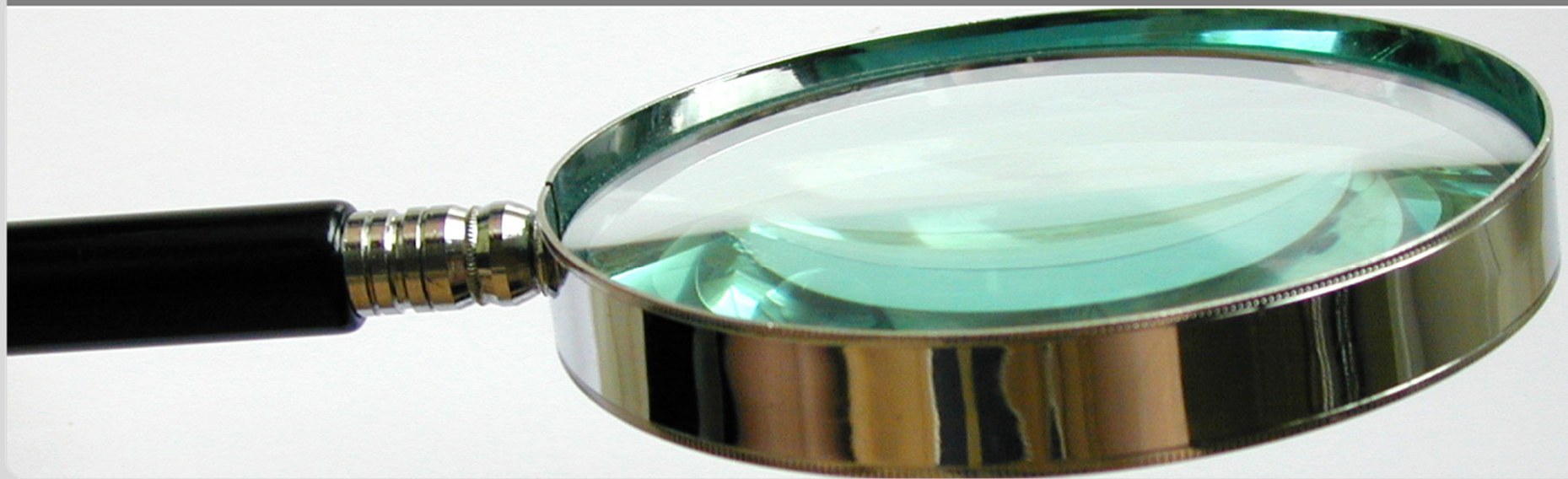


Semantische Technologien und maschinelles Lernen zur Verwaltung und Nutzung von Big Data

Rudi Studer & Andreas Harth & Achim Rettinger & Thanh Tran
Institute AIFB & FZI Research Center for Information Technology

Institute of Applied Informatics and Formal Description Methods (AIFB)



Agenda

- Motivation
- Semantic Data Integration
- Semantic Search
- Statistical Learning
- Conclusions

