Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Linked Data Profiling Introduction

22.4.2009

Felix Naumann

Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff

Information systems team

Paul **Führing**

project **ViQTOR**

DQ Annotation & Assessment

Katrin **Heinrich**

Prof. Felix **Naumann**

Information Integration

Jens **Bleiholder**

Data Fusion

project **fusem**

project **HumMer**

Christoph Böhm

Data Profiling & Cleaning

Information Quality

Armin **Roth**

project **System P**

Peer Data Management Systems

Matching

Service-Oriented Systems

Mohammed **AbuJarour**

Ontologies, Profiling

Frank **Kaufer**

Data Integration for Life Science Data Sources

project **Aladin**

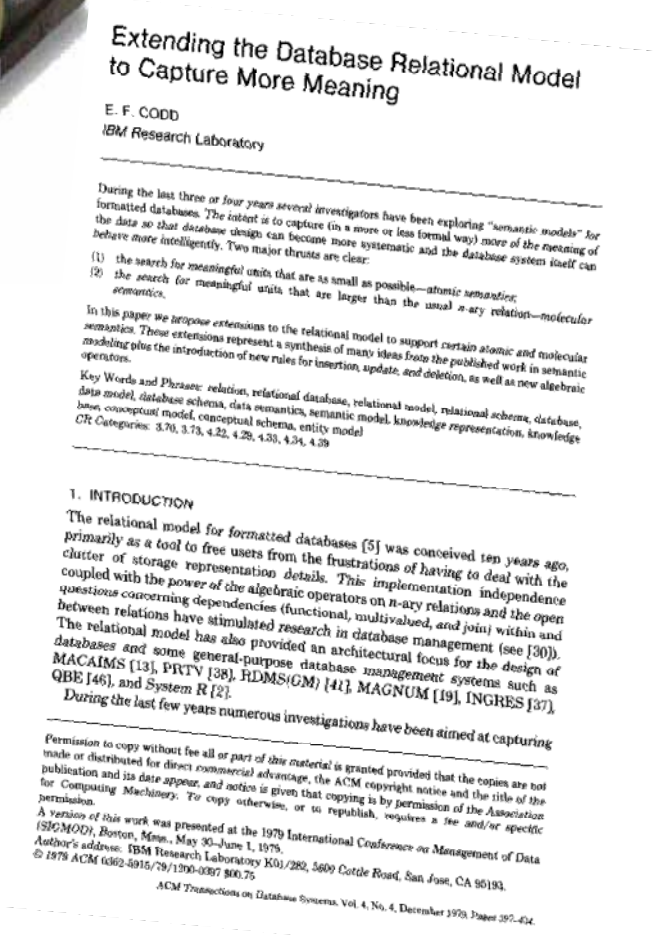
Jana **Bauckmann**

Alexander **Albrecht**

ETL Management

Data Profiling for Schema Management

Other courses in this semester



Extending the Database Relational Model to Capture More Meaning

E. F. CODD

IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

(1) the search for meaningful units that are as small as possible—*atomic semantics*;
(2) the search for meaningful units that are larger than the usual *n*-ary relation—*molecular semantics*.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of new rules for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model
CR Categories: 3.70, 3.73, 4.22, 4.29, 4.33, 4.34, 4.39

1. INTRODUCTION

The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the clutter of storage representation details. This implementation independence coupled with the power of the algebraic operators on *n*-ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS/(GM) [41], MAGNUM [19], INGRES [37], QBE [46], and System R [2].

During the last few years numerous investigations have been aimed at capturing

A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.
Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.

Lectures

- DBS I
- Search engines

Seminars

- Bachelor: Beauty is our Business
- Bachelor: Map/Reduce Algorithms on Hadoop
- Master: Linked Data Profiling
- Forschungsseminar

