



**Hasso  
Plattner  
Institut**

IT Systems Engineering | Universität Potsdam

## Extreme Web Data Integration

August 14, 2010

Felix Naumann

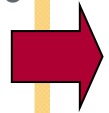
# Acknowledgements

2

- @IBM Almaden
  - Howard Ho, Mauricio Hernandez, Rajasekar Krishnamurthy, Lucian Popa, Roxana Stanoi
- @HPI
  - Christoph Böhm
  - Bachelor project student team
- And elsewhere
  - Antonio Sala (University of Modena)
  - Chis Bizer (dbpedia)
  - Open Data community

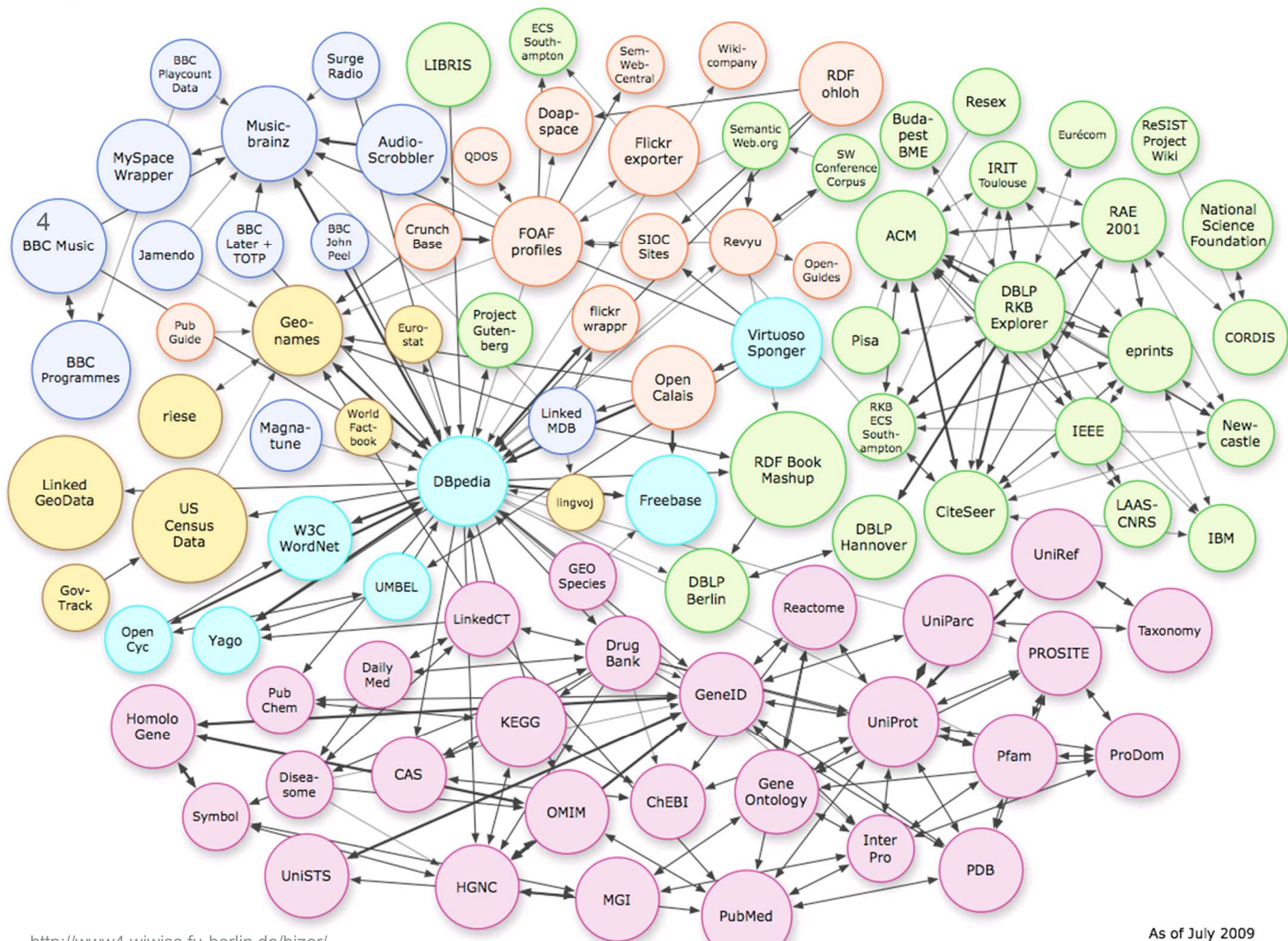
# Overview

3



- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds







# DBpedia - Extraction

5

```
{{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



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

# DBpedia statistics

6

## 1. Core Datasets

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

274 million triples

From English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish and Norwegian versions of Wikipedia

2.6 million things

213,000 persons

328,000 places

57,000 music albums

36,000 films

20,000 companies

<http://wiki.dbpedia.org/Datasets>

