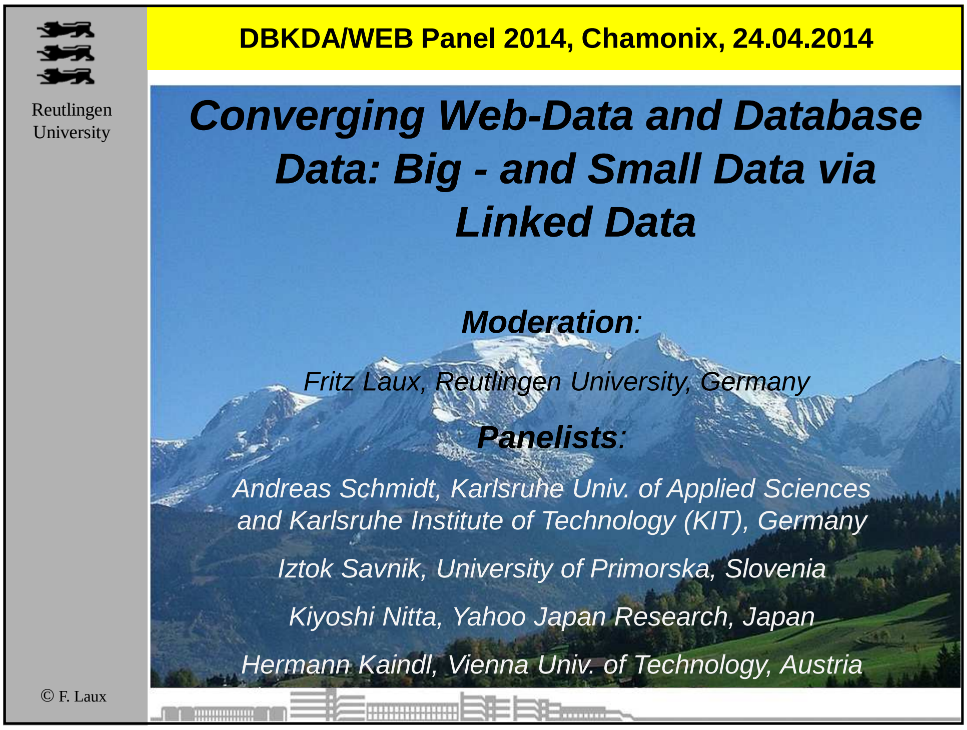


Reutlingen University

DBKDA/WEB Panel 2014, Chamonix, 24.04.2014

Converging Web-Data and Database Data: Big - and Small Data via Linked Data

***Moderation*:**

Fritz Laux, Reutlingen University, Germany

***Panelists*:**

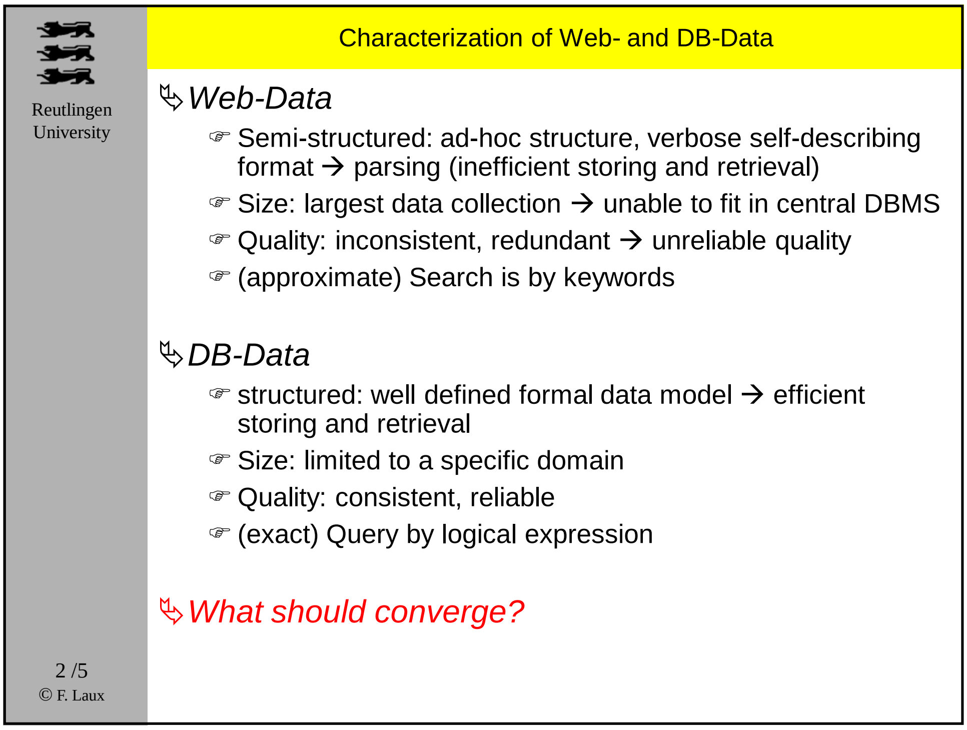
Andreas Schmidt, Karlsruhe Univ. of Applied Sciences and Karlsruhe Institute of Technology (KIT), Germany