MOTIVATION

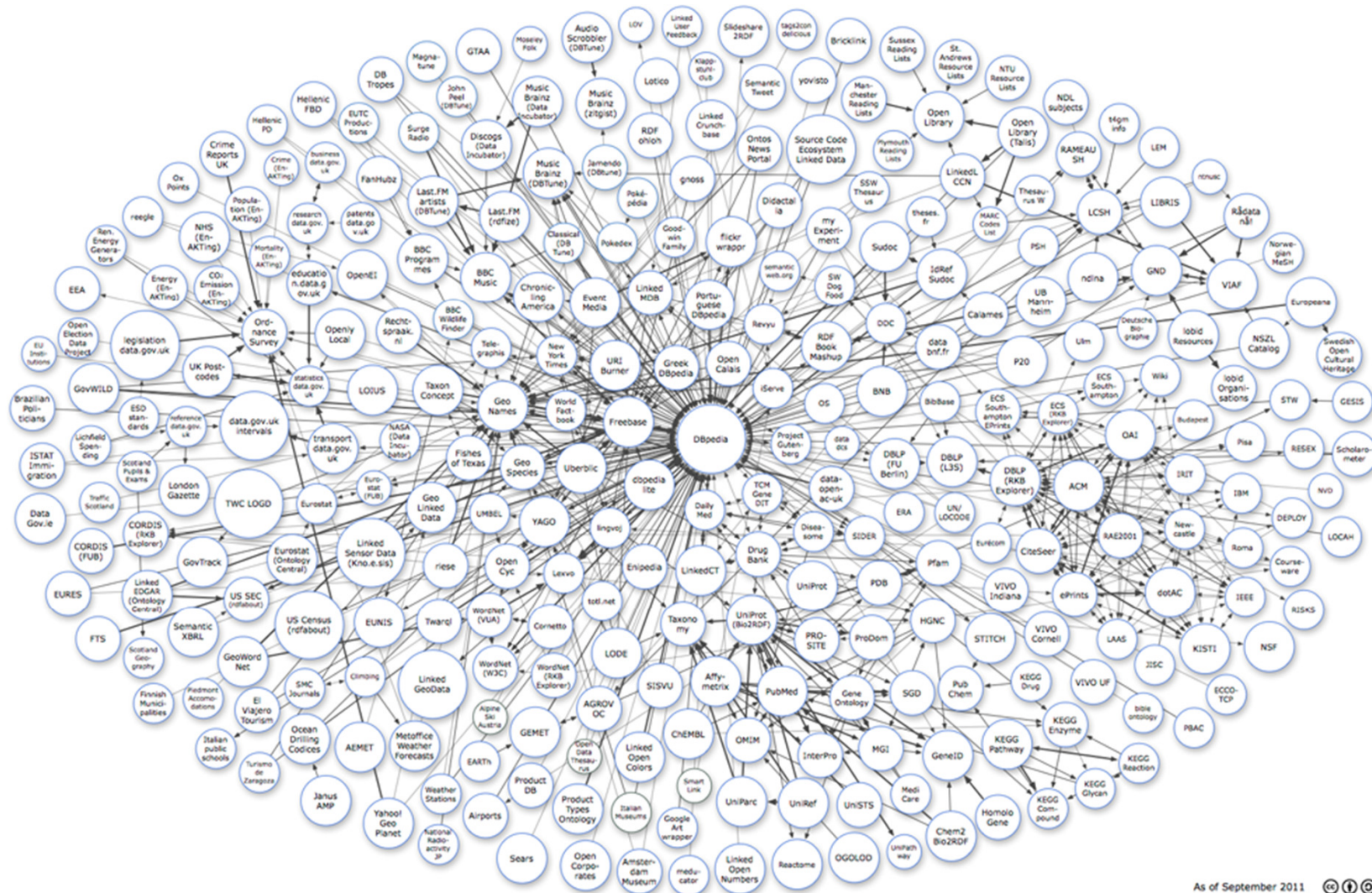
Motivation

- With increased use of computers more and more data is being stored
 - Organisations rely on data to support research and business decisions
 - Data drives policy decisions in government
 - Individuals rely on data from the Web for information and communication

- Data volumes explode
 - More and more data available on the Web is represented in **Semantic Web standards**
 - Linking Open Data (LOD) initiative provides a lot of structured linked data: **Web of Data**

- Various approaches enable insights
 - Data integration on semantic level
 - Semantic interpretation of long-tail queries
 - Deriving new relationships

