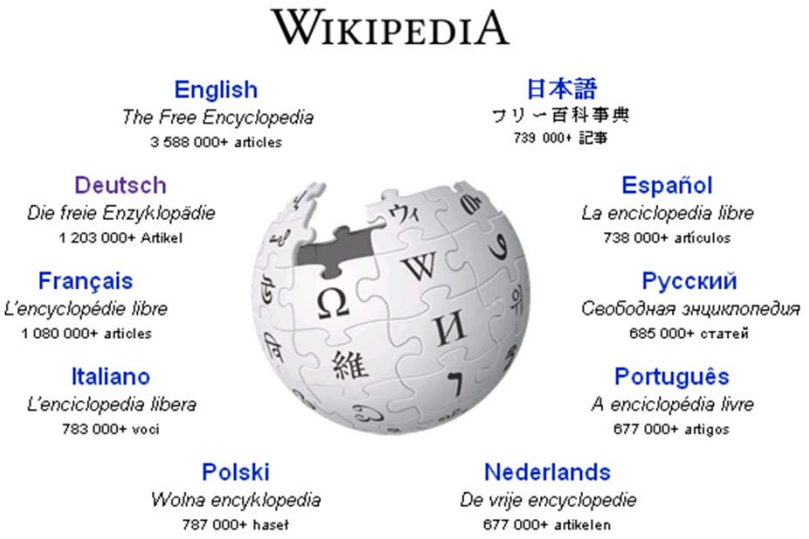
Overview

- 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



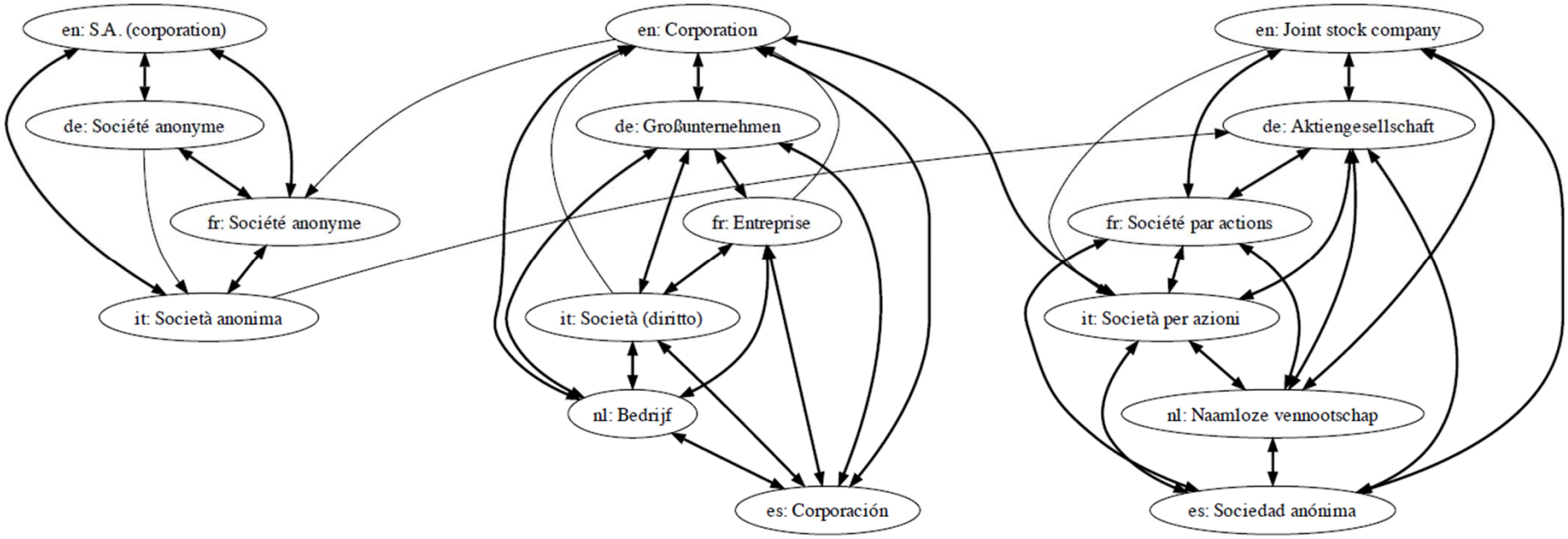
Multi-Lingual Wikipedia

- 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 281 languages (Mar 2011)