Iztok Savnik, University of Primorska, Slovenia

Kiyoshi Nitta, Yahoo Japan Research, Japan

Hermann Kaindl, Vienna Univ. of Technology, Austria

© F. Laux



Reutlingen University

Characterization of Web- and DB-Data

➥ *Web-Data*

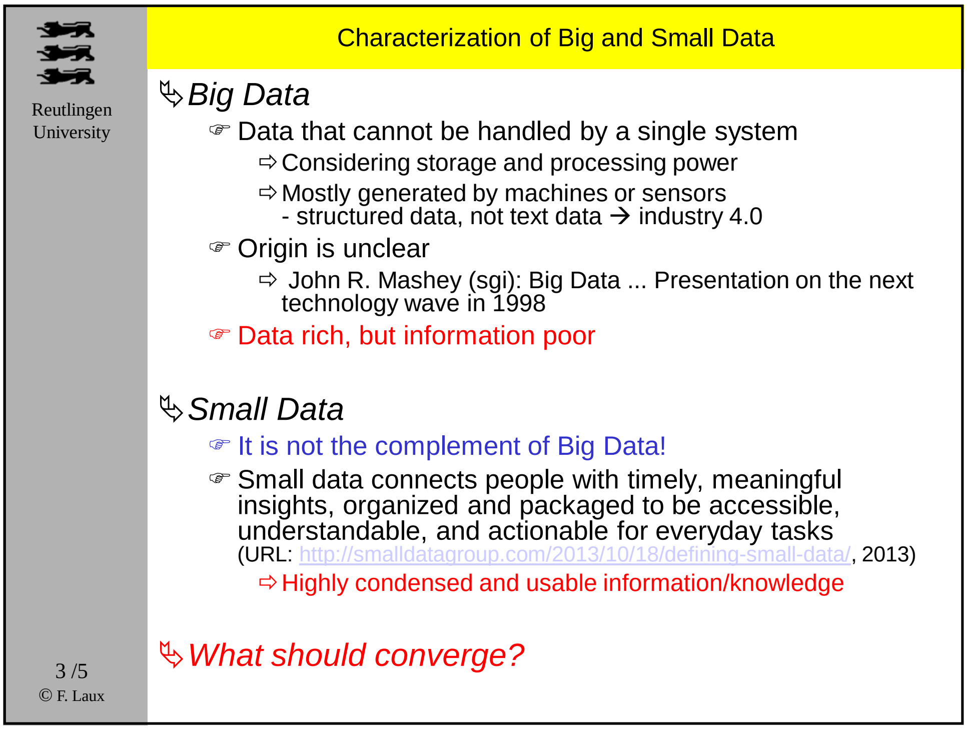
- ☞ Semi-structured: ad-hoc structure, verbose self-describing format → parsing (inefficient storing and retrieval)
- ☞ Size: largest data collection → unable to fit in central DBMS
- ☞ Quality: inconsistent, redundant → unreliable quality
- ☞ (approximate) Search is by keywords

➥ *DB-Data*

- ☞ structured: well defined formal data model → efficient storing and retrieval
- ☞ Size: limited to a specific domain
- ☞ Quality: consistent, reliable
- ☞ (exact) Query by logical expression

➥ *What should converge?*

2 /5
© F. Laux

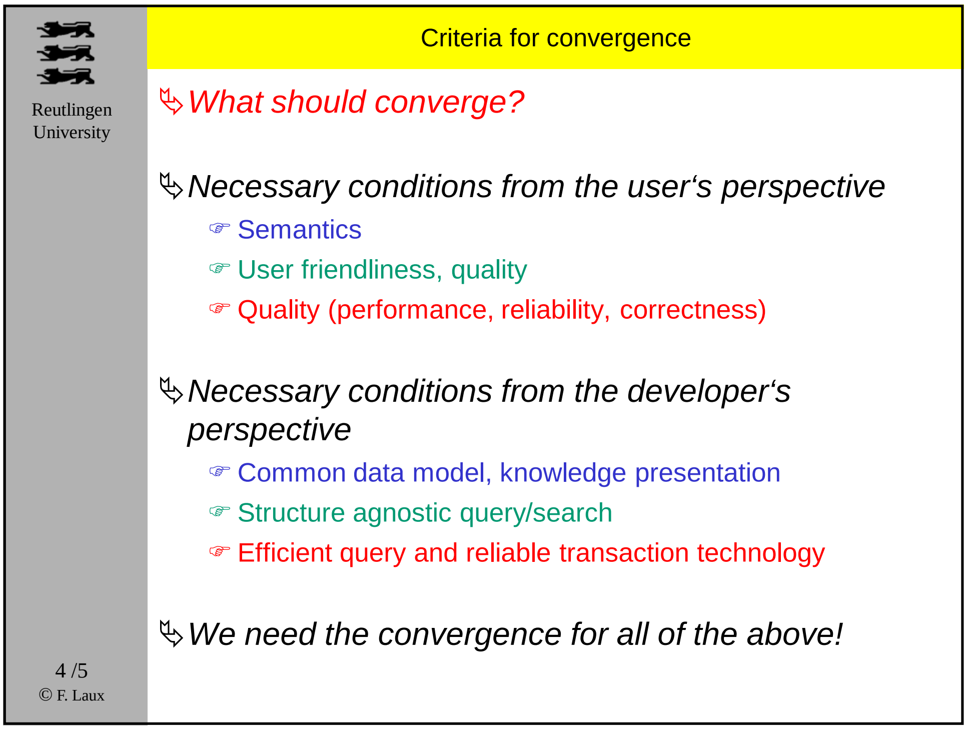


Reutlingen University

Characterization of Big and Small Data

- *Big Data*
  - Data that cannot be handled by a single system
    - Considering storage and processing power
    - Mostly generated by machines or sensors
      - structured data, not text data → industry 4.0
  - Origin is unclear
    - John R. Mashey (sgi): Big Data ... Presentation on the next technology wave in 1998
  - Data rich, but information poor
- *Small Data*
  - It is not the complement of Big Data!
  - Small data connects people with timely, meaningful insights, organized and packaged to be accessible, understandable, and actionable for everyday tasks (URL: http://smalldatagroup.com/2013/10/18/defining-small-data/, 2013)
    - Highly condensed and usable information/knowledge
- *What should converge?*

3 /5
© F. Laux



Reutlingen University

Criteria for convergence

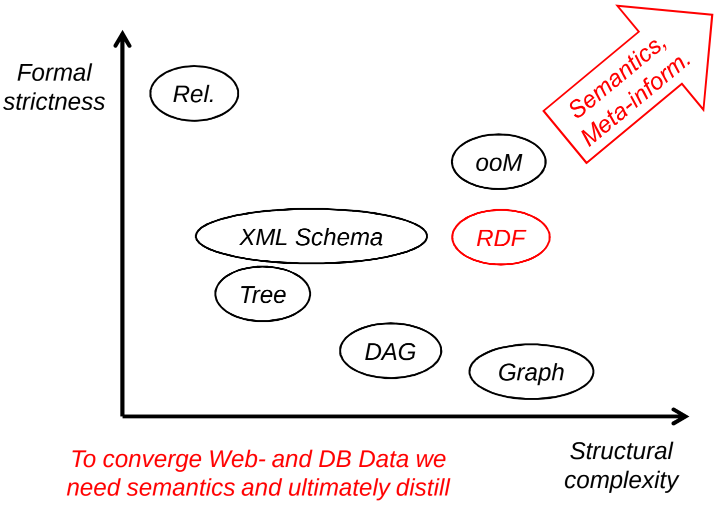
- *What should converge?*
- *Necessary conditions from the user‘s perspective*
  - Semantics
  - User friendliness, quality
  - Quality (performance, reliability, correctness)
- *Necessary conditions from the developer‘s perspective*
  - Common data model, knowledge presentation
  - Structure agnostic query/search
  - Efficient query and reliable transaction technology
- *We need the convergence for all of the above!*