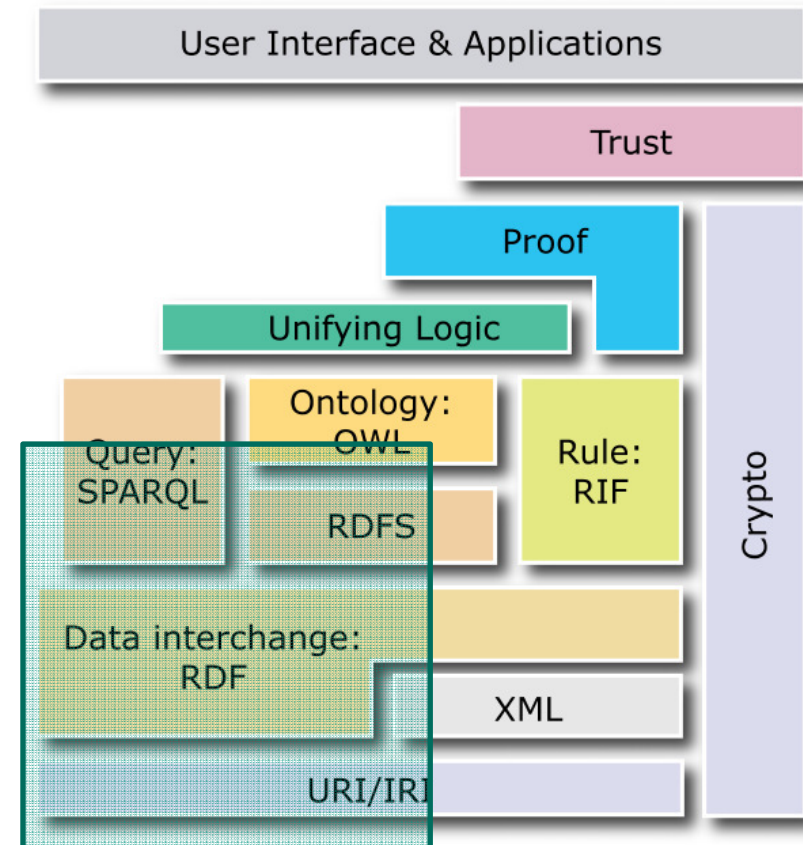
Linked Data on the Web



As of September 2011

Semantic Technologies

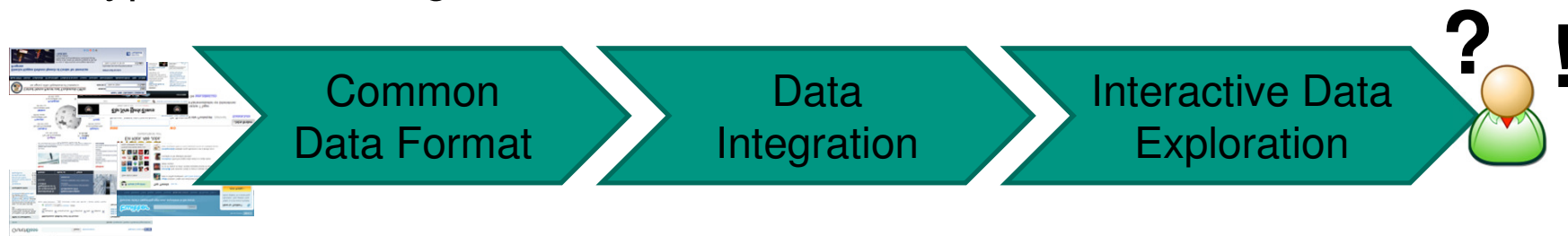
- Semantic Web technologies, standardised by the W3C, are mature:
 - **RDF** recommendation in 1999, update in 2004
 - RDFa (RDF in HTML) note in 2008
 - **RDFS** recommendation in 2004
 - **SPARQL** recommendation in 2008
 - **OWL** recommendation in 2004, update in 2009
 - **RIF Core** recommendation in 2010
- **Linked Data** is a subset of the Semantic Web stack
 - Uniform use of URIs
 - Use of RDF and SPARQL



SEMANTIC DATA INTEGRATION

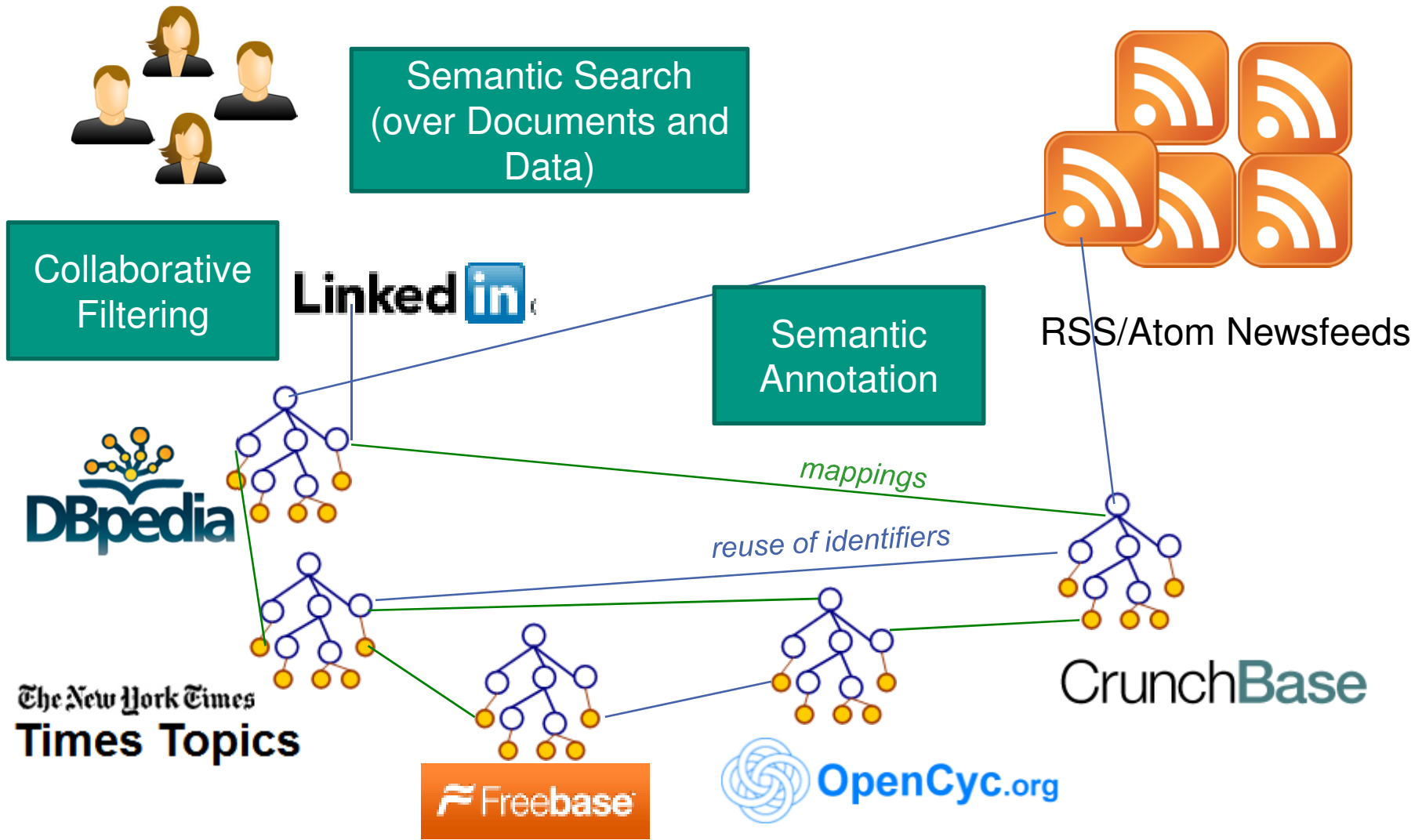
Motivation for Semantic Web Technologies

- Semantic Web/Linked Data technologies are well-suited for data integration
 - **Standard languages** for representing mappings
 - **Linked Data principles** for linking data across datasets, and for publishing and accessing integrated Linked Data
- Typical data integration scenario