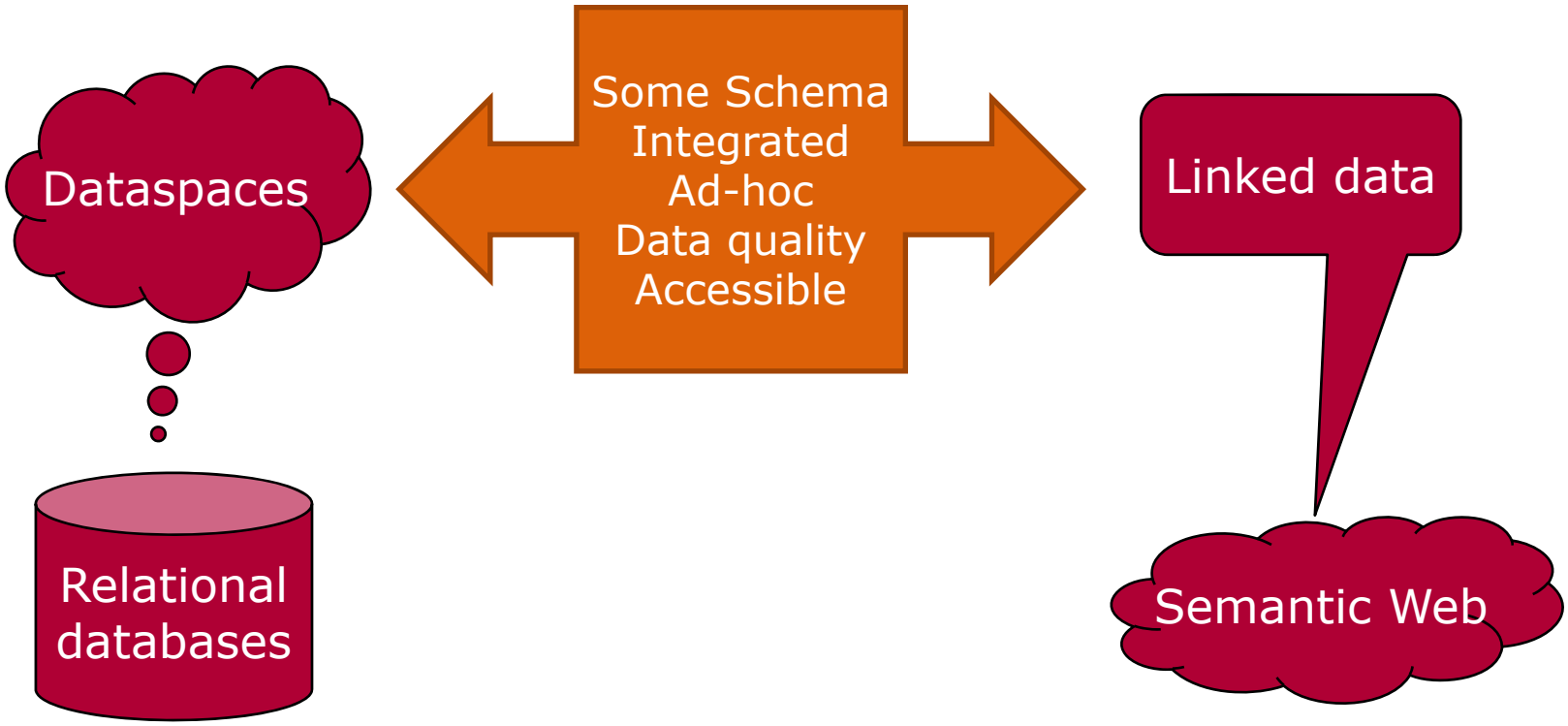
Bachelorproject

- ETL Management

Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff

Linked Data


Dataspaces
Relational databases
Some Schema
Integrated
Ad-hoc
Data quality
Accessible
Linked data
Semantic Web

Linked data – four principles

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

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

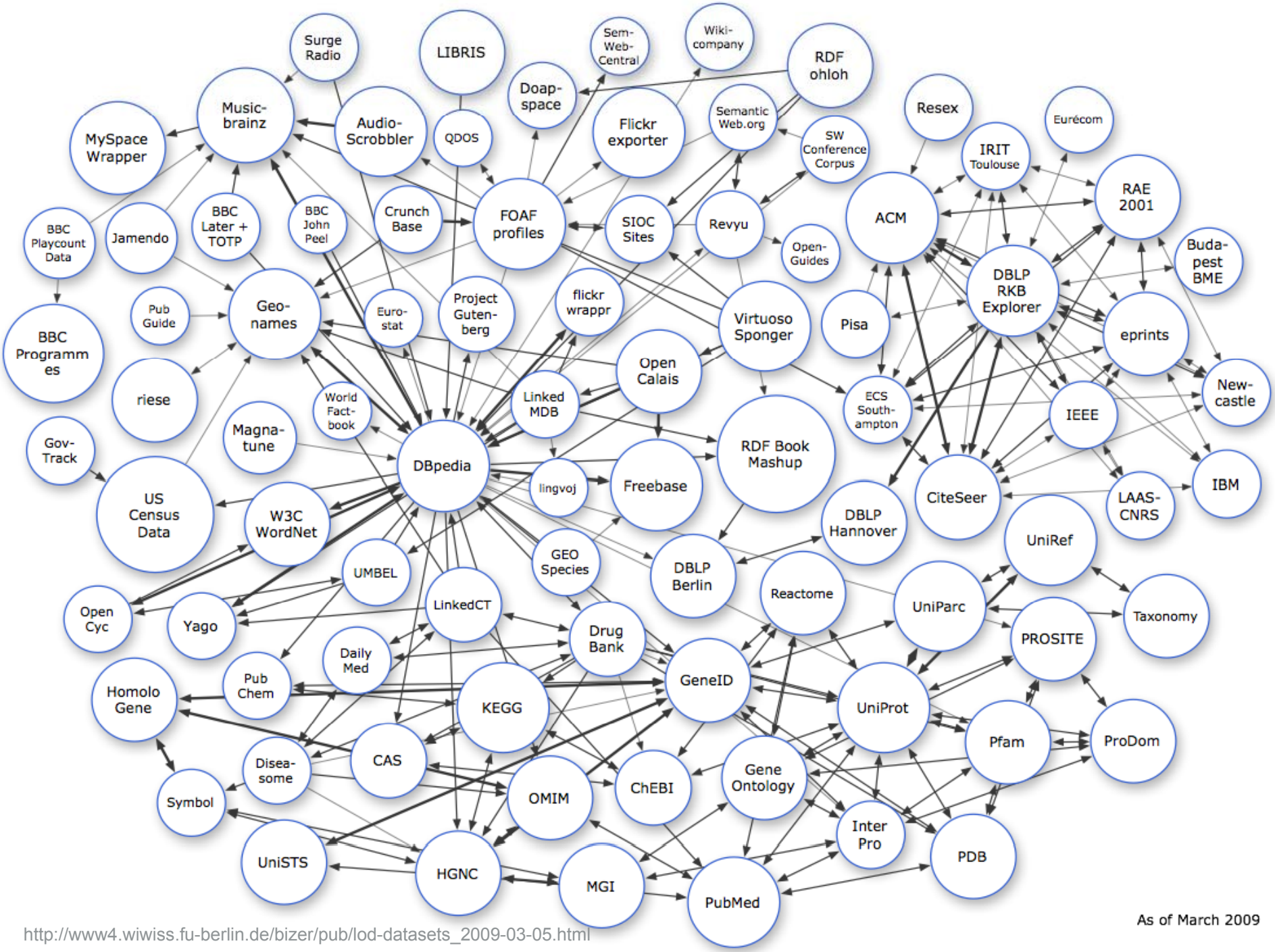
Linked Data Characteristics

- Linked Data comes as triples
  - Subject, predicate, object
  - URI, property, value
- Linked data is often user generated
- Linked data is free

- Linked data is heterogeneous
- Linked data is dirty

Suggested Reading

- Linked Data by Tim Berners-Lee
  - http://www.w3.org/DesignIssues/LinkedData.html
- Interlinking Open Data on the Web by Chris Bizer, Tom Heath, Danny Ayers, and Yves Raimond
  - ESWC 2007
  - http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkingOpenData.pdf
- And many more at
  - http://en.wikipedia.org/wiki/Linked_data
  - http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/
  - http://linkeddata.org/



http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-03-05.html

TED talk by Berners-Lee

- http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html
  - 15 min

DBPedia - Extraction


```
{{Infobox Town AT |
  name = Innsbruck |
  image_coa =  InnsbruckWappen.png |
  image_map = Karte-tirol-I.png |
  state = [[Tyrol]] |
  regbzk = [[Statutory city]] |
  population = 117,342 |
  population_as_of = 2006 |
  pop_dens = 1,119 |
  area = 104.91 |
  elevation = 574 |
  lat_deg = 47 |
  lat_min = 16 |
  lat_hem = N |
  lon_deg = 11 |
  lon_min = 23 |
  lon_hem = E |
  postal_code = 6010-6080 |
  area_code = 0512 |
  licence = I |
  mayor = Hilde Zach |
  website = [http://innsbruck.at] |
}}
```

Innsbruck

Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km²
Population density	1,119 /km²
Elevation	574 m
Coordinates	47°16' N 11°23' E
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at