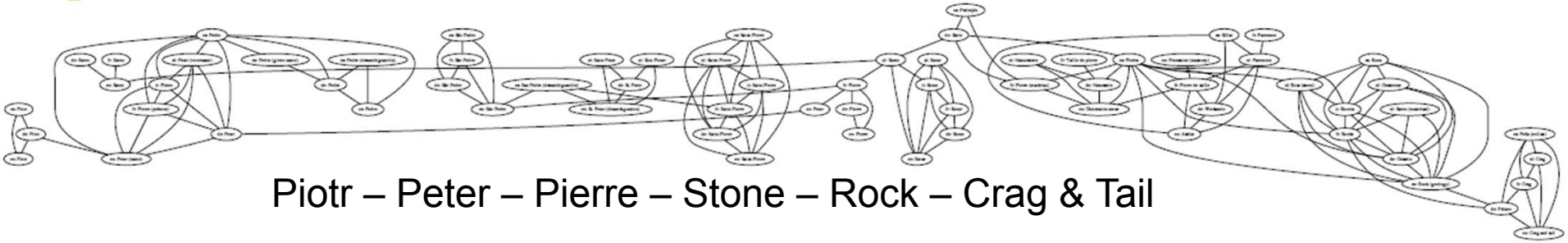


Interlanguage links (ILLs)

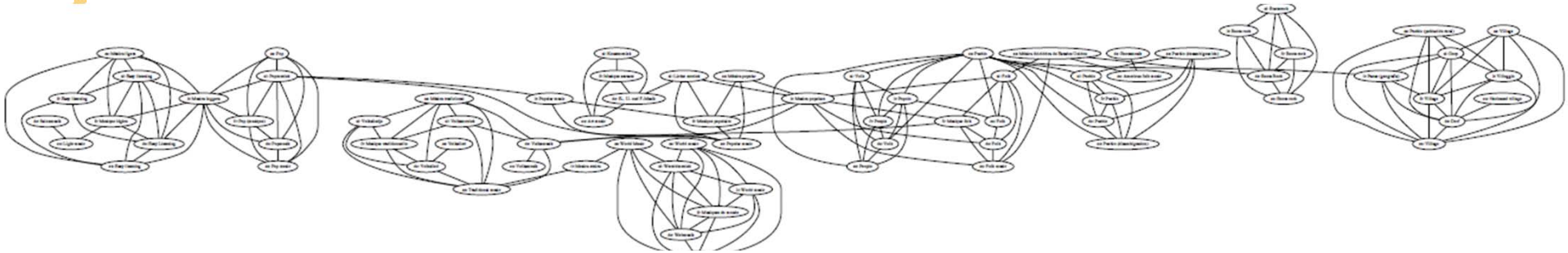
- First, evaluate quality of ILLs and build duplicate clusters
  - Build connected components using cross-language links (restricted to the six largest languages)
- But, largest weakly connected component has 108 articles
  - 26 English, 26 German, 21 French, 13 Italian, 13 Dutch, 9 Spanish articles



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

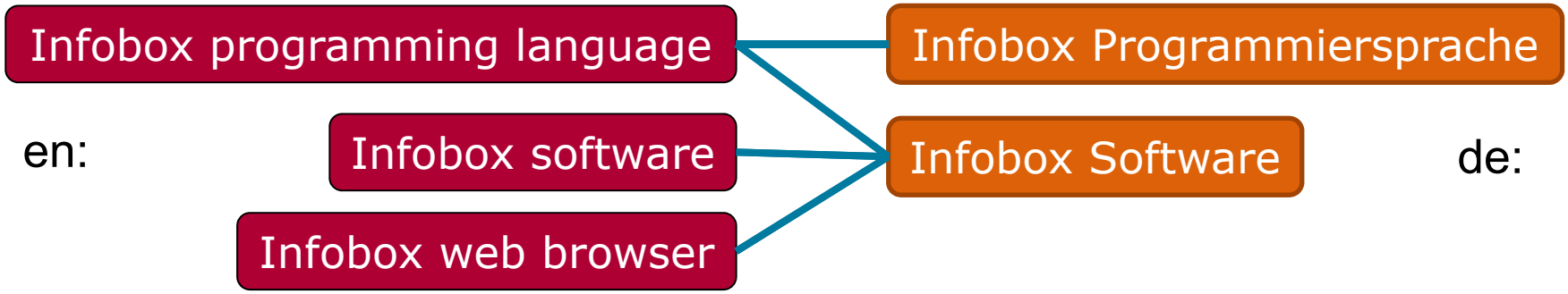
BCC
- Bidirectionally connected components (BCC)
- Undirected graph of bidirectional components is connected
- 4,241 BCCs of which 2,980 are incoherent