## And more sources

7

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

„Raw data now!“



# Use cases

8

- General purpose integration: Create rich knowledge bases
  - Semantic Web
  - Improved Search
  - Link creation
  - Cleansing
- Domain specific integration
  - Creation of high quality data sets: Complete & accurate
  - Enhancement of organization-internal data
  - Create reference data sets



# Two Flavors of Integration

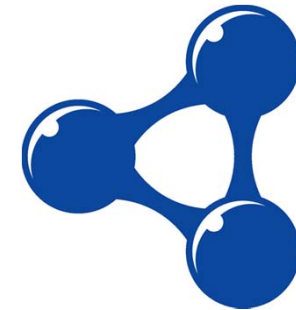
9

- Web Data Integration
  - Structured/semi-structured data
  - Exposed by different data providers
  - Steps: Source selection, data extraction from individual fields/attributes, scrubbing, entity matching, data transformation, data fusion
- Unstructured Entity Integration
  - Unstructured data
  - Small set of focused data sources.
  - Steps: Information extraction (from large text/html documents), Entity Resolution and Data Mapping / Fusion.

# Midas – Integration project with IBM Almaden Research Center

10

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



# Overview

11

- Web Data abounds
  - linked, open, and otherwise
- ➔ ■ Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds



# Challenges: Heterogeneity at all levels

12

## ■ Source

- |             |   |                     |
|-------------|---|---------------------|
| □ Formats   | ↔ | □ File converters   |
| □ Domain    | ↔ | □ Clustering, rules |
| □ Bandwidth | ↔ | □ Patience          |

## ■ Schema

- |             |   |                    |
|-------------|---|--------------------|
| □ Structure | ↔ | □ Schema Mapping   |
| □ Semantics | ↔ | □ Domain knowledge |

## ■ Data

- |              |   |                   |
|--------------|---|-------------------|
| □ Formatting | ↔ | □ Scrubbing       |
| □ Duplicates | ↔ | □ Entity Matching |

# The problem – a format mess

**Commitment position key: SI2.514875.1**

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

## Beneficiary

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

14

## What is a company?

**Def. 1:** Entities having a `companyName`

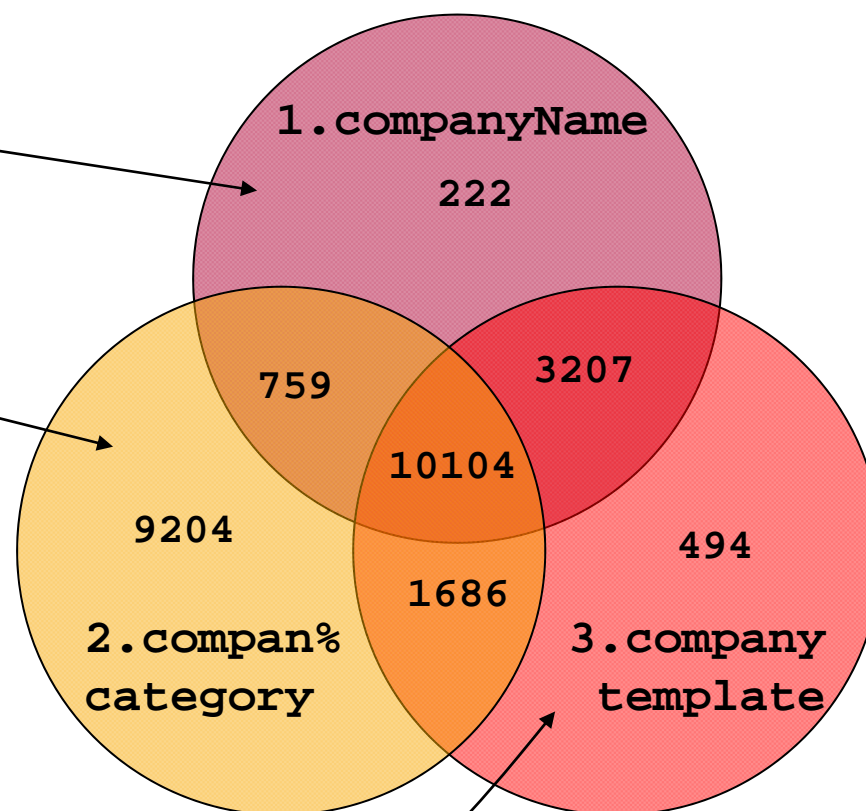
■ 14292 companies

**Def. 2:** Entities having a `category` that starts with `'compan%'`

■ 21753