Infoboxes

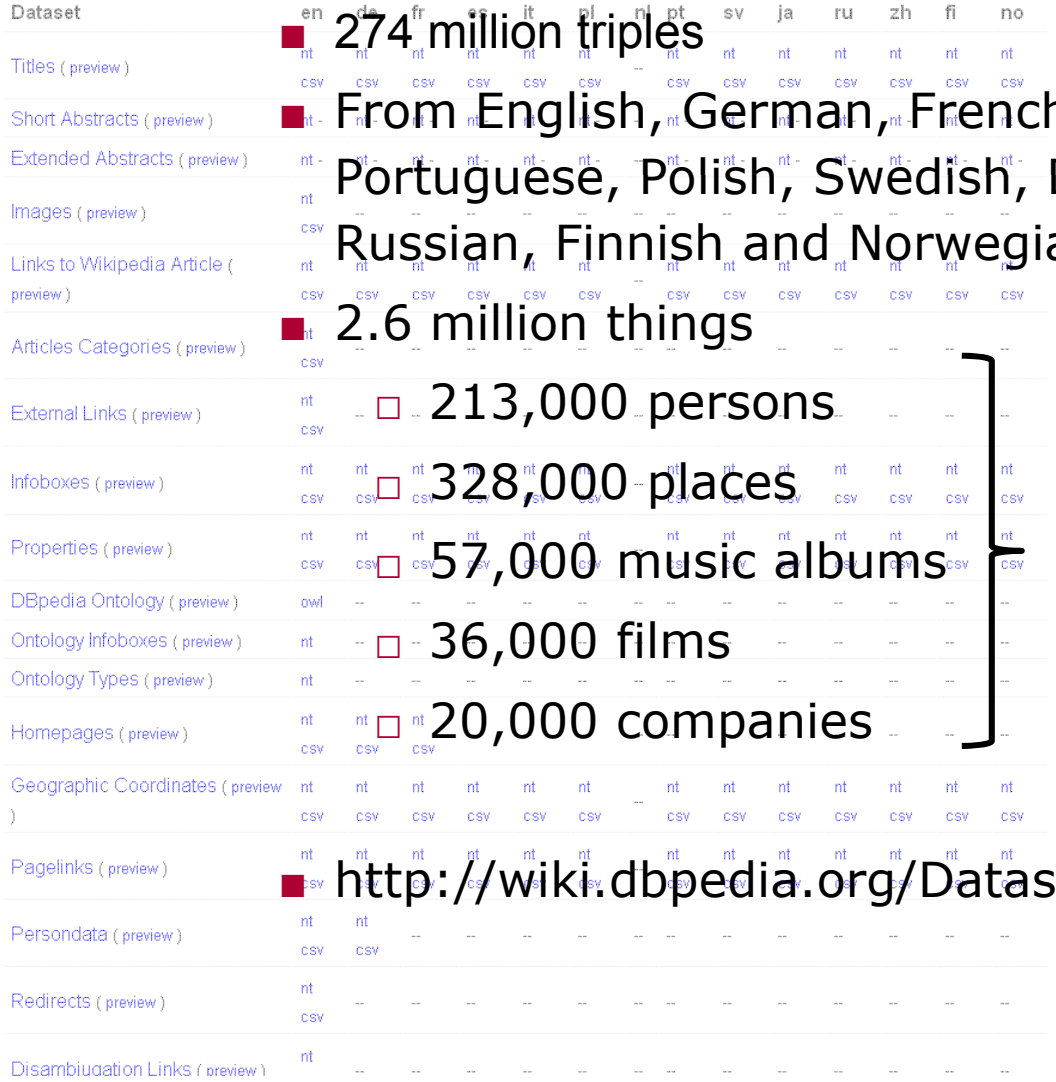
Type	Public (NYSE: IBM)
Founded	Endicott, New York, U.S. (1889, incorporated 1911)
Headquarters	Armonk, New York, USA
Key people	Samuel J. Palmisano, Chairman, President & CEO Mark Loughridge, SVP & CFO Dan Fortin, President (Canada) Jason Oppenheim, President (Asia Pacific) Nick Donofrio, EVP (Innovation & Technology) Mike Rhodin, President IOT Northeast Europe Dominique Cerutti, President IOT Southwest Europe[1]

Industry	Computer hardware Computer software Consulting IT Services
Products	See products listing
Revenue	▲ $98.8 billion USD (+4% FY '07 to '08)
Net income	▲ $10.4 billion USD (+11% FY '07 to '08)[2]
Employees	386,558 (2007)
Subsidiaries	ADSTAR FileNet Informix Iris Associates Lotus Software Rational Software Sequent Computer Systems Telelogic Tivoli Systems, Inc.
Website	www.ibm.com

Categories: Companies listed on the New York Stock Exchange | Dow Jones Industrial Average | IBM | Companies established in 1888 | Companies based in Westchester County, New York | Electronics companies of the United States | Point of sale companies | Semiconductor companies | Computer storage companies | Computer companies of the United States | Computer hardware companies | Software companies of the United States | UML Partners | Multinational companies

DBPedia statistics

1. Core Datasets



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

- 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

How do we know this?

- http://wiki.dbpedia.org/Datasets

Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff

Titles	DBPResource	"rdfsLabel"	WikiPTitle
ShortAbstracts	DBPResource	"rdfsComment"	Abstract(Text)
ExtendedAbstracts	DBPResource	"dbpsAbstract"	Abstract (Text)
Images	DBPResource	"DCrights"; "FOAFdepiction"; "FOAFimg:	Abstract (Text)
LinksToWikiP	DBPResource	"FOAFpage"	WikiP page
ArticlesCategories	DBPResource	"skosSubject"	Category
ExternalLinks	DBPResource	"reference"	URL
InfoBoxes	DBPResource	"name",… (39343 different)	Value
Properties	DBPUrl	"rdfsType"; "rdfsLabel"	rdfsProperty; label
Homepages	DBPResource	"FOAFHomepage"	URL
GeographicCoordinates	DBPResource	8 different from 3 namespaces	Number;URL
Pagelinks	DBPResource	"wikilink"	DBPResource
Persondata	DBPResource	"givenname", "name", "surname", "type" plus 3 others very rare	value
Redirects	DBPResource (almost key, long tail)	"redirect"	DBPResource
DisambiguationLinks	DBPResource	"disambiguates"	DBPResource (almost key, long tail)
WordnetClasses	DBPResource	"wordnettype"	WordnetURL
CategoriesLabels	"Category:"+DBPResource	"rdfsLabel"	Category label (name, also unique)
CategoriesSkos	"Category:"+DBPResource	"skos-broader"; "skos-prefLabel"; "rdfsType"	"Category:"+DBPResource; Categorylabe