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

- Match schemas of **corresponding** infobox templates only.
- Different granularities in templates => n:m mapping
- **Idea:** Count co-occurrences of infobox templates in terms of connected components
- 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 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

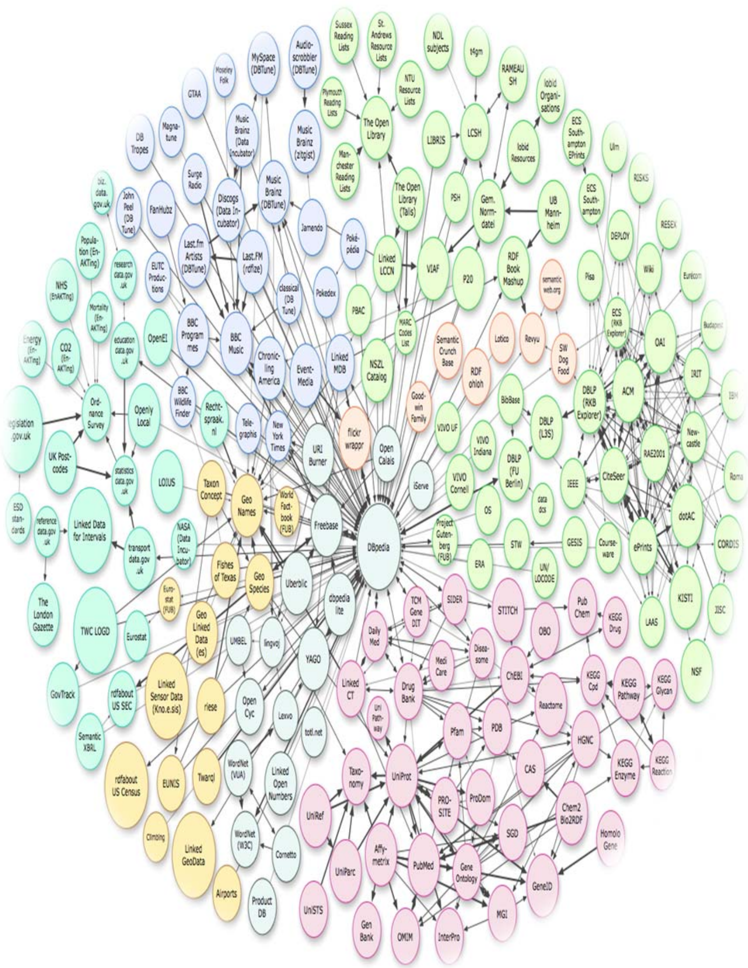
Evaluation

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

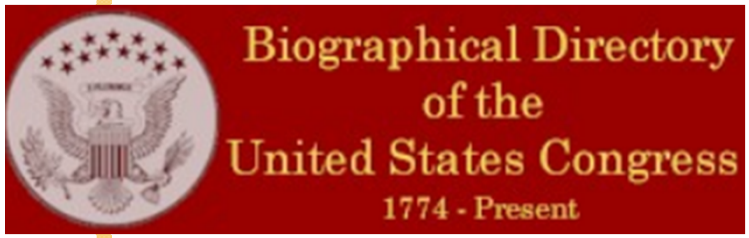
%	en de	en fr	en nl	de fr	de nl	fr nl	Overall
Precision	91.97	92.28	95.15	90.78	91.67	93.85	92.64
Recall	94.17	96.83	94.80	92.06	93.22	92.82	94.21
F_1 Score	93.06	94.50	94.97	91.42	92.44	93.33	93.42