- We show
 - Novel data sets that are published as part of the Web of Data
 - An application showcasing the benefits of Linked Data to end-users
 - Novel generic mechanisms, approaches, and technologies for integration

Scenario: Integration of News with Linked Data



Common Data Format/Access Protocol

- Access to networked data and ontologies is a first step
 - DBpedia, Freebase, NYTimes Topics, CrunchBase already exist as Linked Data and are interlinked

- Next steps:
 - Perform **entity matching** in news feeds (identifying entities in text)
 - Semantic search to enable **complex queries** and collaborative filtering

- Required:
Principled way for integrating data from **services** providing data (e.g., via LinkedIn API) or functionality (e.g., entity matching)

Linked APIs Motivation

- The Web today is not only about serving static data:
 - Data is often **dynamically** created as a result of some calculation carried out over input data (e.g., weather information)
 - Service endpoints, forms and APIs are used to trigger **functionalities** in the **Web** and the **real world** as well (e.g., ordering a pizza or solving a recaptcha)
 - Programmableweb.com lists ~4300 APIs¹



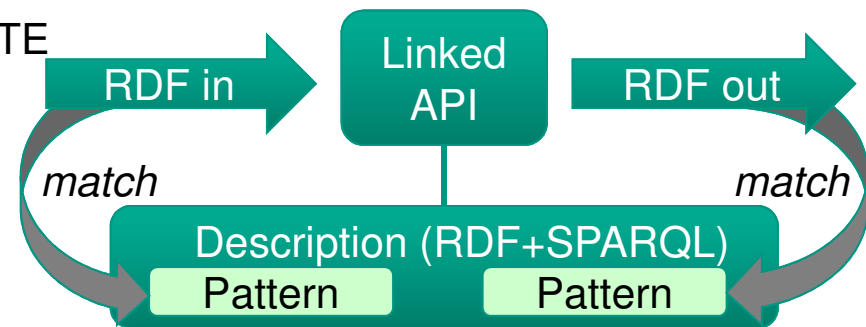
¹<http://programmableweb.com>

Linked APIs

- Web APIs use **heterogenous data formats**, different architectural styles, and are mostly only **textually** described
- Developers have to gain a deep understanding of every API and write **individually tailored code** to consume services in applications

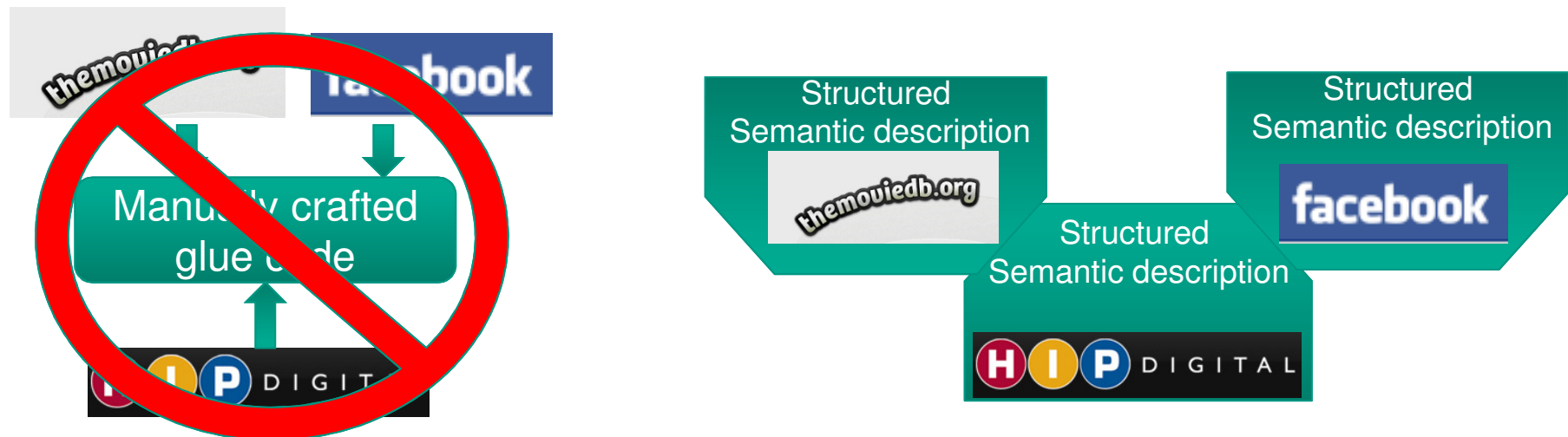
The Linked APIs effort aims to promote a scalable and efficient style of services, by bringing together:

- **RESTful** services (respecting Web architecture)
 - resource-oriented
 - manipulated with HTTP verbs
 - GET, PUT (, PATCH), POST, DELETE
 - Negotiate representations
- **Linked Data**
 - Uniform use of URIs
 - Use of RDF and SPARQL



Combining Linked APIs

- Increased value comes from combinations of services/APIs
 - 6300 mashups on programmableweb.com
 - Manual effort is required for compositions (glue code)
 - Structured service/API descriptions ease the composition process considerably
 - Semantic descriptions allow to execute several tasks automatically (e.g., data matching, discovery, repair)