Almost 1:1, WikiPTitle has approx. 100 duplicates:
(SELECT OBJECT, COUNT(*) AS SUM FROM DBPEDIA.TITLES GROUP BY OBJECT ORDER BY SUM DESC;)

Exactly 1:1; Table is redundant:
SELECT * FROM DBPEDIA.LINKSTOWIKIP
WHERE SUBJECT <> SUBSTR(OBJECT,30);

320007 different categories; average of 23 topics per category; "living_people" most frequent (250000)

Plattner Institut

Some Queries – Pivoting Persons

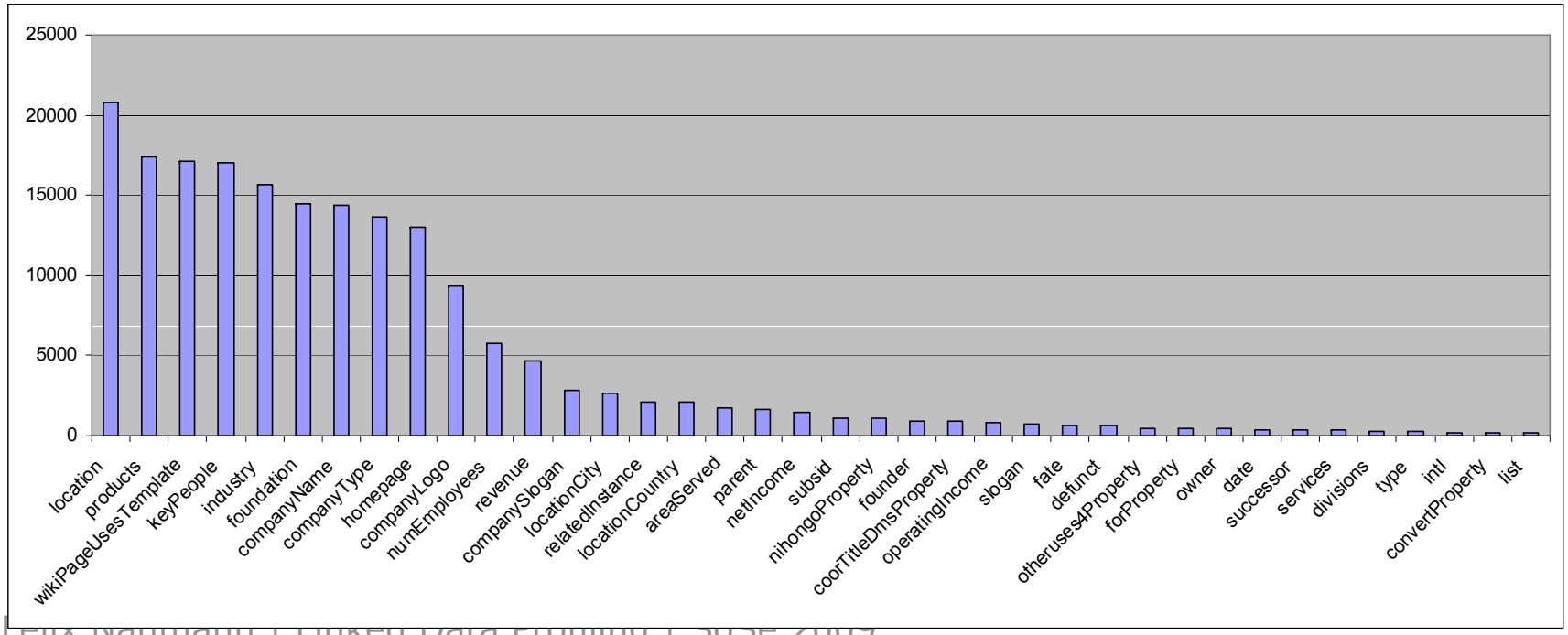

```
SELECT DISTINCT P1.PERSONID AS PERSONID, P2.VALUE AS NAME,
P3.VALUE AS GIVENNAME, p4.VALUE AS SURNAME
FROM DBPEDIA.PERSONDATA AS P1, DBPEDIA.PERSONDATA AS P2,
DBPEDIA.PERSONDATA AS P3, DBPEDIA.PERSONDATA AS P4
WHERE P1.PERSONID = P2.PERSONID
AND P2.PERSONID = P3.PERSONID
AND P3.PERSONID = P4.PERSONID
AND P2.ATTRIBUTE = 'http://xmlns.com/foaf/0.1/name‘
AND P3.ATTRIBUTE = 'http://xmlns.com/foaf/0.1/givenname‘
AND P4.ATTRIBUTE = 'http://xmlns.com/foaf/0.1/surname');
```