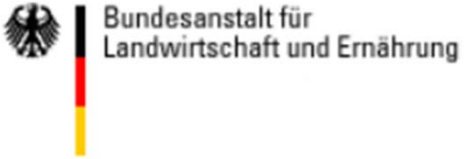
Overview

- 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

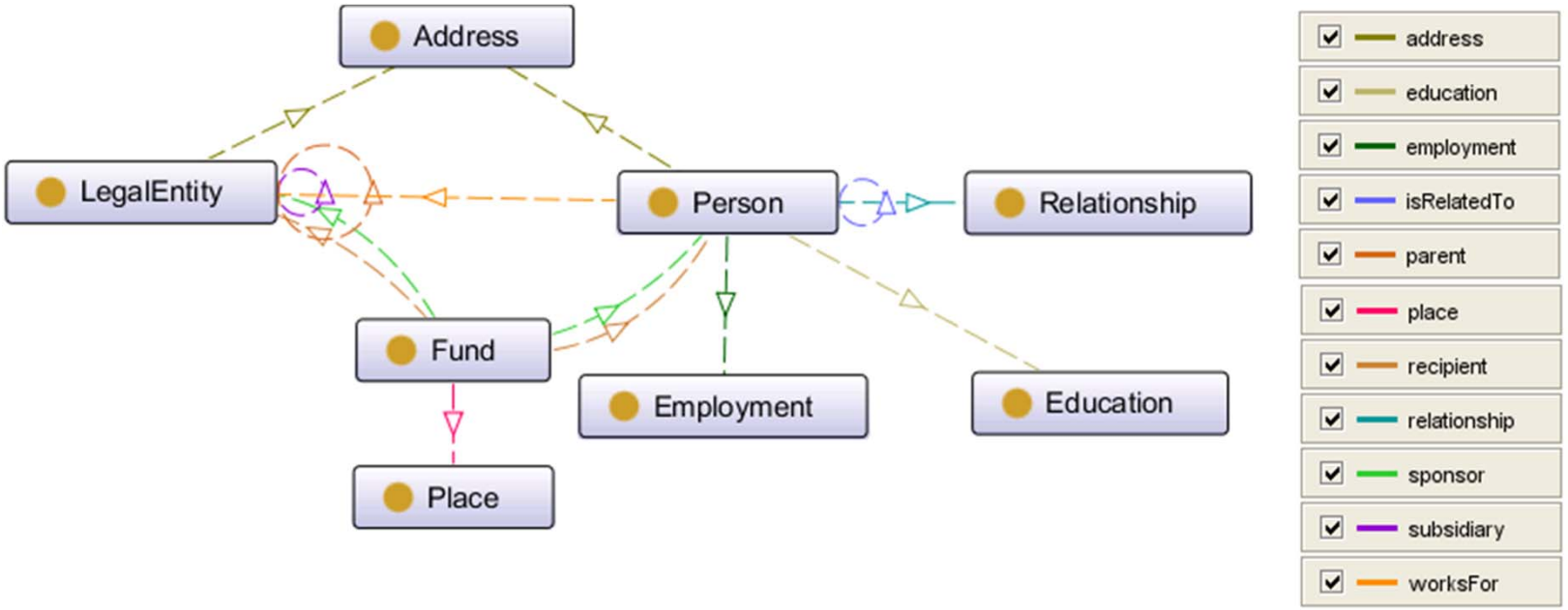


Motivation – Wealth of Open Gov Data



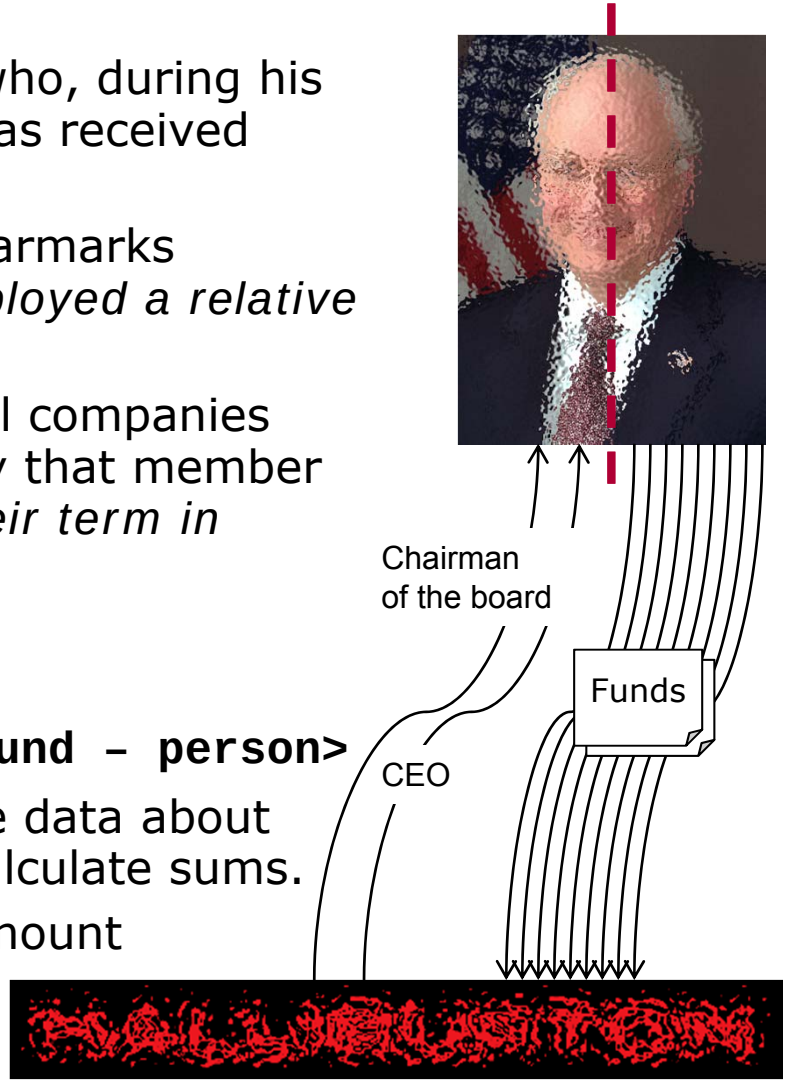


Companies, Agencies, and People