Integration and Interoperation Summary

- Linked Data and Linked APIs as **common abstraction** for accessing **data** and **functionality**
- Linked APIs provide means for publishing and reusing data services on the Web
- Linked Data/Linked APIs can be used in
 - Data processing workflows
 - Query processing
- <http://linkedservices.org/> - community website (KIT, U Ghent, USC ISI, OntoText) with further information and links, reference to mailing list

SEMANTIC SEARCH

Motivation for Semantic Search

- Common queries solved
 - navigational, entity search with unambiguous named entity mention
- But long tail queries...
- Several **problematic cases** (long tail queries)
 - **Ambiguous / imprecise queries (entity queries)**
 - “George Bush” the beer brewer from Germany
 - **Complex queries (aggregated, relational queries)**
 - “digital camera under 300 dollars *produced by canon* in 1992”

Use **semantics** captured by thesauri, ontologies, semantic meta(data) to obtain **precise understanding**, to **aggregate information** from different sources, and to retrieve relevant results!

Semantic Search Solution

Search Intent Interpretation, Refinement and Exploration



Search Results

Keywords

queen single

Click on one of the suggestions to initiate translation! (can take a few seconds)

queen single
Set searchfield to "queen single"

A (queen) is a Single

B writer A (queen)
B is a Single

A is a Single
A producer B (queen)

Query Completions

queen single.php
queen singled
queen singlar
queen singlarpt
queen singles
queen singles-1997-2007
queen singles/2002/03/04/the
queen sinales/2002/03/25/new

Term Completions

Queen (band)
Queen (band)
Queen (band)
Queen (band)
Queen (band)
Queen (band)
Queen (band)

- Initial Query**
See Entire Query
- ?sx1**
- A Kind of Magic (song)
 - Another One Bites the Dust
 - Back Chat
 - Bicycle Race
 - Body Language (song)
 - Calling All Girls
 - Crazy Little Thing Called Love
 - Fat Bottomed Girls
 - Good Old-Fashioned Lover Boy
 - Hammer to Fall
 - Heaven for Everyone
 - I Want to Break Free
 - It's Late
 - It's a Hard Life
 - Keep Yourself Alive
 - Killer Queen
 - Las Palabras de Amor
 - Liar (Queen song)
 - Long Away
 - Mustapha

RESULT COLUMN1

producer

Range: All Values (43)

type

Range: All Values (43)

writer

Range: All Values (42)


- Musical Artist (42)
 - Brian May (13)
 - Frank Musker (1)
 - Freddie Mercury (14)
 - John Deacon (7)
 - Roger Meddows-Taylor (7)

Facets

Semantic Search Solution

Result Inspection, Analysis and Browsing

Earthquake
Search

Login / Register


IWB Tabs

Semantic Wiki | Table | Graph

View | Blog | Edit | Revisions

An **earthquake** (also known as a **tremor** or **tembler**) is the result of a sudden release of energy in the **Earth's crust** that creates **seismic waves**. Earthquakes are recorded with a **seismometer**, also known as a seismograph. The **moment magnitude** of an earthquake is conventionally reported, or the related and mostly obsolete **Richter** magnitude, with magnitude 3 or lower earthquakes being mostly **imperceptible** and magnitude 7 causing serious damage over large areas. Intensity of shaking is measured on the modified **Mercalli scale**.

At the Earth's surface, earthquakes manifest themselves by shaking and sometimes displacing the ground. When a large earthquake **epicenter** is located offshore, the seabed sometimes suffers sufficient displacement to cause a **tsunami**. The shaking in earthquakes can also trigger landslides and occasionally volcanic activity.

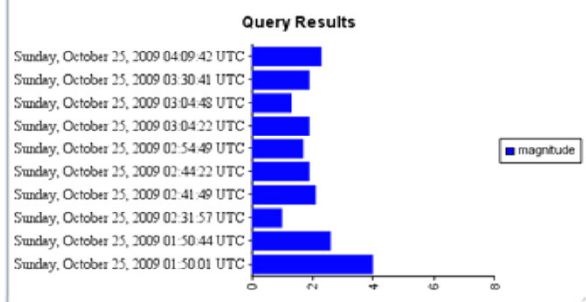
In its most generic sense, the word *earthquake* is used to describe any seismic event — whether a natural **phenomenon** or an event caused by humans — that generates seismic waves. Earthquakes are caused mostly by rupture of geological **faults**, but also by volcanic activity, landslides, mine blasts, and nuclear experiments. An earthquake's point of initial rupture is called its **focus** or **hypocenter**. The term **epicenter** refers to the point at ground level directly above the hypocenter.