4 /5
© F. Laux

Reutlingen University

Semantics: Model convergence

- *Tension field between schema strictness and structural complexity*



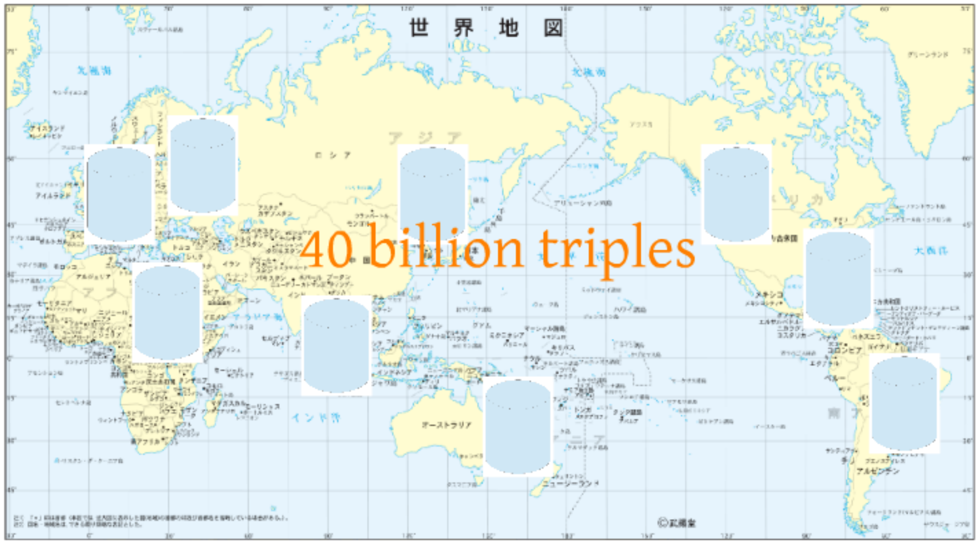
To converge Web- and DB Data we need semantics and ultimately distill valuable knowledge out of the data!

5 /5

© F. Laux

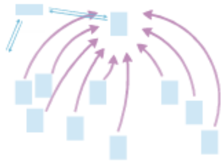
INTRODUCTION

Resource description framework (RDF) data are widely used in the Internet and their volume is growing steadily. The linked open data (LOD) project promotes the acceleration of the accumulation of RDF data to provide freely accessible on-line resources



(A-1) Local Cache Approach

gather a subset of RDF data on local computational resources



(A-2) Federated Search Approach

distribute sub-queries to several search services distributed over the Internet

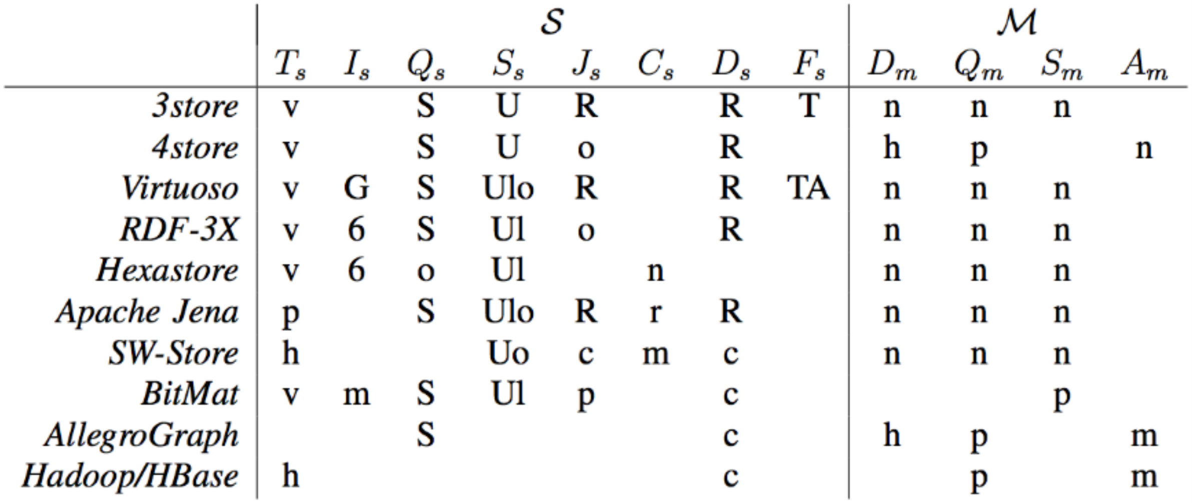
play an important role for query process efficiency

CLASSIFICATION OF RDF STORAGE MANAGERS

RDF storage managers in the local cache approach can be classified in accordance with several aspects.

$$RSM(\mathcal{S}, \mathcal{M})$$

PROPERTIES OF RDF STORAGE MANAGERS



	$\mathcal{S}$								$\mathcal{M}$			
	T_s	I_s	Q_s	S_s	J_s	C_s	D_s	F_s	D_m	Q_m	S_m	A_m
3store	v		S	U	R		R	T	n	n	n	
4store	v		S	U	o		R		h	p		n
Virtuoso	v	G	S	Ulo	R		R	TA	n	n	n	
RDF-3X	v	6	S	Ul	o		R		n	n	n	
Hexastore	v	6	o	Ul		n			n	n	n	
Apache Jena	p		S	Ulo	R	r	R		n	n	n	
SW-Store	h			Uo	c	m	c		n	n	n	
BitMat	v	m	S	Ul	p		c				p	
AllegroGraph			S				c		h	p		m
Hadoop/HBase	h						c			p		m

CHALLENGES

More varied values with S attributes than with M attributes

- Researches so far have succeeded in achieving good performances by developing single process technologies.
- While practical semantic web applications tend to process large-scale data sets, solutions based on data distribution parallelism have become more popular.

Caching techniques have not been researched that much

- Only Apache Jena and SW-Store reported confirming the efficiency of caching techniques.
- Technologies for automatic investigation and classification of processing queries might become important to utilize caching technologies.

Many researches have been carried out for developing efficient join algorithms with index structures

- This area has a long history in the research of database management systems.
- While the accumulated RDF data-set is rapidly growing and SPARQL queries are basically constructed from joins of triple patterns, join operations will be applied more strongly in semantic web applications.