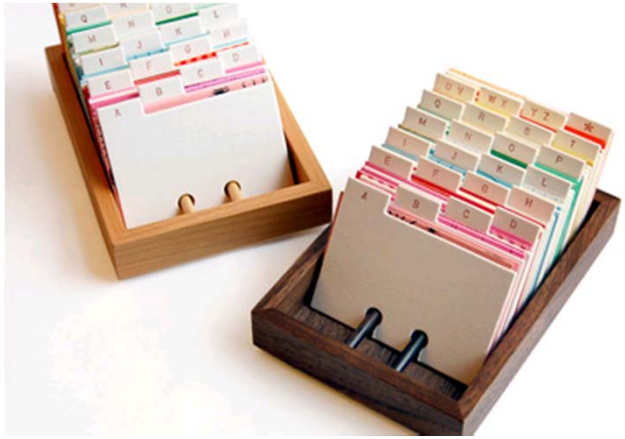
Interesting queries

- 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 …"



Five steps for integration

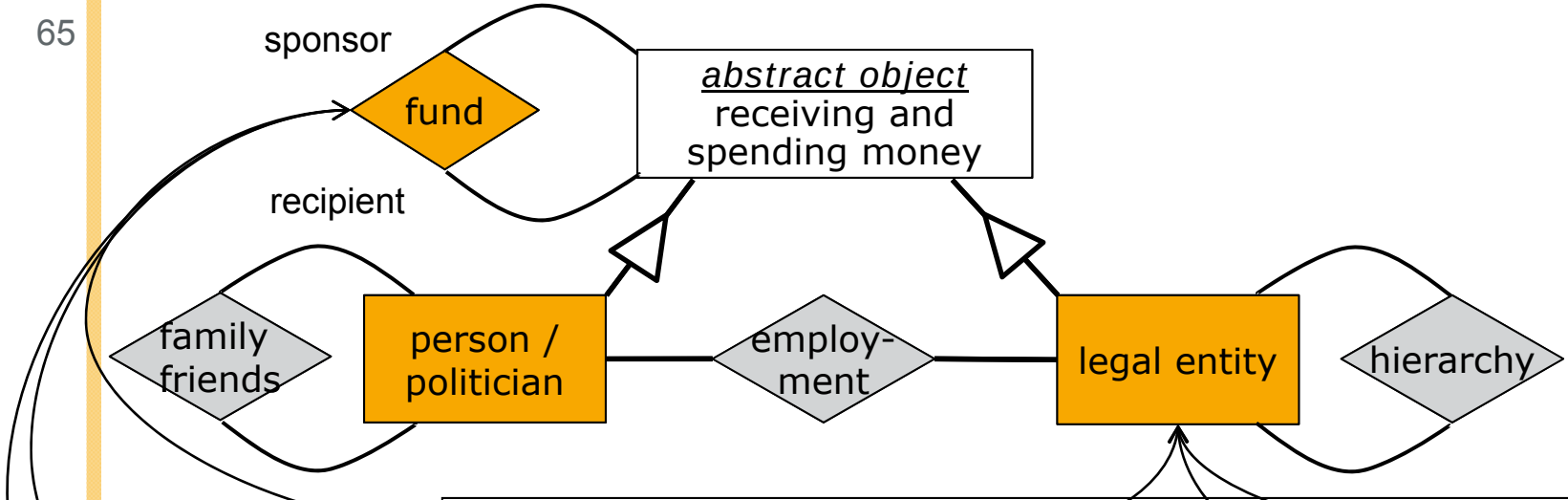
- Source Selection
- Schema Matching & Mapping
- Data Extraction & scrubbing
- Entity Matching
- Data Fusion