Some Queries – Profiling Persons

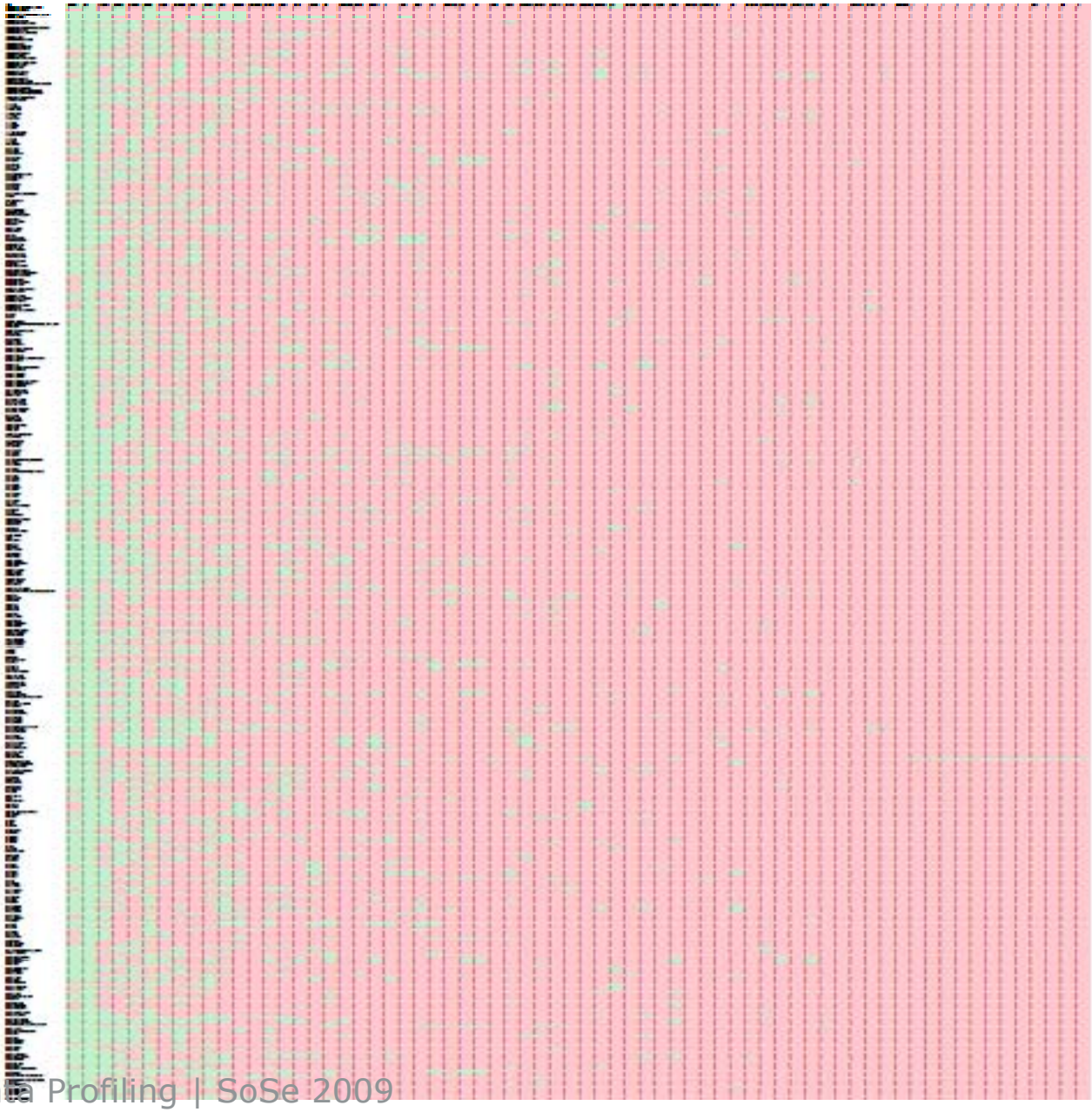

```
SELECT *
FROM  DBPEDIA.PIVOTPERSONDATA
WHERE NAME <> GIVENNAME || ' ' || SURNAME;

SELECT COUNT(*)
FROM  DBPEDIA.PIVOTPERSONDATA
WHERE NAME = '';

SELECT *
FROM  DBPEDIA.PIVOTPERSONDATA AS P1, DBPEDIA.PIVOTPERSONDATA AS P2
WHERE ARESTRINGSSIMILAR(P1.SURNAME, P2.SURNAME, 0.1) = 1
AND ARESTRINGSSIMILAR(P1.GIVENNAME, P2.GIVENNAME, 0.1) = 1
AND   P1.PERSONID <> P2.PERSONID
ORDER BY P1.surname, P1.GIVENNAME
```

Companies

- 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

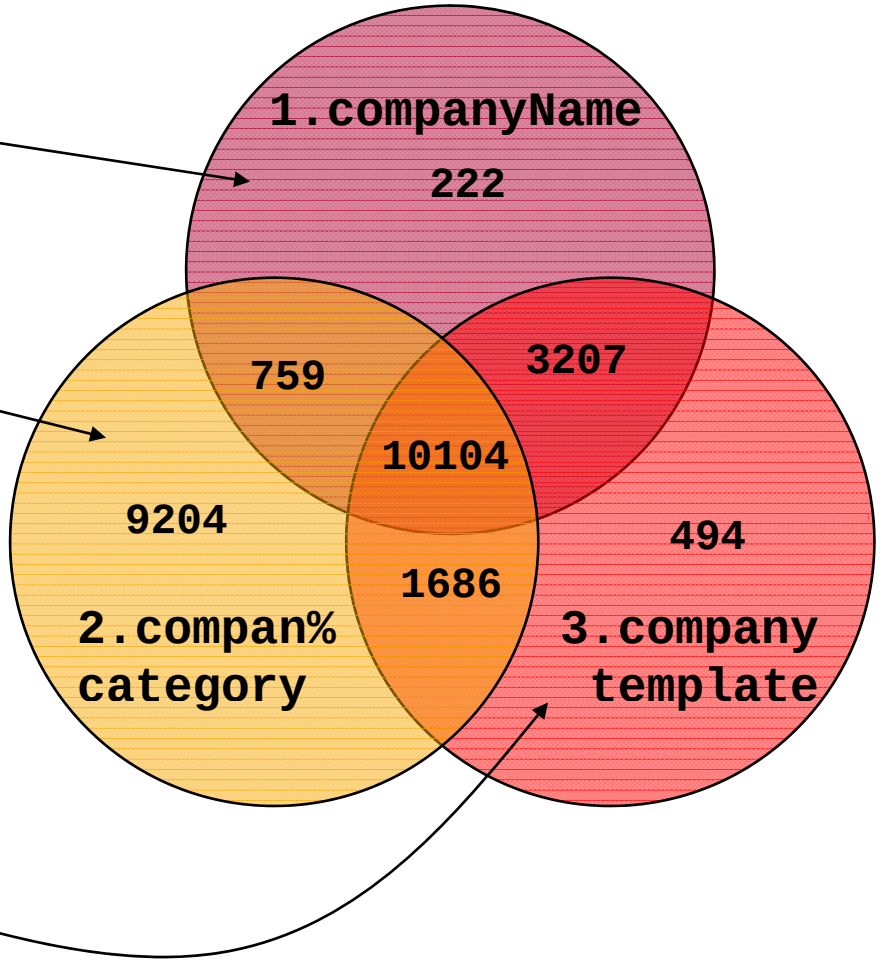


Some more about companies

Def. 1: Topics having a **`companyName`**
- 14292 companies

Def. 2: Topics having a **`category`** that starts with **`'compan%'`**
- 21753

Def. 3: Topics having a **`wikiPageUsesTemplate`** with value **`Template:infobox_company`**
- 15491



The queries…

- SELECT COUNT(*) FROM (
  - (SELECT DISTINCT I.TOPIC
    FROM DBPEDIA.INFOBOXES I
    WHERE I.ATTRIBUTE = 'companyName')
    - INTERSECT
  - (SELECT DISTINCT I.TOPIC
    FROM DBPEDIA.INFOBOXES I, DBPEDIA.ARTICLESCATEGORIES C
    WHERE I.TOPIC = C.SUBJECT
    AND UPPER(C.OBJECT) LIKE '%COMPAN%')
    - INTERSECT
  - (SELECT DISTINCT I.TOPIC
    FROM DBPEDIA.INFOBOXES I
    WHERE I.VALUE = 'Template:infobox_company')
)