Contents

- Naturally occurring earthquakes
 - Earthquake fault types
 - Earthquakes away from plate boundaries
 - Shallow-focus and deep-focus earthquakes
 - Earthquakes and volcanic activity
- Earthquake clusters
 - Aftershocks
 - Earthquake swarms
 - Earthquake storms
- Size and frequency of occurrence
- Induced seismicity
- How to measure and locate an earthquake
- Effects/impacts of earthquakes
 - Shaking and ground rupture
 - Landslides and avalanches
 - Fires
 - Soil liquefaction
 - Tsunami
 - Floods
 - Human impacts
- Preparation
- History
 - Pre-Middle Ages

Query Results


Input: datetime | Output: magnitude | Aggregation: None



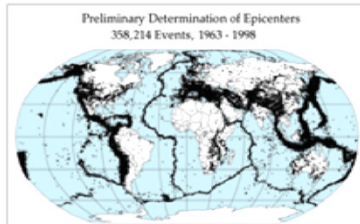
Query Results

Event	Approximate Magnitude
Sunday, October 25, 2009 04:09:42 UTC	~4.5
Sunday, October 25, 2009 03:30:41 UTC	~4.0
Sunday, October 25, 2009 03:04:48 UTC	~3.5
Sunday, October 25, 2009 03:04:22 UTC	~3.5
Sunday, October 25, 2009 02:54:49 UTC	~3.5
Sunday, October 25, 2009 02:44:22 UTC	~3.5
Sunday, October 25, 2009 02:41:49 UTC	~3.5
Sunday, October 25, 2009 02:31:57 UTC	~3.5
Sunday, October 25, 2009 01:50:44 UTC	~4.0
Sunday, October 25, 2009 01:50:01 UTC	~4.5

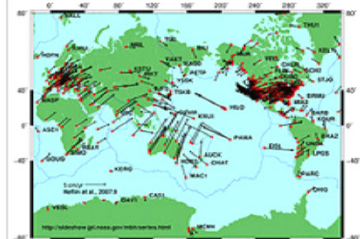
GMap Earthquake



Preliminary Determination of Epicenters
358,214 Events, 1963 - 1998



Global earthquake **epicenters**, 1963-1998



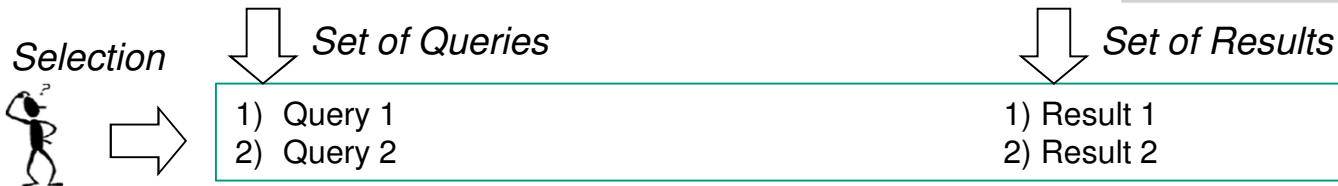
Keyword Query Processing: Problem

“Articles of researchers at Stanford with Turing Award” „Stanford Article Turing Award“



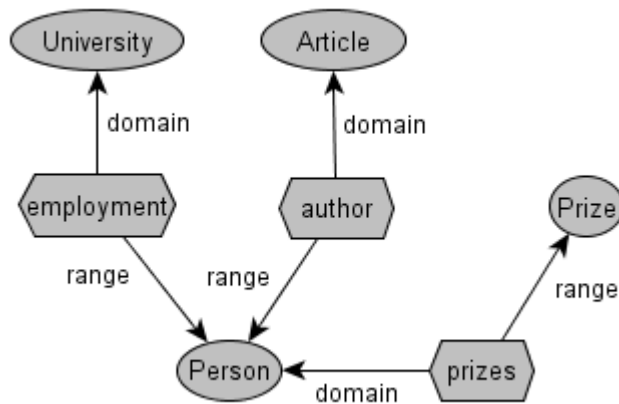
Specification

- Keywords might produce large number of matching elements in the data graph
- The **data graph might be large in size**
- Search complexity increases substantially with the size of the graph
- **Large number of results**

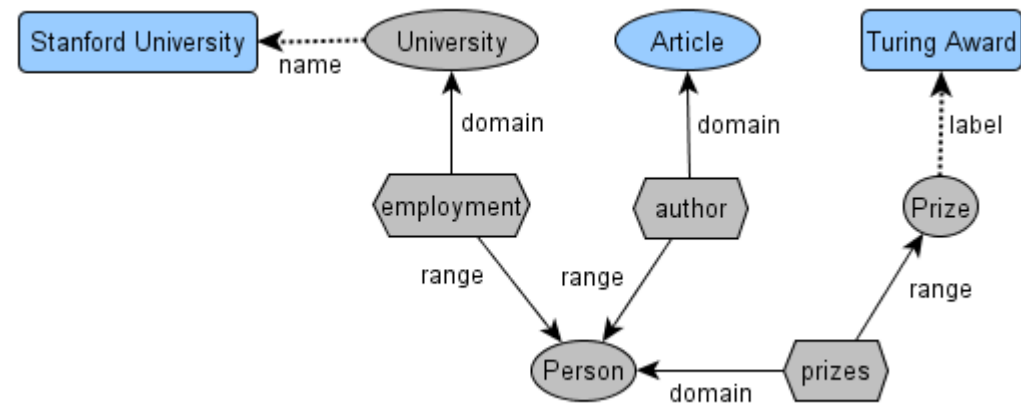


Ideas behind Solution: Top-k over Query Space

Schema graph



Query space



- Decreased complexity through exploration on much **reduced summary model** called query space
- Top-k procedure for graph exploration to **compute only top-k results**
- Principle approach for ranking based on **language models for structured results**

Benefits of Semantic Search

- Solve queries in the long tail
 - Ambiguous queries
 - Complex queries
- Addressing **complex information needs of end-users**
 - Complex results to intuitive keyword-based queries
 - Both documents and facts

...Selected Challenges

■ Hybrid content management

- Indexing hybrid content (structured data & text)
- Processing hybrid queries
- **Ranking hybrid results** (facts combined with text)

■ Querying paradigm for complex retrieval tasks

- Querying at once vs. iterative exploration
- Combination of keywords, NL and facets?