Most RDF storage managers can accept SPARQL queries

- SPARQL-based RDF storage managers rarely cause semantic mismatch due to the existence of RDF algebras described in the W3C recommendation.
- While OPTIONAL operator was introduced to make the query language convenient enough, efficient processing of such queries will be one of the most crucial challenges.

DBKDA-2014 Panel

The Sixth International Conference on Advances in Databases, Knowledge, and Data Applications

April 20 - 26, 2014 - Chamonix, France

Advances on Converging WEB Data and Database Data: Big Data and Small Data via Linked Data

Andreas Schmidt

(1)
Department of Informatics and Business Information Systems
University of Applied Sciences Karlsruhe
Moltkestraße 30
76133 Karlsruhe
Germany

(2)
Institute for Applied Sciences
Karlsruhe Institute of Technologie
PO-box 3640
76021 Karlsruhe
Germany

Terminology

- What is ...
  - Big Data ?
  - Small Data ?
  - Linked Data ?

Terminology

- What is ...
  - Big Data
    - ... a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. [Snijders, C., Matzat, U., & Reips, U.-D. (2012). ‘Big Data’: Big gaps of knowledge in the field of Internet. International Journal of Internet Science, 7, 1-5]
    - Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers" [Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue]
  - Small Data
  - Linked Data

Terminology

- What is ...
  - Big Data
  - Small Data
    - “Small data is the amount of data you can conveniently store and process on a single machine, and in particular, a high-end laptop or server” (Rufus Pollok, Open Knowledge Foundation)
    - Small data connects people with timely, meaningful insights (derived from big data and/or “local” sources), organized and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks (Alan Bonde)]
  - Linked Data

Terminology

- What is ...
  - Big Data
  - Small Data
  - Linked Data
    - ... describes a method of publishing structured data so that it can be inter-linked and become more useful [http://en.wikipedia.org/wiki/Linked_data]
    - It builds upon standard Web technologies such as HTTP, RDF and URIs []
    - Origin: Open (Government) Data

Terminology

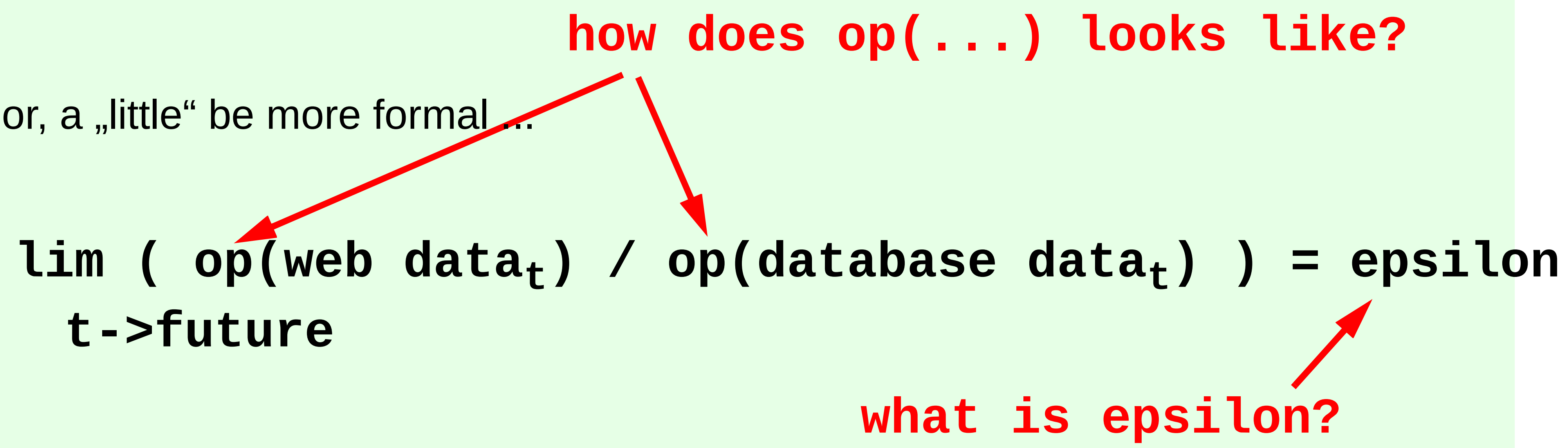
- Converging

```
f(x) = 2x^3 - 7x^2 + 12
g(x) = (12x^5-14x + 3)/(4x^2 + 2x)


lim (f(x) / g(x)) = 2/3
  x->oo
```

- „Converging WEB Data and Database Data“

$$\lim_{t\text{->}future} \left(\text{web data}_t \;/\; \text{database data}_t \right) = \text{epsilon}$$



how does op(...) looks like?

or, a „little“ be more formal ...

$$\lim_{t\text{->}future} \left(op(\text{web data}_t) \;/\; op(\text{database data}_t) \right) = \text{epsilon}$$

what is epsilon?

- „Big Data and Small Data via Linked Data“

```
linkify(Big Data) => linked data

linkify(Small Data) => linked data
```

```
Database Data => Web Data
```

- Bizer, Christian; Heath, Tom; Berners-Lee, Tim (2009). "Linked Data—The Story So Far". International Journal on Semantic Web and Information Systems 5 (3): 1–22
- Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue
- Snijders, C., Matzat, U., & Reips, U.-D. (2012). ‘Big Data’: Big gaps of knowledge in the field of Internet. International Journal of Internet Science, 7, 1-5
- http://en.wikipedia.org/wiki/Linked_data