Infoboxes with CompanyTemplate

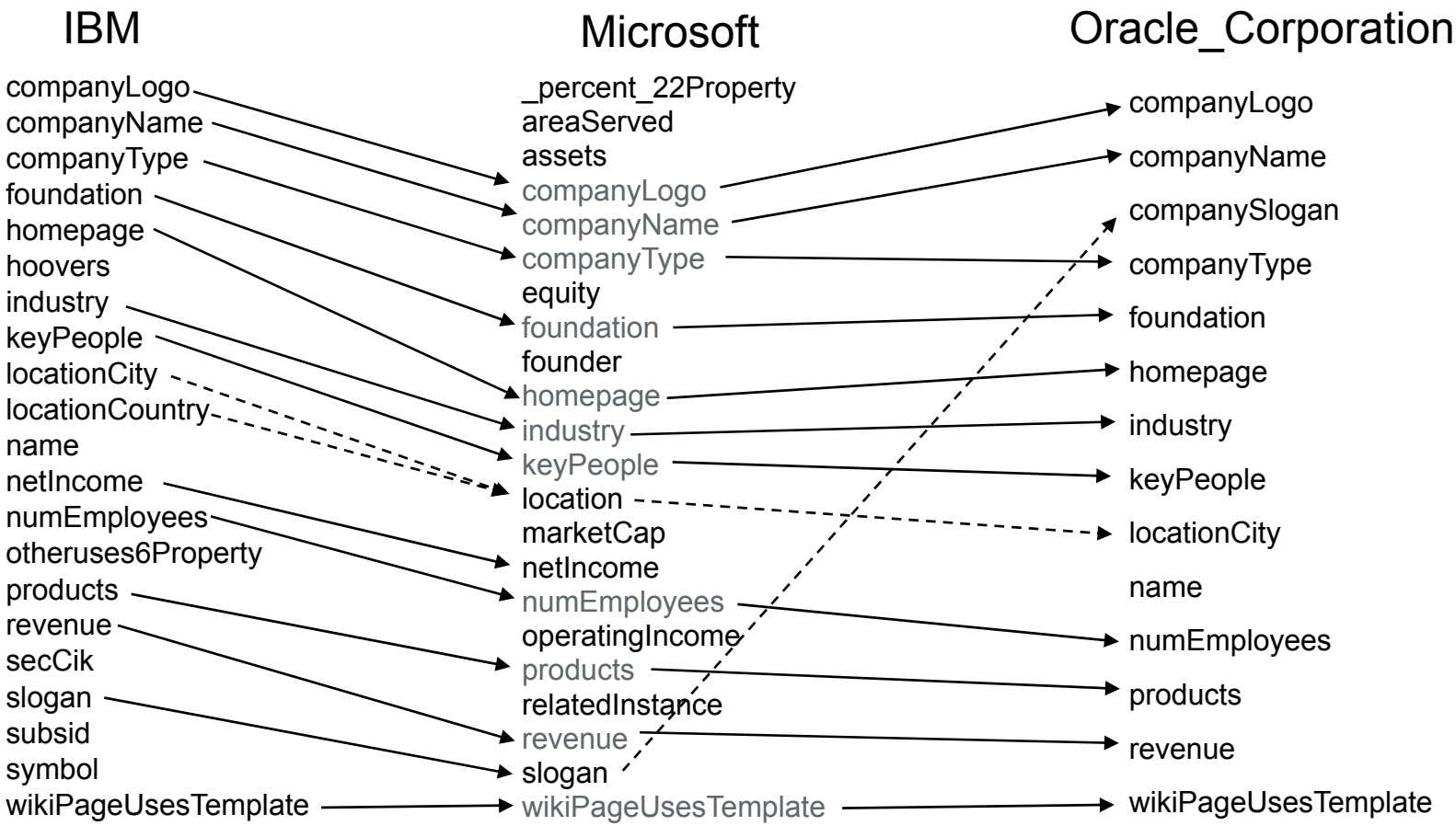
- 1083 different attributes
  - H-index 56
  - 499 appear only once
- Of the 1083 attr., 39 distinct ones contain ‘name’ as substring
- 273 companies without any name attribute

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

Zooming into IBM

otheruses6Property	IBM (disambiguation)
otheruses6Property	Big Blue
wikiPageUsesTemplate	Template:otheruses6
companyName	International Business Machines Corporation
companyLogo	http://upload.wikimedia.org/wikipedia/commons/5/51/IBM_logo.svg
companyType	Public_company
foundation	Endicott%2C_New_York
foundation	New_York
foundation	United_States
locationCity	Armonk%2C_New_York
locationCity	New_York
locationCountry	USA
slogan	On Demand Business, in demand people
keyPeople	Samuel_J._Palmisano
keyPeople	Mark_Loughridge
keyPeople	Dan_Fortin
keyPeople	Nick_Donofrio
industry	Computer_hardware
industry	Computer_software
industry	Consultant
industry	IT_Service_Management
products	List_of_IBM_products
revenue	98800000000
netIncome	10800000000
numEmployees	386558
subsid	ADSTAR
subsid	FileNet
subsid	Informix
subsid	Iris_Associates
subsid	Lotus_Software
subsid	Rational_Software
subsid	Sequent_Computer_Systems
subsid	Tivoli_Systems%2C_Inc.
homepage	http://www.ibm.com/
wikiPageUsesTemplate	Template:infobox_company
name	IBM Corp.
symbol	IBM
secCik	51143
hoovers	10796
wikiPageUsesTemplate	Template:finance_links