■ Semantics for broader search context/process:

from querying to browsing to intuitive presentation, supporting complex analysis of data / results

STATISTICAL LEARNING

Motivation for Statistical Learning

Statistical learning is a method that can help to solve tasks for data integration, semantic search, e.g.:

1. Textmining:

- Extract structure (facts) from unstructured sources (text)
- Link extracted facts to knowledge bases

2. Statistical Analysis:

- Cluster semantic data (find similar entities, facts, events,...)
- Predict facts and events, analyze trends

1. Textmining: Solutions I

Unsupervised Semantic Parsing (USP):

- Identify similar terms
- Identify similar syntactical structures

Microsoft buys Powerset

Microsoft acquires semantic search engine Powerset

Powerset is acquired by Microsoft Corporation

The Redmond software giant buys Powerset

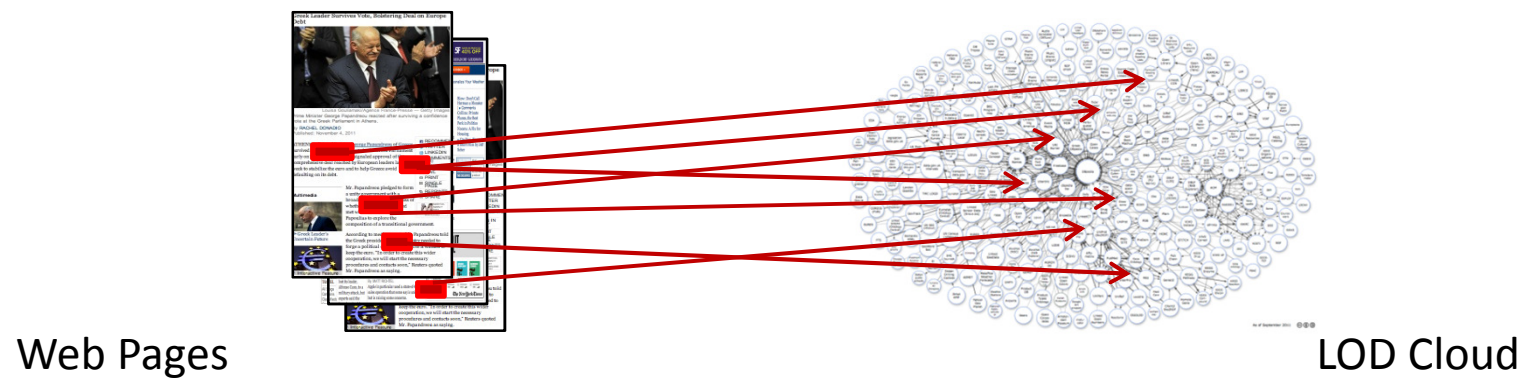
Microsoft's purchase of Powerset, ...

Automatically cluster synonymic expressions

1. Textmining: Solutions II

Semantic Annotation:

- Link text fragments to rich background knowledge (Wikipedia / Dbpedia)