**Def. 3:** Entities having a `wikiPageUsesTemplate` with value `Template:infobox_company`

■ 15491



# The problem – a schema mess

15

- Wikipedia/DBpedia: Triples and ill-defined templates invite disaster.
- Schema chaos: Many attribute synonyms
  - Hundreds of different attributes
  - **companyName** vs. **organizationName** vs. **name** vs. **company**
- 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

# Profiling Companies

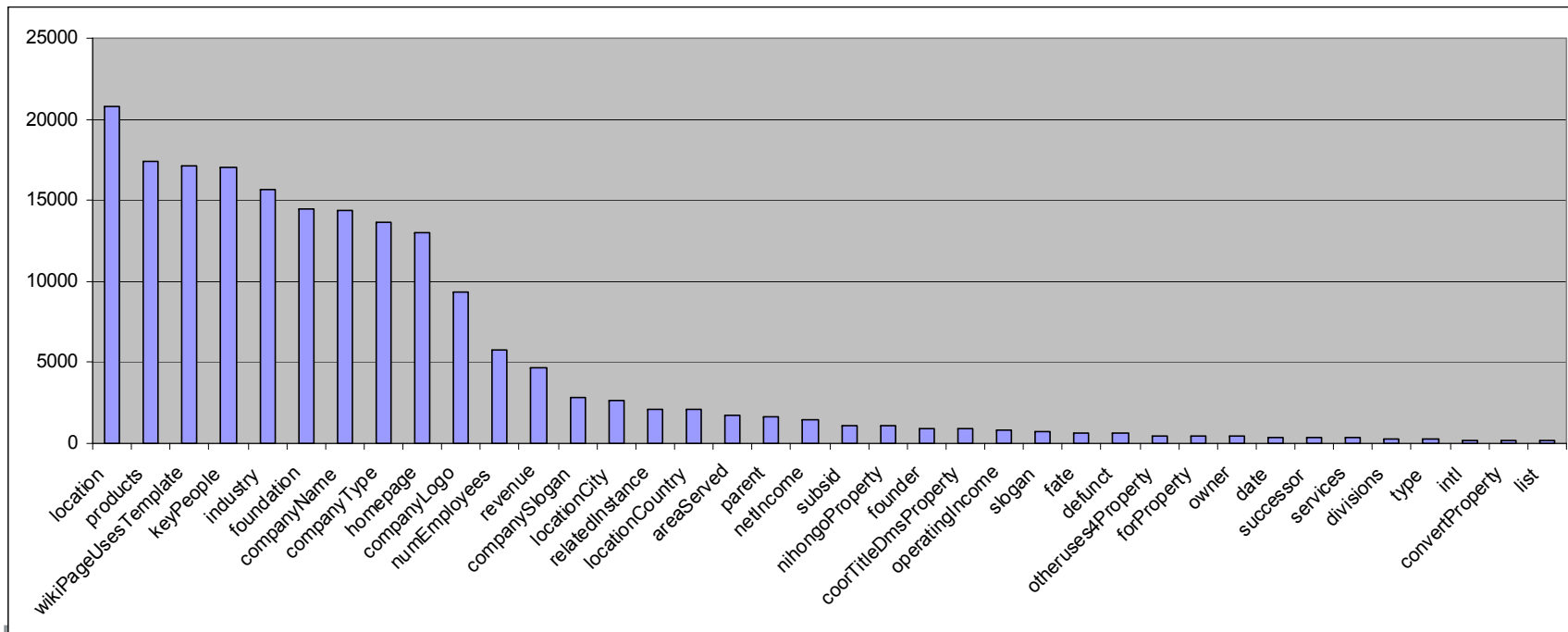
16

- Definition of companies?

- `SELECT DISTINCT TOPIC FROM DBPEDIA.INFOBOXES WHERE ATTRIBUTE = 'companyName'`

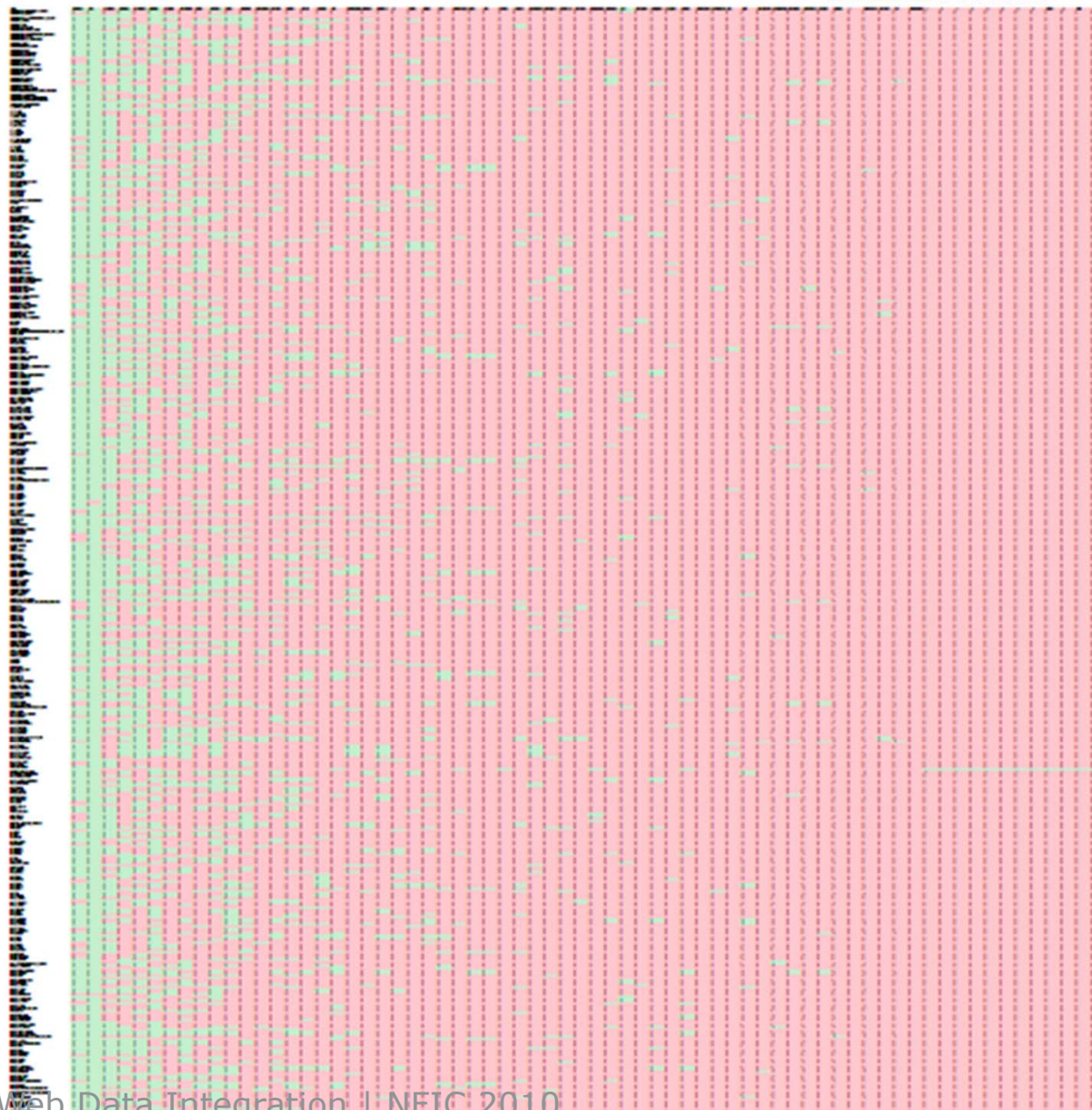
- Schema?

- `SELECT ATTRIBUTE, COUNT(*) AS SUM`
  - `FROM DBPEDIA.INFOBOXES`
  - `WHERE TOPIC IN`
  - `(SELECT DISTINCT TOPIC FROM DBPEDIA.INFOBOXES WHERE ATTRIBUTE = 'companyName')`
  - `GROUP BY ATTRIBUTE`
  - `ORDER BY SUM DESC;`



# Company attribute distribution

17



# Infoboxes with CompanyTemplate

18

- 1083 different attributes
  - H-index 56
  - 499 appear only once
- Of the 1083 attr., 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



# Company Template

19