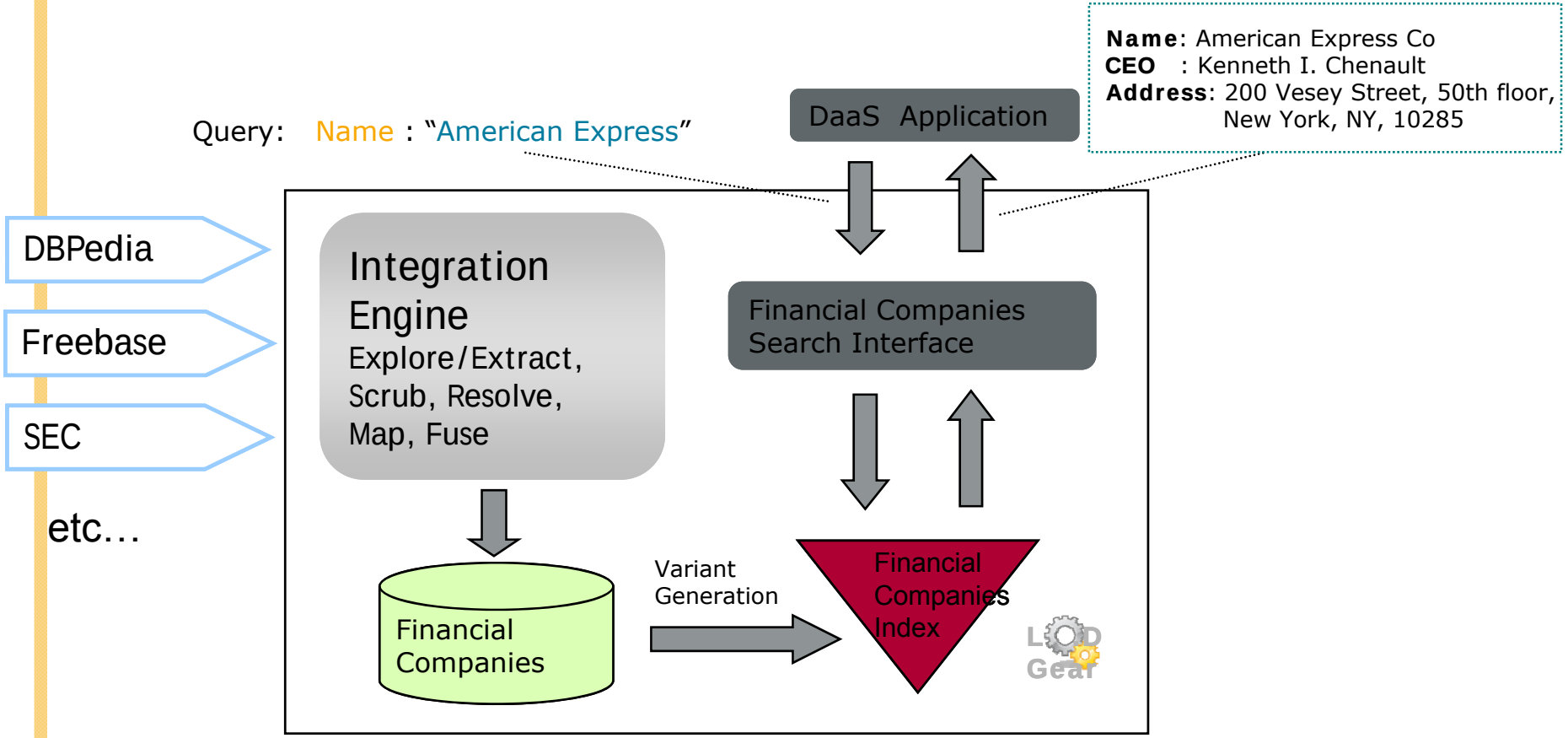
Schemata of IBM, Microsoft, and Oracle



The problem – a schema and data mess

- Triples and ill-defined templates invite disaster.
- Extract attribute values from various text fields in the source
  - 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
  - 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: ? or Undisclosed or Image:green_up.png or Assets exceed £4 billion GBP or Wal-Mart or € bn (as of 2004) or http://www.credit-suisse.com/investors/en/reports/2007_results_q4.jsp

LOD Gear: System for building and maintaining high quality Linked-Open-Data



Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff

Three themes

- Data heterogeneity
- Semantic heterogeneity
- Object heterogeneity

Project Goals:
Profiling Data Heterogeneity

- In DBPedia there are companies, people, artists, buildings, and many more different concepts.

- Goal:
  - find out such concepts or cluster of objects
  - display them with useful /cool statistics
  - allow inter- concept / cluster comparisons
  - allow comparison with existing classifications

Project Goals:
Profiling Semantic Heterogeneity

- Given existing concepts / classes of objects, e.g. all subjects that use a specific Wikipedia template.

- Goal:
  - Find out interesting / cool statistics
  - Find Patterns
    - longitude=73° → time zone=UTC-5
    - company name attribute → industry attribute
  - Explore Schema Matchings
    - suggestions
    - what if *name* and *family name* were merged
  - Explore Join Paths
    - location – river / city / country

Project Goals:
Profiling Object Heterogeneity

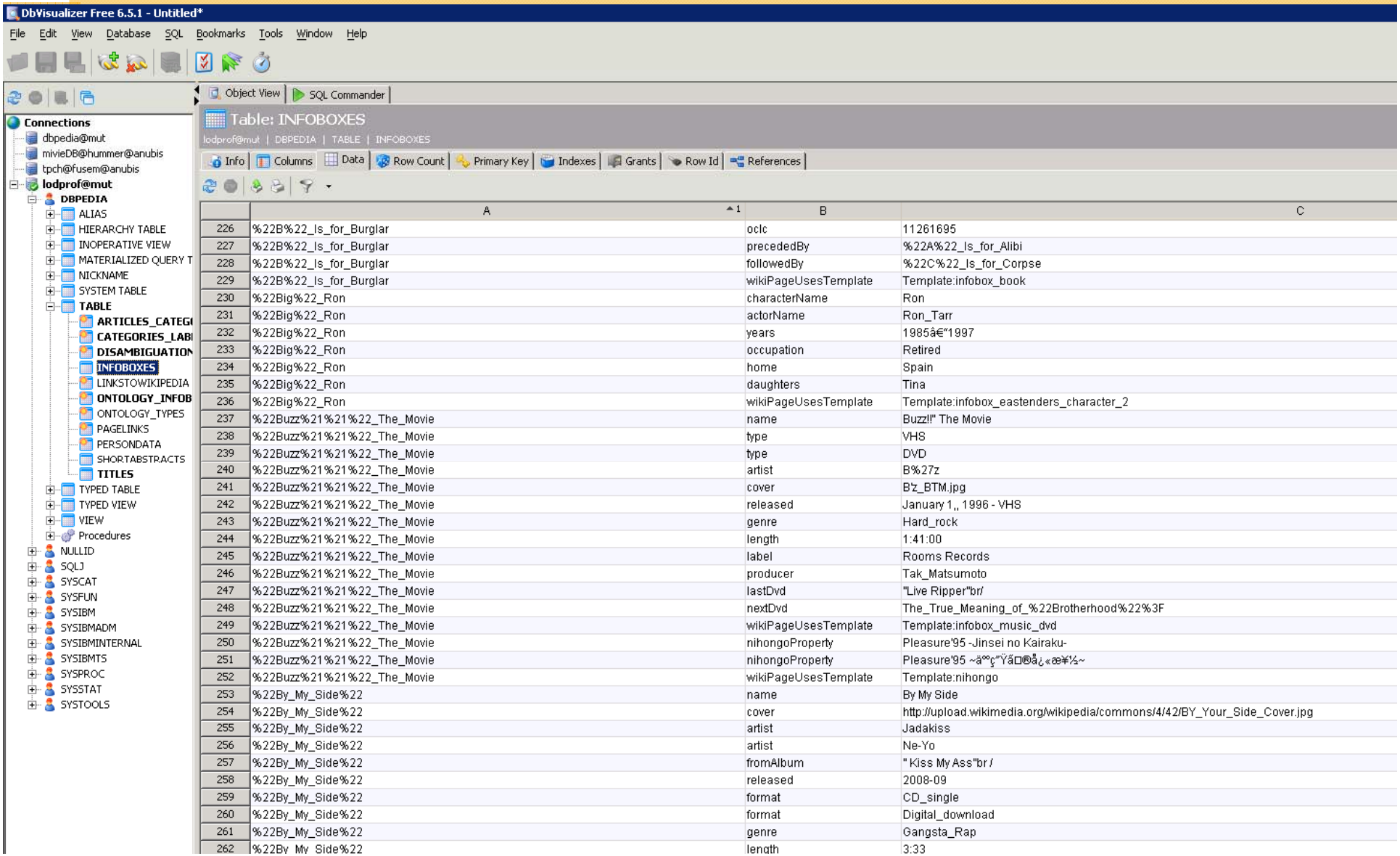
- Given existing concepts / classes of objects, e.g. all subjects that use a specific Wikipedia template.