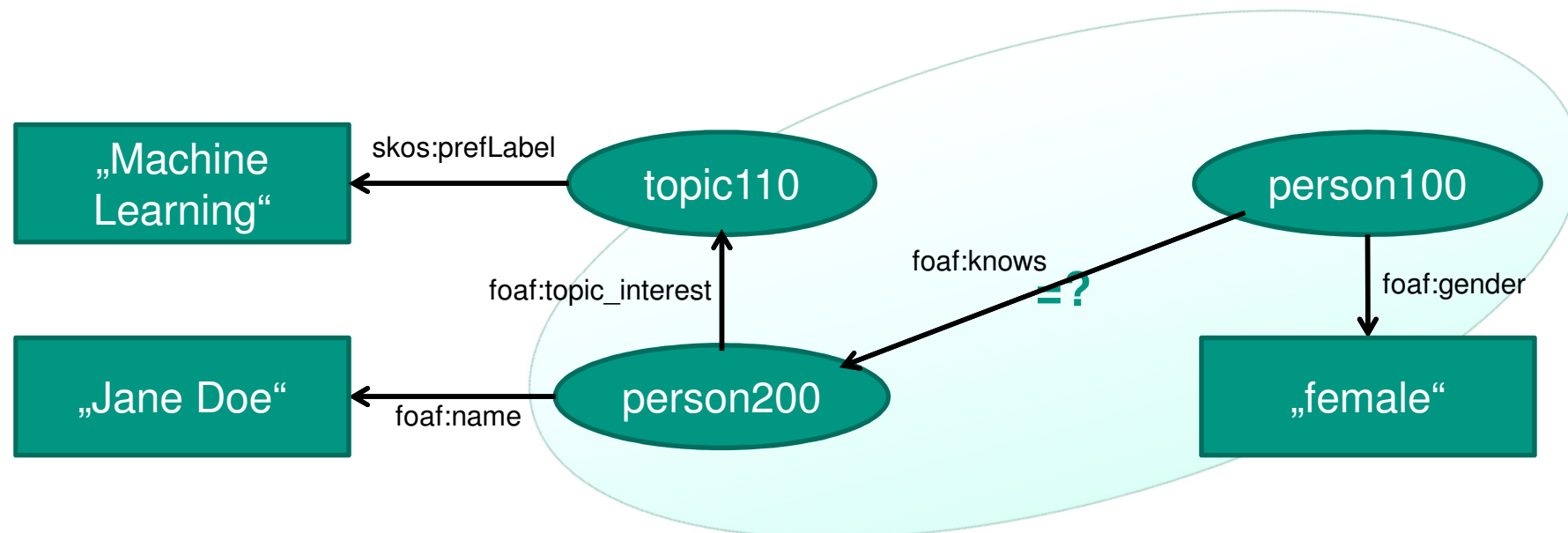
1. Textmining: Challenges / Benefits

- Benefits: Existing tools work well for
 - **fixed** set of entity types (Persons, Institutions,...)
 - **popular** domains (mainstream news, Wikipedia,...; cannot be directly applied to special domains: e.g. nanotechnology)
 - **major** languages (English, Spanish,...)
 - domains where annotated corpora are available

- Challenges for current research
 - **Cross-lingual** (cover and bridge between many languages)
 - see EU project: X-LIKE
 - **Non-standard** language (cover e.g. twitter feeds)
 - **Unsupervised** approaches (Data driven, do not require extensive annotation efforts; but don't scale, results are often hard to interpret by users)

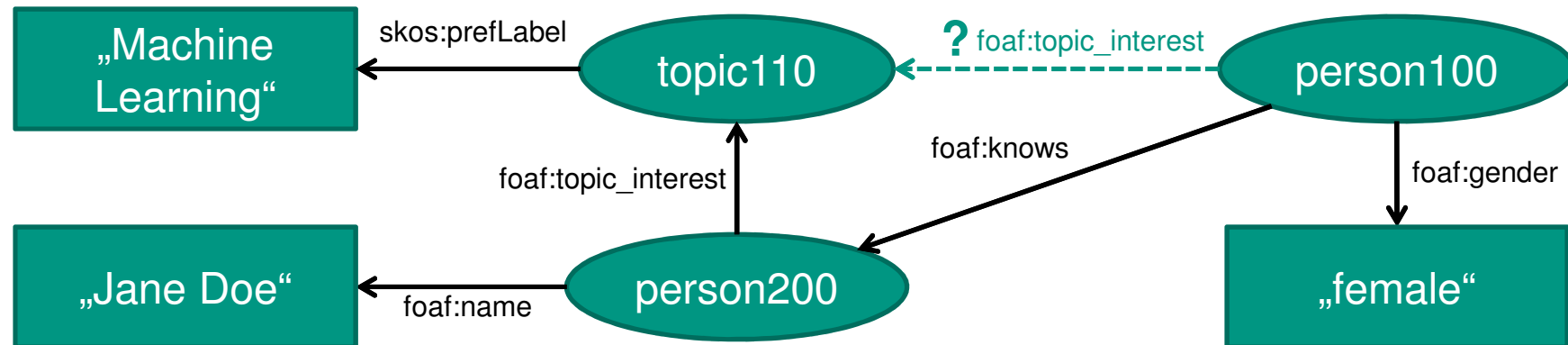
2. Statistical Analysis: Solutions I

- Find similar entities (clustering)
- Find identical entities (entity resolution)



2. Statistical Analysis: Solutions II

- Prediction (recommendation engines)



2. Statistical Analysis: Challenges / Benefits

- Research challenges
 - **Big data** analytics
 - **Very sparse** data sets (little information about a specific instance)
 - **Rich contextual** knowledge available (temporal, location)
 - Extract / Predict **complex events**
 - Provide **anytime** feedback for exploratory data analysis
 - Analyze data streams

- Potential benefits of new approaches
 - Scale to huge data sizes
 - Can deal with high dimensional sparse data sets
 - Incorporate temporal information
 - Make personalized recommendations

CONCLUSION

Conclusion

- Large and increasing amounts of semantic data
- Semantic data / technologies provide **added value**
 - **Data Integration:**
Linked Data / Linked APIs provide standards-based means for **publishing and reusing data / data services** on the Web
 - **Semantic Search:**
Addressing **complex information needs** in the **long tail**,
providing complex results to intuitive keyword-based queries
 - **Statistical Analysis/Learning:**
Extract and **predict complex events** and **links** over high-dimensional sparse Big Semantic Data
 - Combine ML with intelligent complex event processing
- Integrate methods for building up **proactive infrastructures**

Questions / Comments?

<http://www.aifb.kit.edu>

<http://www.ksri.kit.edu>

<http://www.fzi.de>