Data sources so far

Source	Num. of entities	Num. of attributes	Format	Content
US Spending	1.7m	122	XML	all gov spending
US Earmarks	20,000	37	CSV	anonymous garrantees
US Congress	12,000	8	HTML	members of congress since 1744, incl. bio
DE Party Donations	1,500	4	HTML	Donations > 20,000 €
EU Finance	122,000	11	HTML	EU spending
EU Agric. Subventions	207,000	8	HTML	EU spending
EU Parliam. Data	900	14	HTML	members of parliament
Freebase Person Data	1,8m	32	TSV	person data

Data – Mapping and Scrubbing



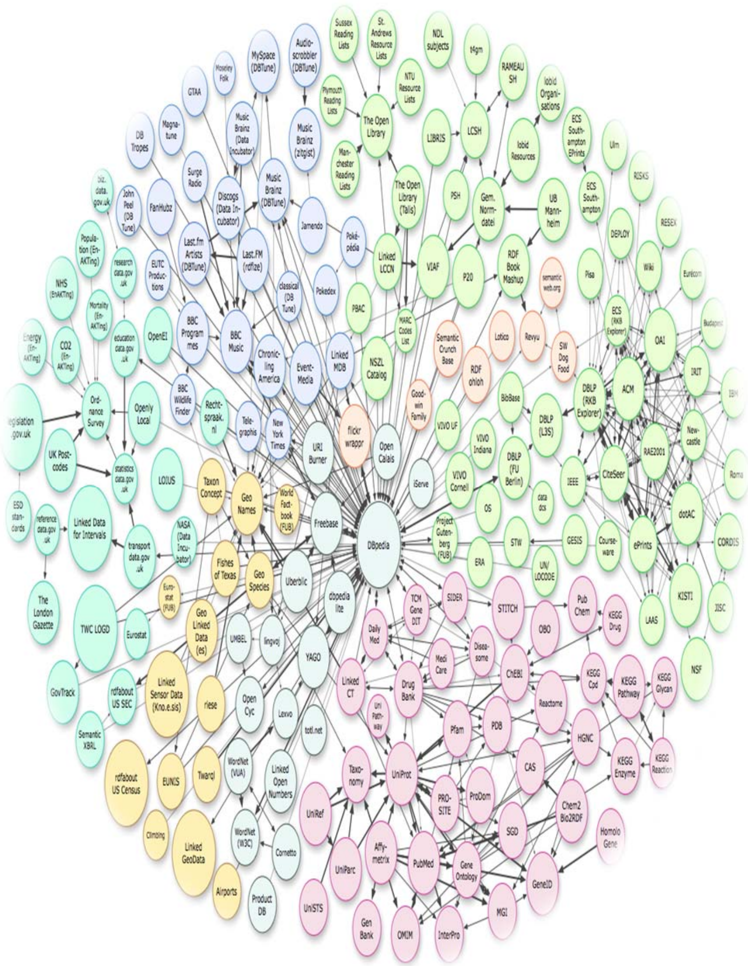

```
{ "_id" : "euFinance#28994",
  "year" : 2008,
  "nameOfBeneficiary" : "ROBERT BOSCH GMBH*",
  "coordinator" : false,
  "countryTerritory" : "Germany 70049 STUTTGART",
  "coFinancingRate" : "67,51 %",
  "amount" : 3199959.00,
  "commitmentPositionKey" : "F13.A22622.1",
  "subjectOfGrantOrContract" : "MULTISPECTAL TERAHERTZ, INFRARED ...
  "responsibleDepartment" : "Information Society and Media",
  "budgetLineNameAndNumber" : "Support for research ...
}
```

Data – Cleansing

- Deduplication / Entity Matching
  - Intra Source Consolidation
  - Intra Source Duplicate Detection
    - Duplicate Detection Toolkit – DuDe
    - Hundreds of duplicates within original sources
  - Entity Matching across Sources
    - Augment discovered Person Data with Freebase Info
    - Jaro-Winkler and Monge-Elkan distance
- Entity Fusion
    - Dempster-Shafer-Theory