```
{{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
<pre>{{Infobox Company   name           =   logo           =   type           =   genre          =   fate           =   predecessor    =   successor     =   foundation     =   founder        =   defunct       =   location_city  =   location_country =   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          = }}</pre>	<p>REQUIRED</p> <p>REQUIRED</p> <p>REQUIRED</p> <p>REQUIRED</p> <p>REQUIRED</p>

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 Purchase 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	Convenience and Economy
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

21

- 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:

?

Wal-Mart

Undisclosed

Assets exceed £4 billion GBP

[http://www.credit-suisse.com/investors/en/reports/2007\\_results\\_q4.jsp](http://www.credit-suisse.com/investors/en/reports/2007_results_q4.jsp)

Image:green\_up.png



€ bn (as of 2004)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	fieldName	example	type	info	usaS pendi ng	usEar mark s	usCon gress	euPar tyDon ations	euFi nanc e	euA grar	euPa rliam ent	Free bas e	usSec		
1	fieldName	example	type	info	usaS pendi ng	usEar mark s	usCon gress	euPar tyDon ations	euFi nanc e	euA grar	euPa rliam ent	Free bas e	usSec		
2	_id:	usEarmarks#123_P	string		not releva nt				rele vant?				not relevant		
3	firstName		string			x	x	x		x	x				
4	middleName		string			x	x	x		x	x				
5	lastName		string			x	x	x		x	x				
6	nameAddition	["Jr."]	[String]	Dictionary		x	x	x		x	x				
7	originals	[ "fileName#876867", ... ]	[string]	ID only		x									
8	addresses: [{		[record]												
9	street	"5th Avenue, 1"	string	other countries?							x				
10	additionalLine	"post office: 123"	string								x				
11	city	"New York"	string							x	x				
12	zipCode	"012G4"	string							x	x				
13	state	"New York"	string								x				
14	district	2	int												
15	country	"United States"	string	<a href="http://so.org/iso/english_country_names_and">so.org/iso/en glish_country _names_and</a>		US					x				
16	}]														
17	birth: {year: month: day:}	{year: 1950, month: 3, day: 12}	record with int				x				x				
18	death: {year: month: day:}	{year: 1950, month: 3, day: 12}	record with int				x								
19	gender	"male", "female"	string												
20	placeOfBirth		string				bio				x				
21	placeOfDeath		string				bio								
22	nationalities	[France]	[string]								x				
23	religion	http://www.vitobonsignor	string									x			
24	websites		[string]								x				
25	biography		text												
26	bibiliography		text				x				x				
27	ethnicity	White	string									x			
28	professions	["politician"]	[string]			politi cian	politici an				(politi cian)				

# Overview

23

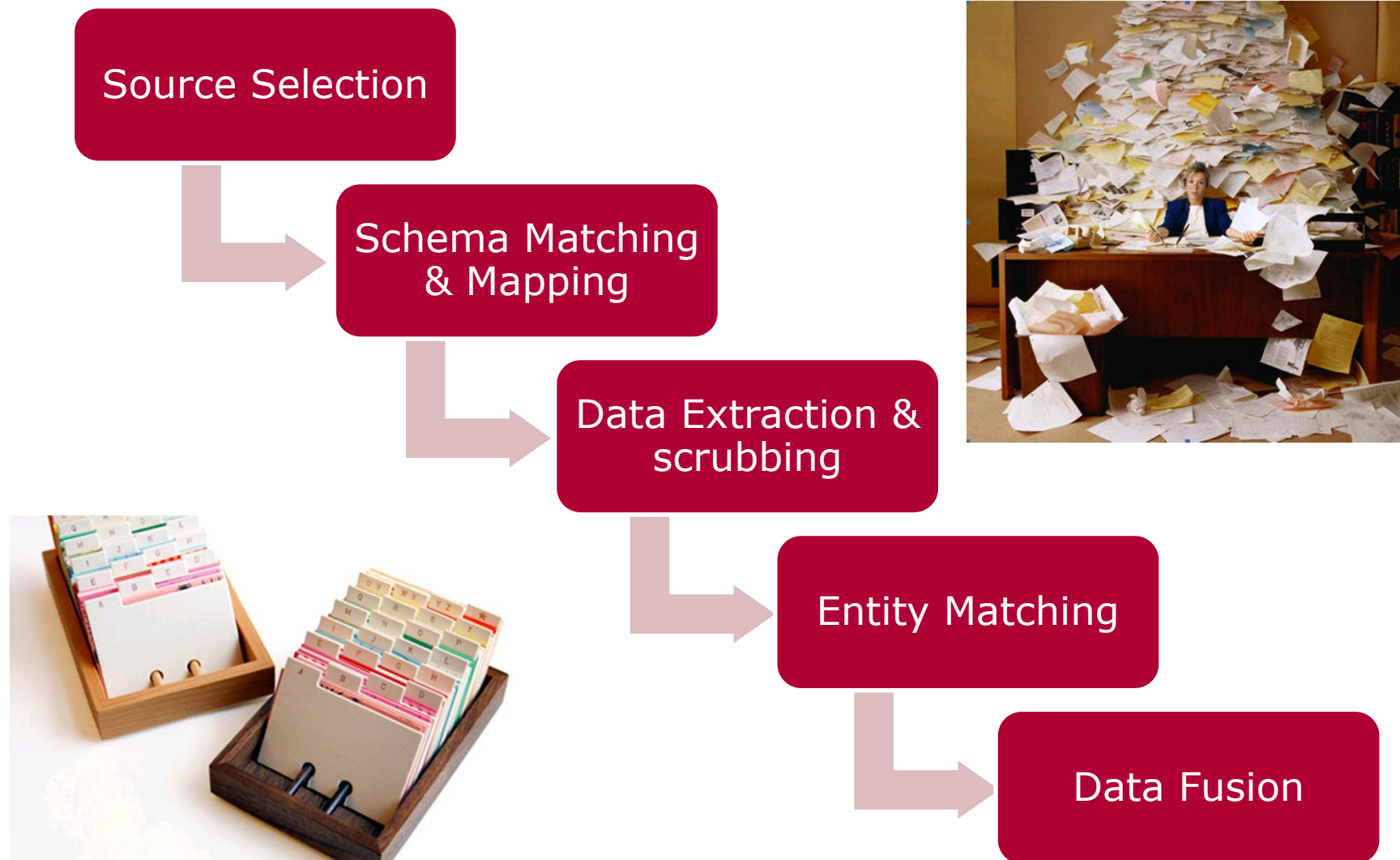
- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- ➔ ■ Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds





# Five steps for integration

24



# Five steps – Source selection

25

- 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)

---

• <a href="#">Alternative</a> (4,700)	• <a href="#">Medicine</a> (10,070)
• <a href="#">Conditions and Diseases</a> (14,289)	• <a href="#">Mental Health</a> (4,577)
• <a href="#">Healthcare Industry@</a> (5,652)	• <a href="#">Regional</a> (0)

---

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

dmoz.org

# Five steps – Schema matching and schema mapping

26

- 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
locationCity	locationCity, location	BusinessAddress, MailingAddress	/location/ mailing_address/pos
locationCountry	locationCountry, location, showflag	BusinessAddress, MailingAddress	/location/ mailing_address/city
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

27

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

## Five steps – Entity matching

28

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

**Find People**Find People | [Find a Business](#)

First Name	*Last Name	City, State or ZIP	<b>Find</b>
<input type="text" value="Michael"/>	<input type="text" value="Jordan"/>	<input type="text" value="CA"/>	


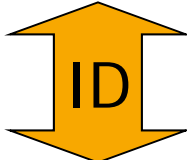

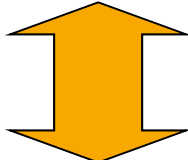




**Whoa!** Over 100 Results Found



# Five steps – Data fusion

29

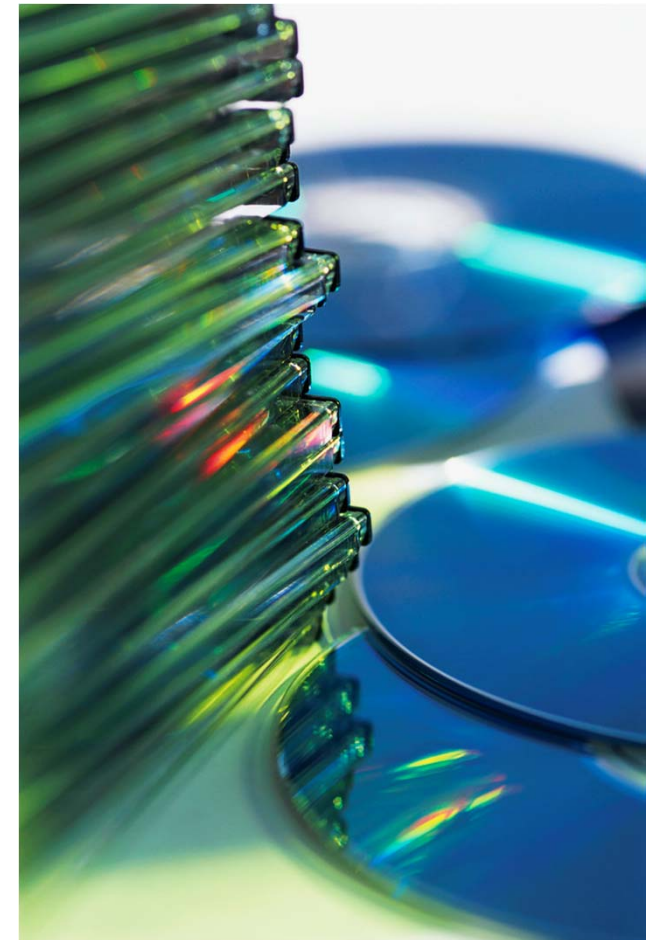
- Combine multiple representations of real-world entities
  - Survivorship, consolidation, etc.
- Resolve data conflicts
  - Conflict resolution functions
  - Reputation / accuracy / freshness -> “truth discovery”
- Retain data lineage