TU
WIEN

Data and Knowledge: One Man's Opinion

Hermann Kaindl

Vienna University of Technology, Austria

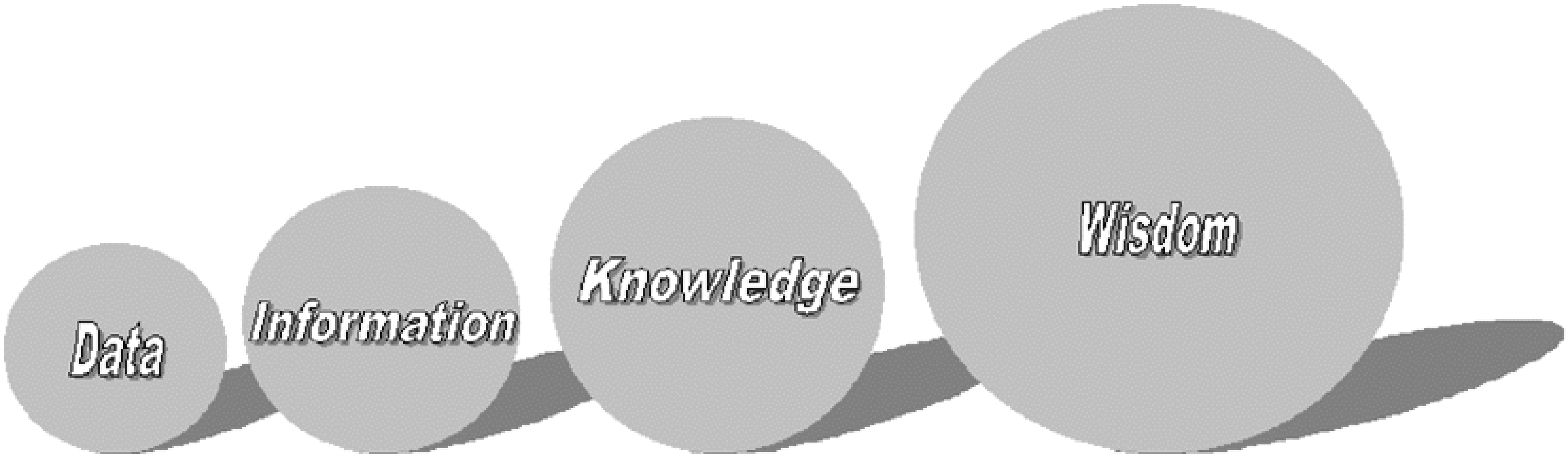
Institut für
Computertechnik

ICT

Institute of
Computer Technology

Value Chain

- Knowledge Management



Miscellenious

- „Datentechnik“ (data technology)
- From WWW to Semantic Web
- From GPS-driven navigation tools to Google cars
- Data in the Cloud
- Data looked outdated for a while, when everything seemed to be knowledge, but now data seem to be ubiquitous!

Thank you for your attention!

???

Current state of graph databases

Iztok Savnik
University of Primorska & Jožef Stefan Institute

Panel:
Big Data and Small Data via Linked Data
DBKDA, 2014

Terminology

- Linked data
  - Linked Open Data
- Open data
- Graph databases
- Knowledge bases
- Knowledge graphs

Wordnet

- Princeton's large lexical database of English.
  - Cognitve synonims: synsets ≡ concepts
    - 117,000 synsets
  - Synsets are linked by:
    - conceptual-semantic relationships, and
    - lexical relationships.
    - Include definitions of synsets.
  - Main relationships:
    - Synonymy, hyponymy (ISA), meronymy (part-whole), antonymy

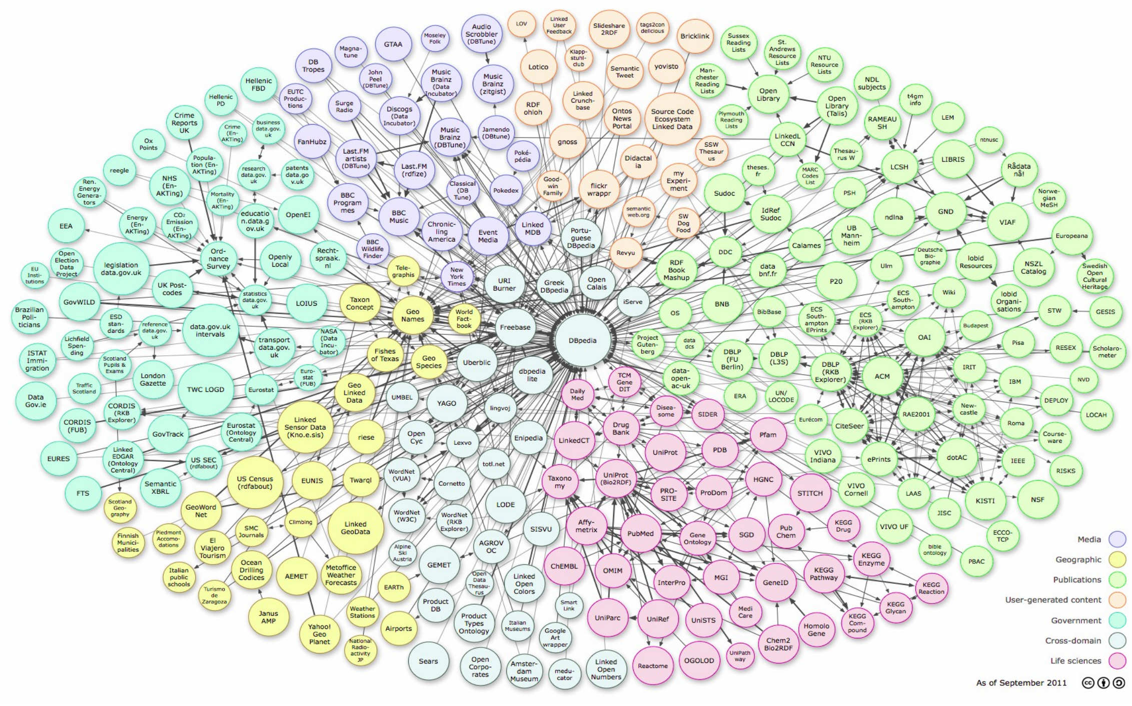
Linked Open Data

LINKINGOPENDATA
W3C SWEO Community Project

- Datasets are represented in RDF
  - Wikipedia, Wikibooks, Geonames, MusicBrainz, WordNet, DBLP bibliography
- Number of triples: 33 Giga (10^9) (2011)
- Governments:
  - USA, UK, Japan,

    Austria, Belgium,

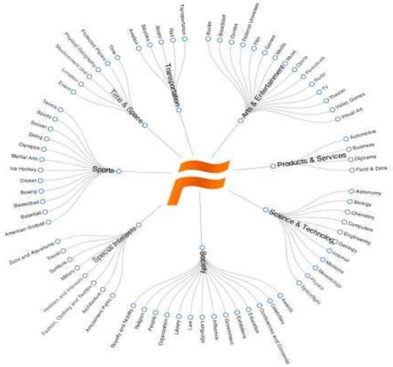
    France, Germany, ...
- Active community