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



http://govwild.org

- 200,000 persons
- 248,000 legal entities
- 1,000,000 funds

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

GovWILD

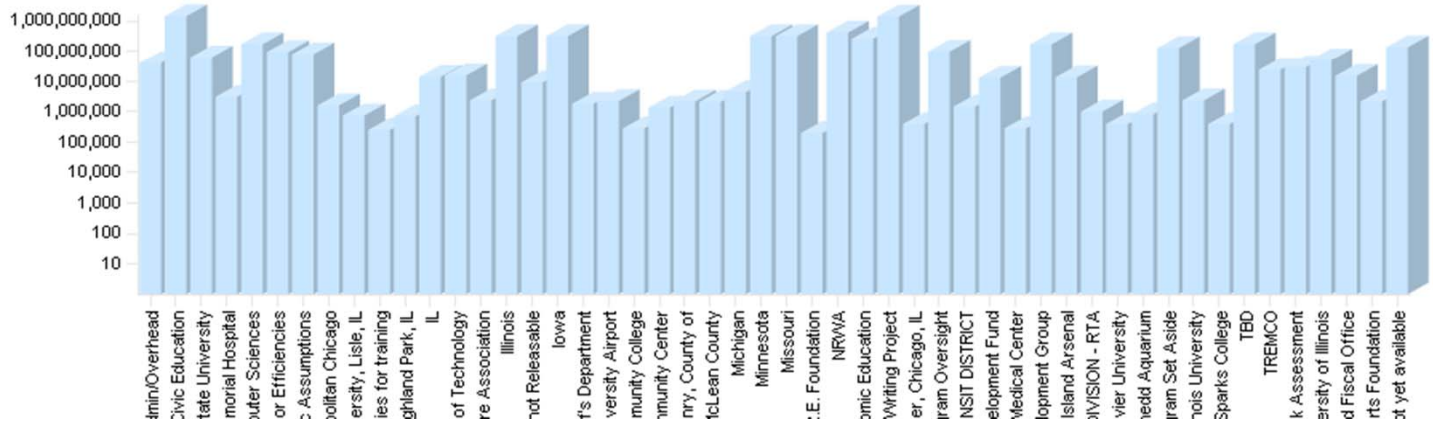
Barack Obama

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

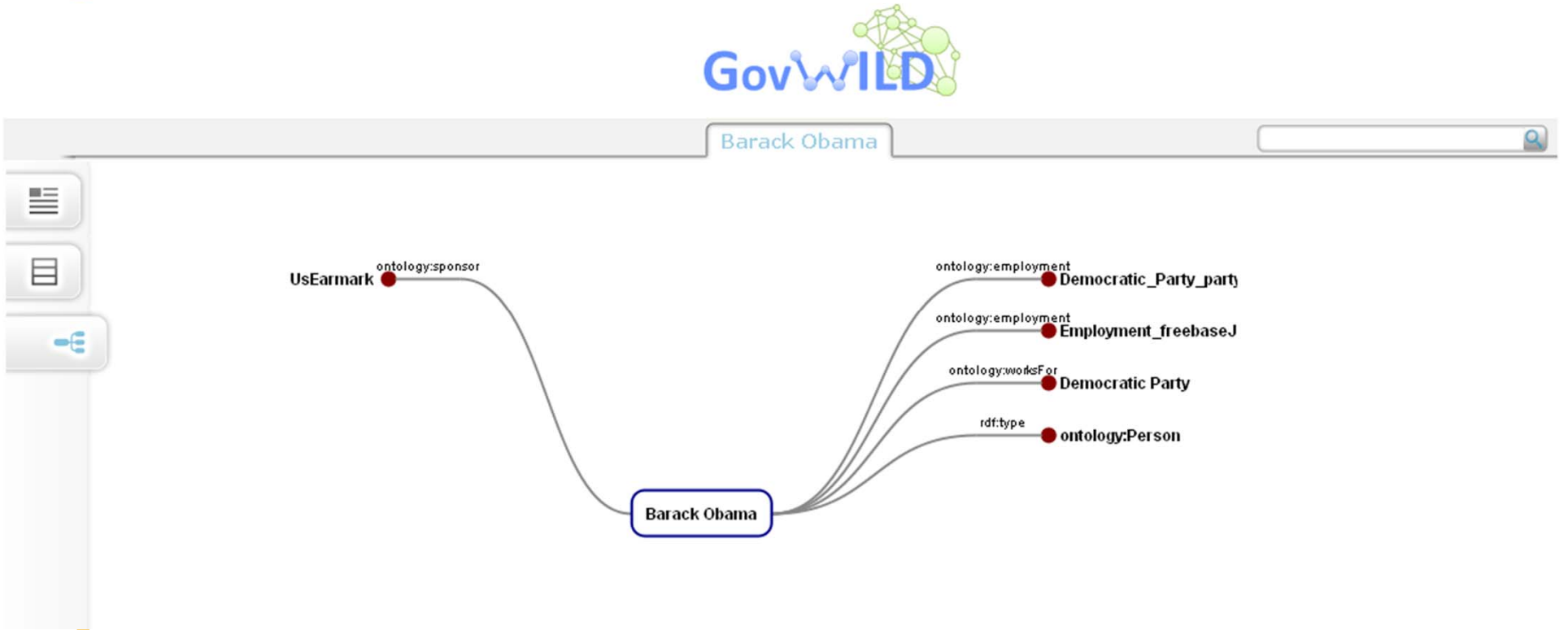
Earmarks



GovWILD

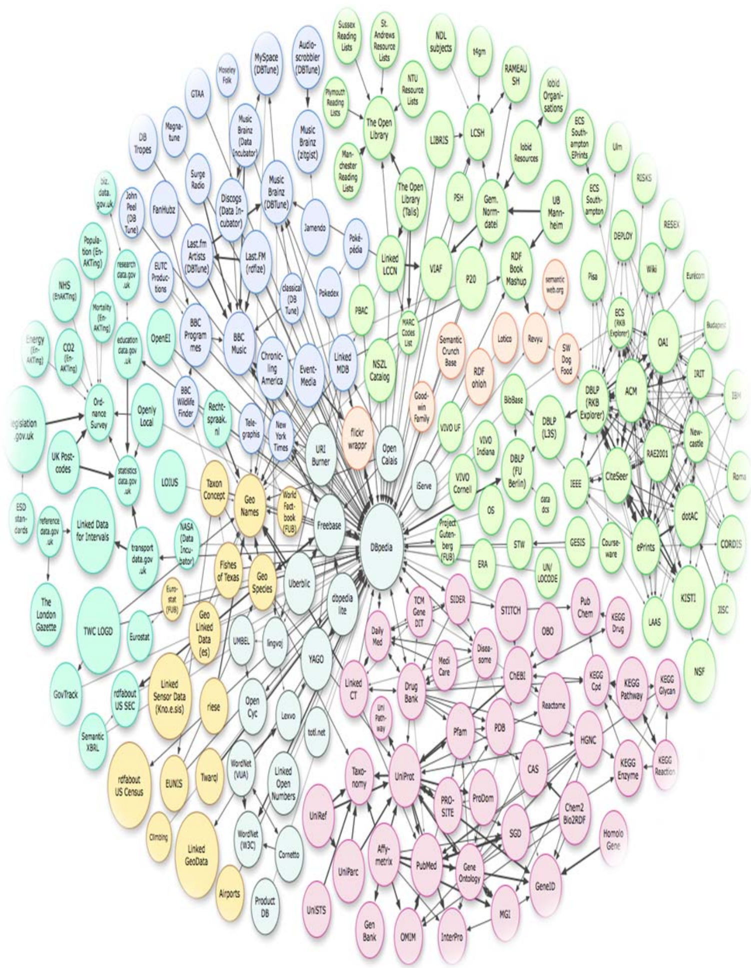
Barack Obama

Predicate	Value	Info
ontology:placeOfBirth	Honolulu, Hawaii	
ontology:birthYear	1961	
ontology:lastName	Obama	
rdf:type	ontology:Person	
ontology:employment	Democratic_Party_party_269	
ontology:employment	Employment_freebaseJoinedPersons_478_usStates_99	
rdfs:label	Barack Obama	
ontology:worksFor	Democratic Party	
ontology:birthDay	4	
ontology:birthMonth	8	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	
ontology:sponsor of	UsEarmark	


GovWILD
Barack Obama
UsEarmark
ontology:sponsor
Barack Obama
ontology:employment
Democratic_Party_party
ontology:employment
Employment_freebaseJ
ontology:worksFor
Democratic Party
rdf:type
ontology:Person

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