- Goal:
  - How much information do objects provide ?
  - How do attribute values correspond to Wikipedia Texts ?
  - What can we find out by looking at external resources ?
  - Rank objects …

Infrastructure

- mut/sokar.hpi.uni-potsdam.de
- DB2 database **lodprof** (requires tuning!)
  - DBPEDIA.INFOBOXES
  - DBPEDIA.ARTICLES_CATEGORIES
  - DBPEDIA.CATEGORIES_LABEL
  - DBPEDIA.DISAMBIGUATIONLINKS
  - DBPEDIA.LINKSTOWIKIPEDIA
  - DBPEDIA.ONTOLOGY_INFOBOXES
  - DBPEDIA.ONTOLOGY_TYPES
  - DBPEDIA.PAGELINKS
  - DBPEDIA.PERSONDATA
  - DBPEDIA.SHORTABSTRACTS
  - DBPEDIA.TITLES
- User/Pwd: see whiteboard
- Java 1.6
- Google Web Toolkit

DBVisualizer



	A	B	C
226	%22B%22_Is_for_Burglar	oclc	11261695
227	%22B%22_Is_for_Burglar	precededBy	%22A%22_Is_for_Alibi
228	%22B%22_Is_for_Burglar	followedBy	%22C%22_Is_for_Corpse
229	%22B%22_Is_for_Burglar	wikiPageUsesTemplate	Template:infobox_book
230	%22Big%22_Ron	characterName	Ron
231	%22Big%22_Ron	actorName	Ron_Tarr
232	%22Big%22_Ron	years	1985â€“1997
233	%22Big%22_Ron	occupation	Retired
234	%22Big%22_Ron	home	Spain
235	%22Big%22_Ron	daughters	Tina
236	%22Big%22_Ron	wikiPageUsesTemplate	Template:infobox_eastenders_character_2
237	%22Buzz%21%21%22_The_Movie	name	Buzz!!" The Movie
238	%22Buzz%21%21%22_The_Movie	type	VHS
239	%22Buzz%21%21%22_The_Movie	type	DVD
240	%22Buzz%21%21%22_The_Movie	artist	B%27z
241	%22Buzz%21%21%22_The_Movie	cover	B'z_BTM.jpg
242	%22Buzz%21%21%22_The_Movie	released	January 1,, 1996 - VHS
243	%22Buzz%21%21%22_The_Movie	genre	Hard_rock
244	%22Buzz%21%21%22_The_Movie	length	1:41:00
245	%22Buzz%21%21%22_The_Movie	label	Rooms Records
246	%22Buzz%21%21%22_The_Movie	producer	Tak_Matsumoto
247	%22Buzz%21%21%22_The_Movie	lastDvd	"Live Ripper"br/
248	%22Buzz%21%21%22_The_Movie	nextDvd	The_True_Meaning_of_%22Brotherhood%22%3F
249	%22Buzz%21%21%22_The_Movie	wikiPageUsesTemplate	Template:infobox_music_dvd
250	%22Buzz%21%21%22_The_Movie	nihongoProperty	Pleasure'95 -Jinsei no Kairaku-
251	%22Buzz%21%21%22_The_Movie	nihongoProperty	Pleasure'95 ~ä°ç"Ÿå□®å¿«æ¥½~
252	%22Buzz%21%21%22_The_Movie	wikiPageUsesTemplate	Template:nihongo
253	%22By_My_Side%22	name	By My Side
254	%22By_My_Side%22	cover	http://upload.wikimedia.org/wikipedia/commons/4/42/BY_Your_Side_Cover.jpg
255	%22By_My_Side%22	artist	Jadakiss
256	%22By_My_Side%22	artist	Ne-Yo
257	%22By_My_Side%22	fromAlbum	" Kiss My Ass"br /
258	%22By_My_Side%22	released	2008-09
259	%22By_My_Side%22	format	CD_single
260	%22By_My_Side%22	format	Digital_download
261	%22By_My_Side%22	genre	Gangsta_Rap
262	%22By_My_Side%22	length	3:33

Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff

Homework!

- Explore DBPedia (and Wikipedia)!
  - Appreciate problem and profiling need
  - Next week: Mr. DBPedia is here
    - 29.4. 13 – 17 Uhr
- See TED-talk by Tim Berners-Lee
- Choose problem (email with top 2 choices)
  - Infrastructure (6x 1/6 person)
  - Domain heterogeneity (2 persons)
  - Schematic heterogeneity (2 persons)
  - Object heterogeneity (Discovery per Object) (2 persons)
- Identify solution directions

Questions, wishes, …

- Now, or ...
- Office: A.1-13
- Consultations: Tuesdays 15-16 Uhr
  or by arrangement
- Email: naumann@hpi.uni-potsdam.de
- Phone: (0331) 5509 280