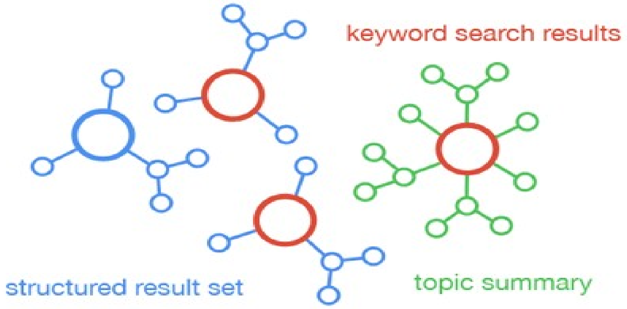
http://en.wikipedia.org/wiki/Open_Data

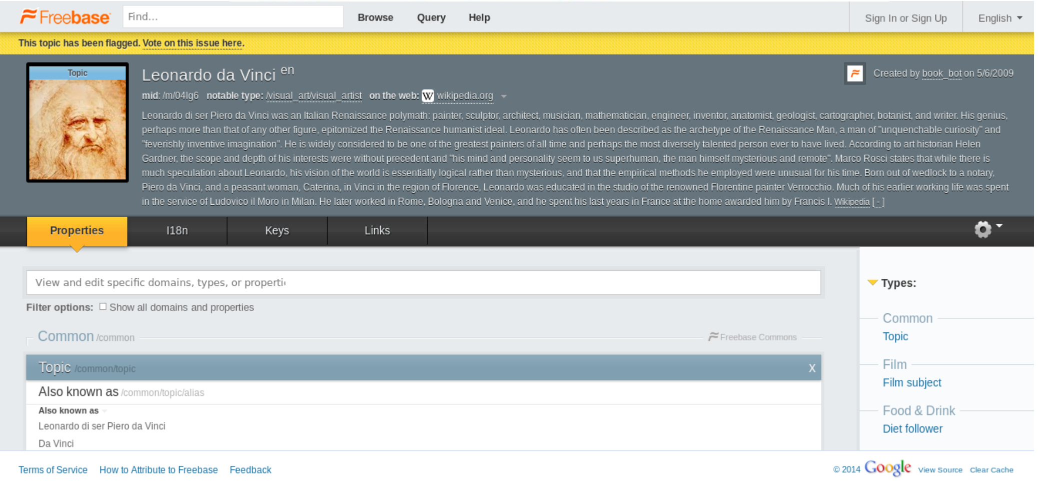
http://www.w3.org/LOD

Freebase



- Free, knowledge graph:
  - people, places and things,
  - 2,478,168,612 facts, 43,459,442 topics
- Semantic search engines are here !





This topic has been flagged. Vote on this issue here.

Leonardo da Vinci en

Created by book_bot on 5/6/2009

mid: /m/04lg6 notable type: /visual_art/visual_artist on the web: wikipedia.org

Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer. His genius, perhaps more than that of any other figure, epitomized the Renaissance humanist ideal. Leonardo has often been described as the archetype of the Renaissance Man, a man of "unquenchable curiosity" and "feverishly inventive imagination". He is widely considered to be one of the greatest painters of all time and perhaps the most diversely talented person ever to have lived. According to art historian Helen Gardner, the scope and depth of his interests were without precedent and "his mind and personality seem to us superhuman, the man himself mysterious and remote". Marco Rosci states that while there is much speculation about Leonardo, his vision of the world is essentially logical rather than mysterious, and that the empirical methods he employed were unusual for his time. Born out of wedlock to a notary, Piero da Vinci, and a peasant woman, Caterina, in Vinci in the region of Florence, Leonardo was educated in the studio of the renowned Florentine painter Verrocchio. Much of his earlier working life was spent in the service of Ludovico il Moro in Milan. He later worked in Rome, Bologna and Venice, and he spent his last years in France at the home awarded him by Francis I. Wikipedia [-]

View and edit specific domains, types, or properti

Filter options: ☐ Show all domains and properties

Common /common Freebase Commons

Topic /common/topic x

Also known as /common/topic/alias

Also known as

Leonardo di ser Piero da Vinci

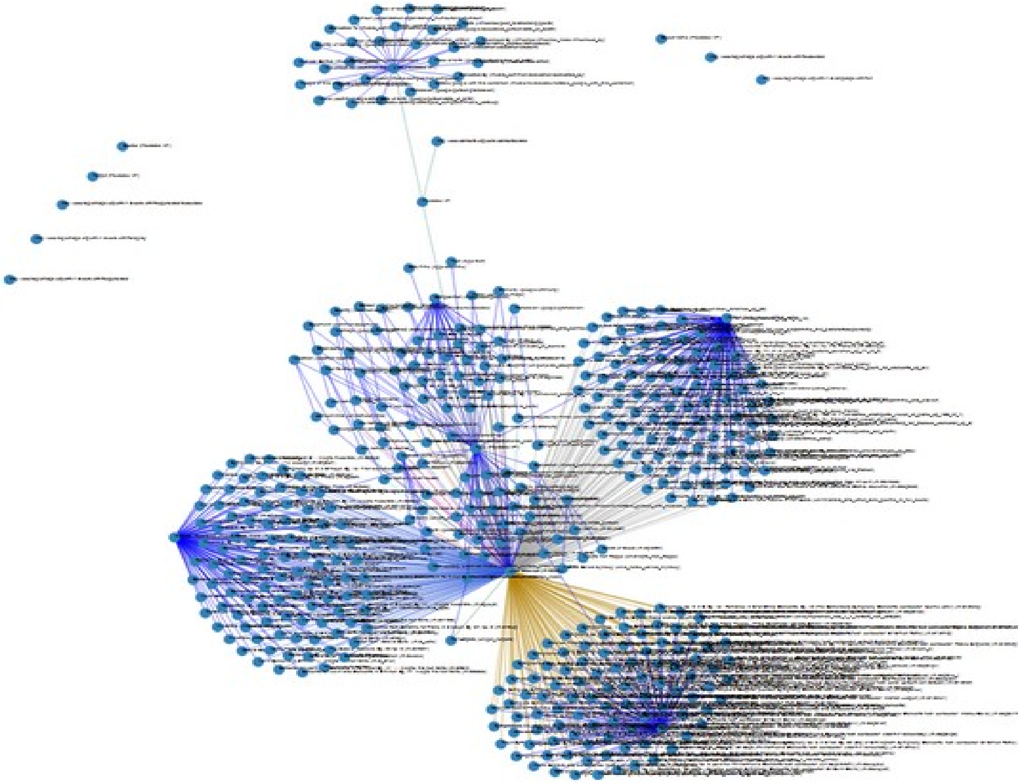
Da Vinci

Types:

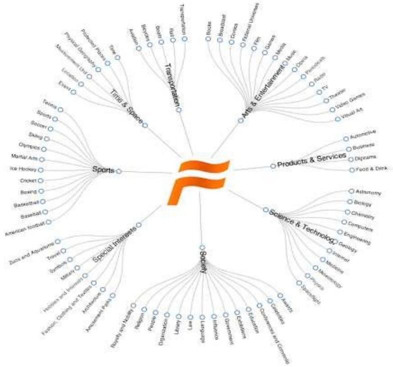
Common
Topic

Film
Film subject

Food & Drink
Diet follower



Freebase

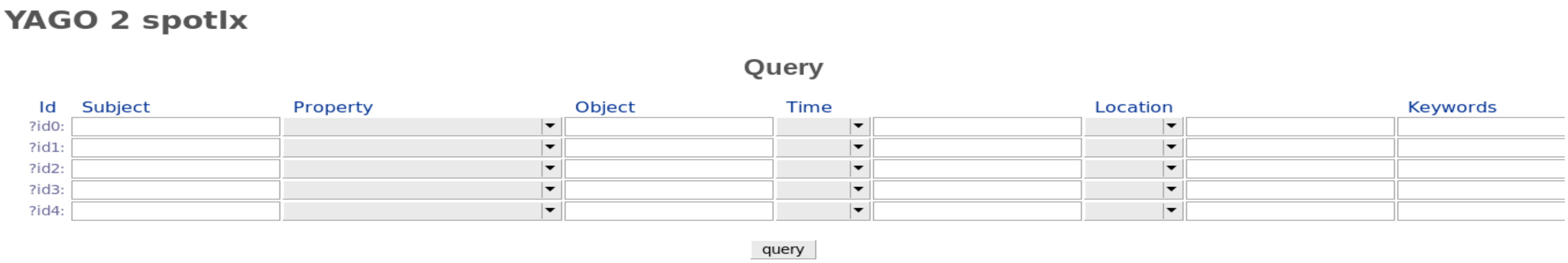


- Based on graphs:
  - nodes, links, types, properties, namespaces
- Google use of Freebase

  - Knowledge graph
  - Words become concepts
  - Semantic questions
  - Semantic associations
  - Browsing knowledge
  - Knowledge engine
- Available in RDF

YAGO

- 10 Mega (10^6) concepts
  - Max Planc Institute, Informatik
  - Accuracy of 95%
- Includes:
  - Wikipedia, WordNet, GeoNames
  - Links Wordnet to Wikipedia taxonomy (350K concepts)
  - Anchored in time and space



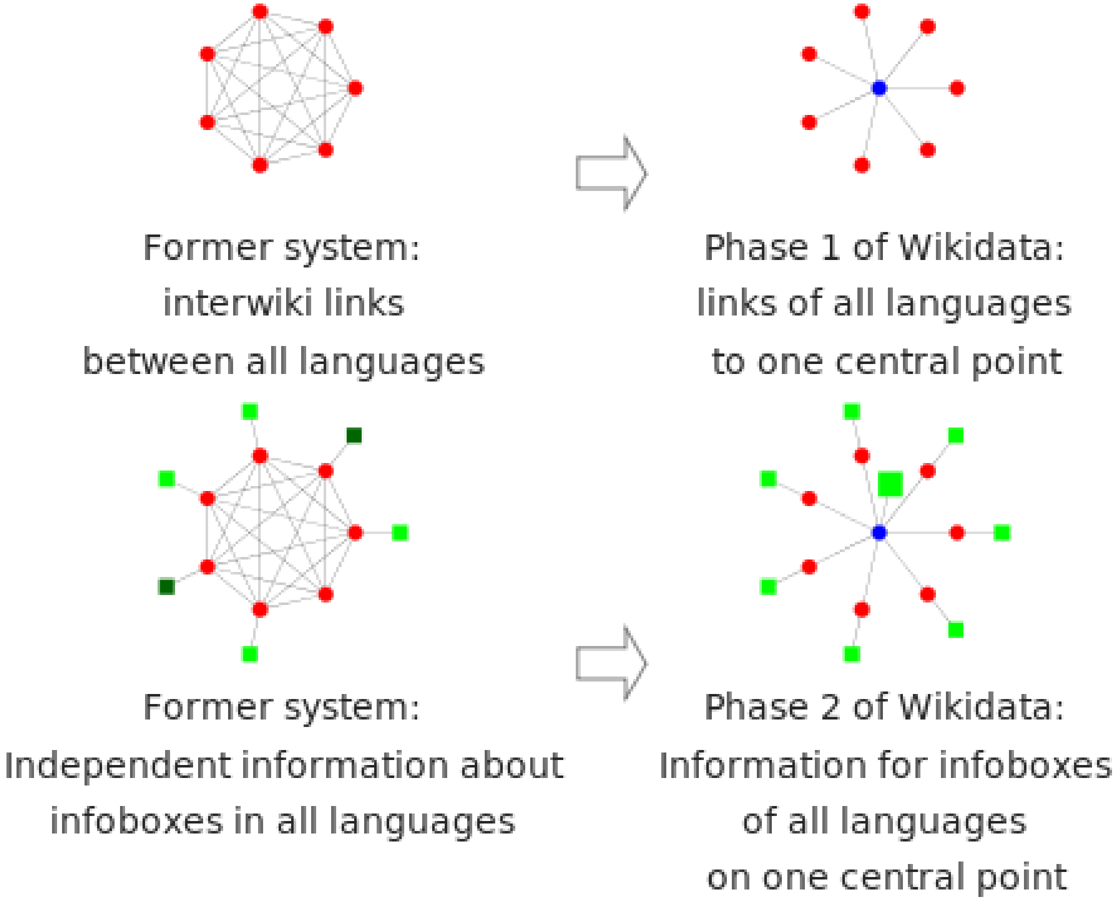
YAGO 2 spotlx

Query

Id	Subject	Property	Object	Time	Location	Keywords
?id0:						
?id1:						
?id2:						
?id3:						
?id4:						

Wikidata

- Free knowledge base with 14,550,852 items
- Collecting structured data
- Properties of
  - person, organization, works, events, etc.