0766607194	H. Melville		\$3.98	
				
0766607194	Herman Melville	Moby Dick	\$5.99	 

# Overview

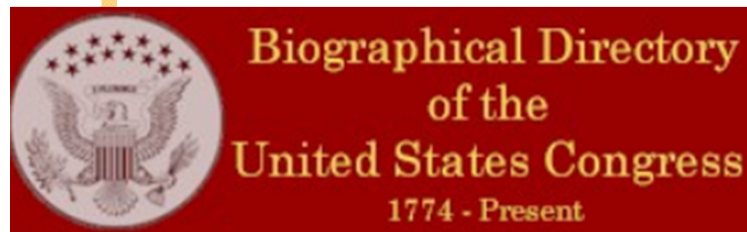
30

- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- ➔ ■ The GovWILD experience
  - politicians, friends, and funds



# Motivation – Wealth of Open Gov Data

31



## Interesting queries

32

- Find all *classmates* of George W. Bush who, during Bush's 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 member of congress, find all companies that have received funding supported by that member and have *employed him/her after their term in congress*.
- Goal: Demonstrate the power of
  - *Sets*: Instead of researching individuals, write queries against large sets of persons
  - *Joins*: Make unknown connections, for instance connecting persons through their universities or connecting persons with companies in multiple ways (employment and funding)
  - *Grouping and aggregation*: Combine information about parties, companies, and persons and find averages and sums.
  - *Sorting*: Order results by funding amount to find top results.

## Data sources so far

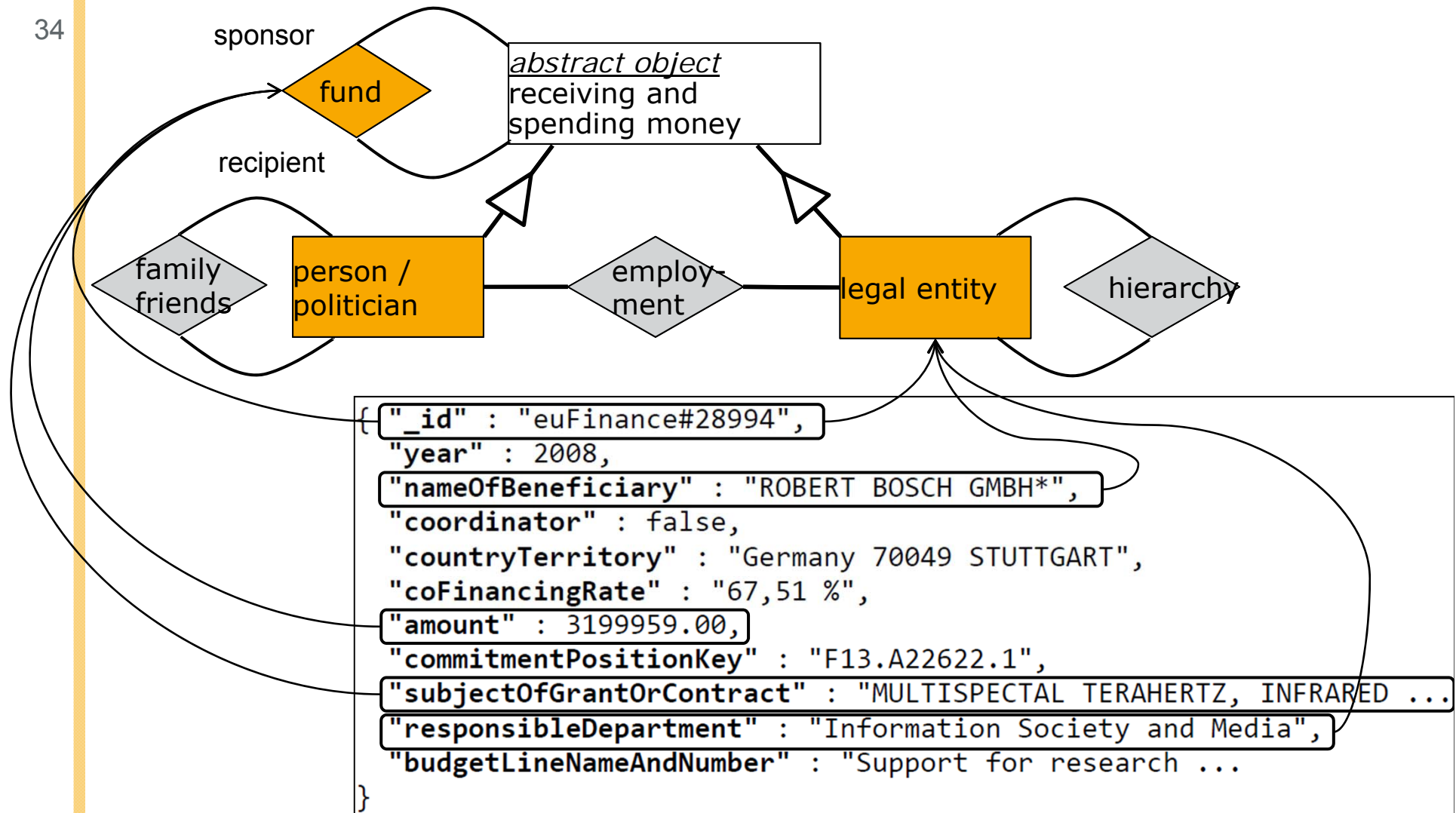
33

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

34



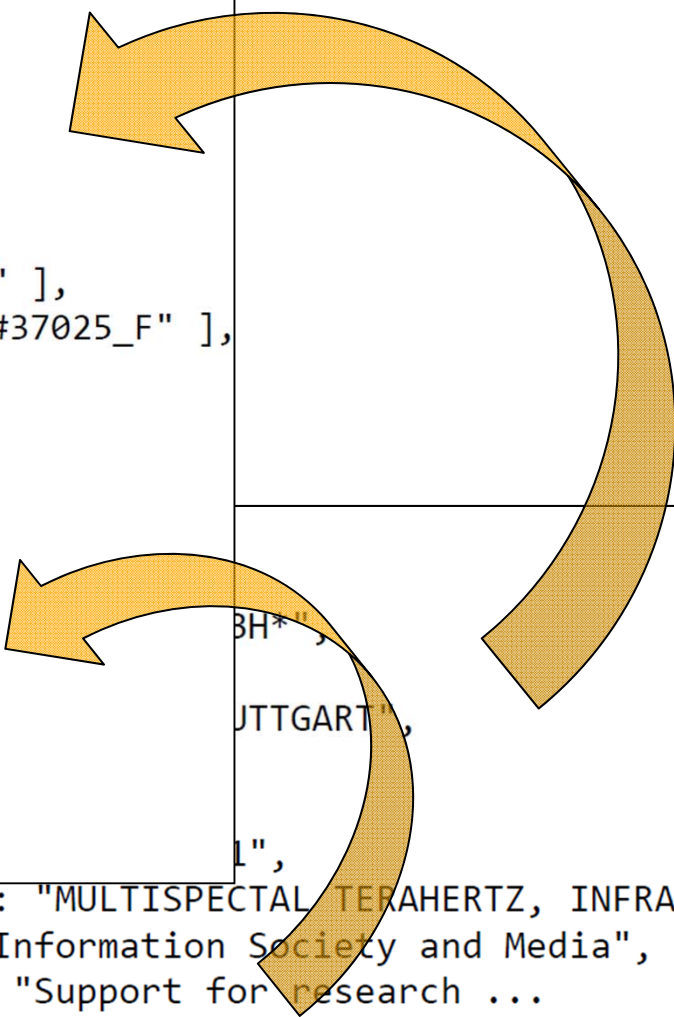
# Data - Transformation

```

legal_entity: {
  "_id": "euFinance#28994_L1",
  "addresses": [
    { "country": "Germany",
      "zipCode": "70049",
      "city" : "Stuttgart" } ],
  "name": "Robert Bosch",
  "originals": [ "euFinance#28994", "euFinance#37025" ],
  "receivedFunds": [ "euFinance#28994_F", "euFinance#37025_F" ],
  "type": { "form": "GmbH",
            "category": "company" }
}
fund: {
  "_id": "euFinance#28994_F",
  "amount": 3199959,
  "currency": "EUR",
  "date": { "year": 2008 },
  "originals": [ "euFinance#28994" ],
  "recipients": [ "euFinance#28994_L1" ],
  "sponsors": [ "euFinance#42090_L2" ]
}

"subjectOfGrantOrContract" : "MULTISPECTAL TERAHERTZ, INFRARED ...
"responsibleDepartment" : "Information Society and Media",
"budgetLineNameAndNumber" : "Support for research ..."

```




# Data – Cleansing

36

- Deduplication
  - 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

<http://govwild.org>

37

- 200,000 persons
  - 248,000 legal entities
  - 1,000,000 funds
- 
- The image shows the GovWILD logo in a stylized, golden, serif font. Below the logo is a white rectangular search input field with a thin grey border. To the right of the input field is a black rectangular button with the word "Search" in white text.
- Keyword Queries
  - Linked Data Interface (dereference URIs)
  - Exploration of entities mentioned in New York Times articles
  - Data Download (RDF, SQL Dump, JSON files)

# Summary

38

- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds