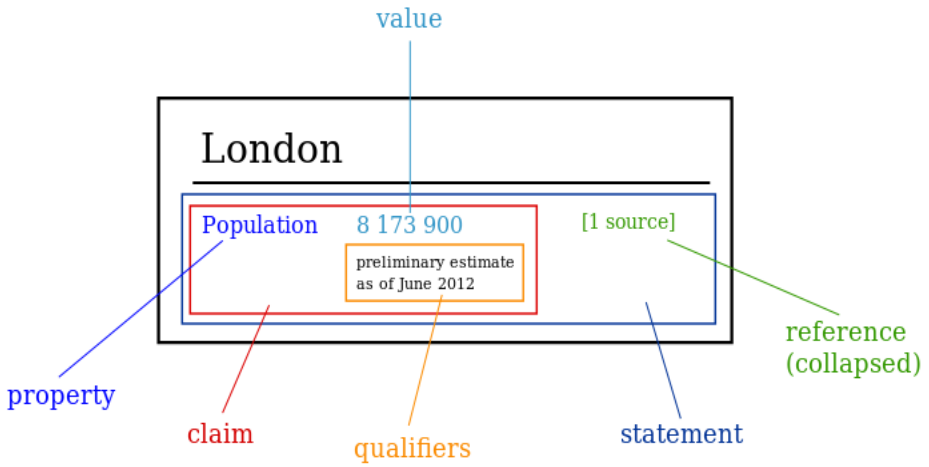


Former system: interwiki links between all languages

Phase 1 of Wikidata: links of all languages to one central point

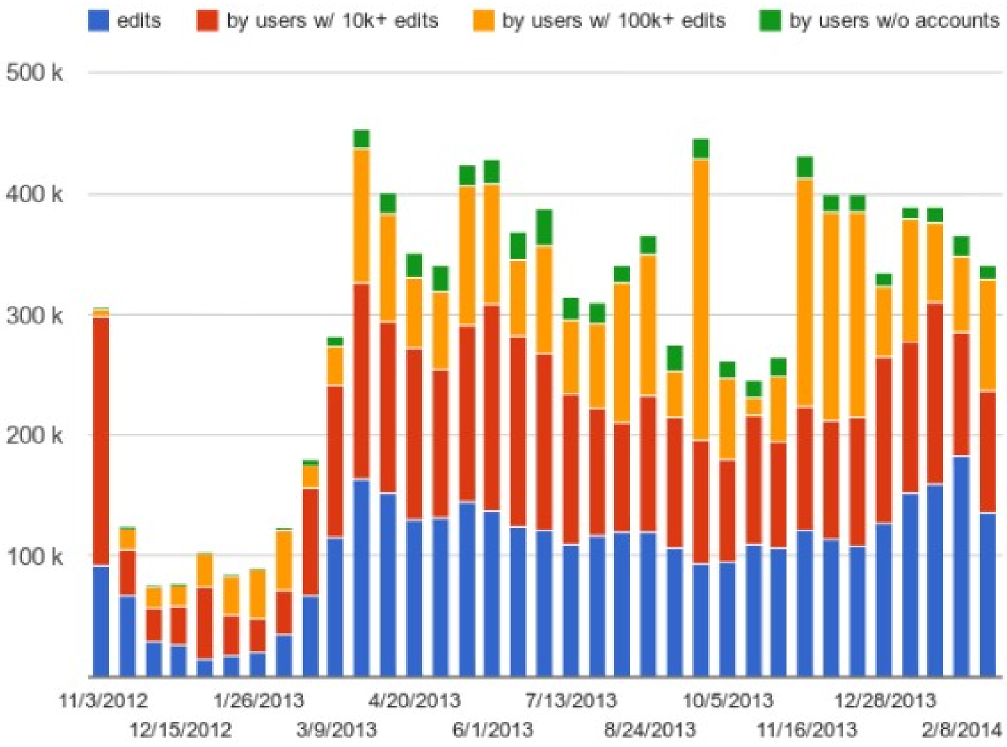
Former system: Independent information about infoboxes in all languages

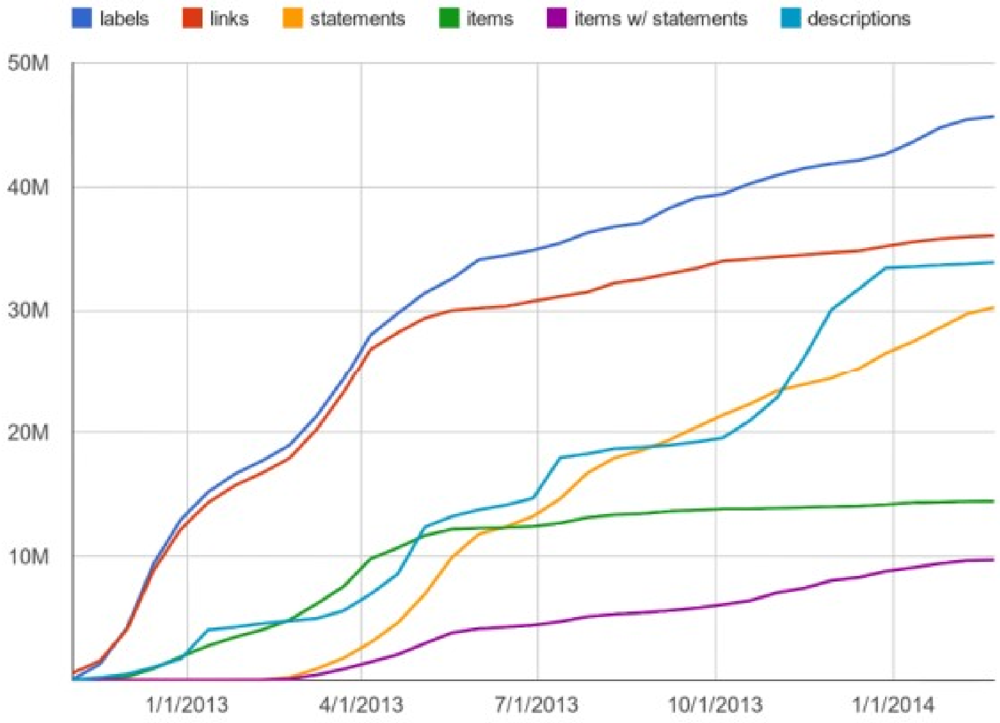
Phase 2 of Wikidata: Information for infoboxes of all languages on one central point



Wikidata

- Free knowledge base with 14,550,852 items





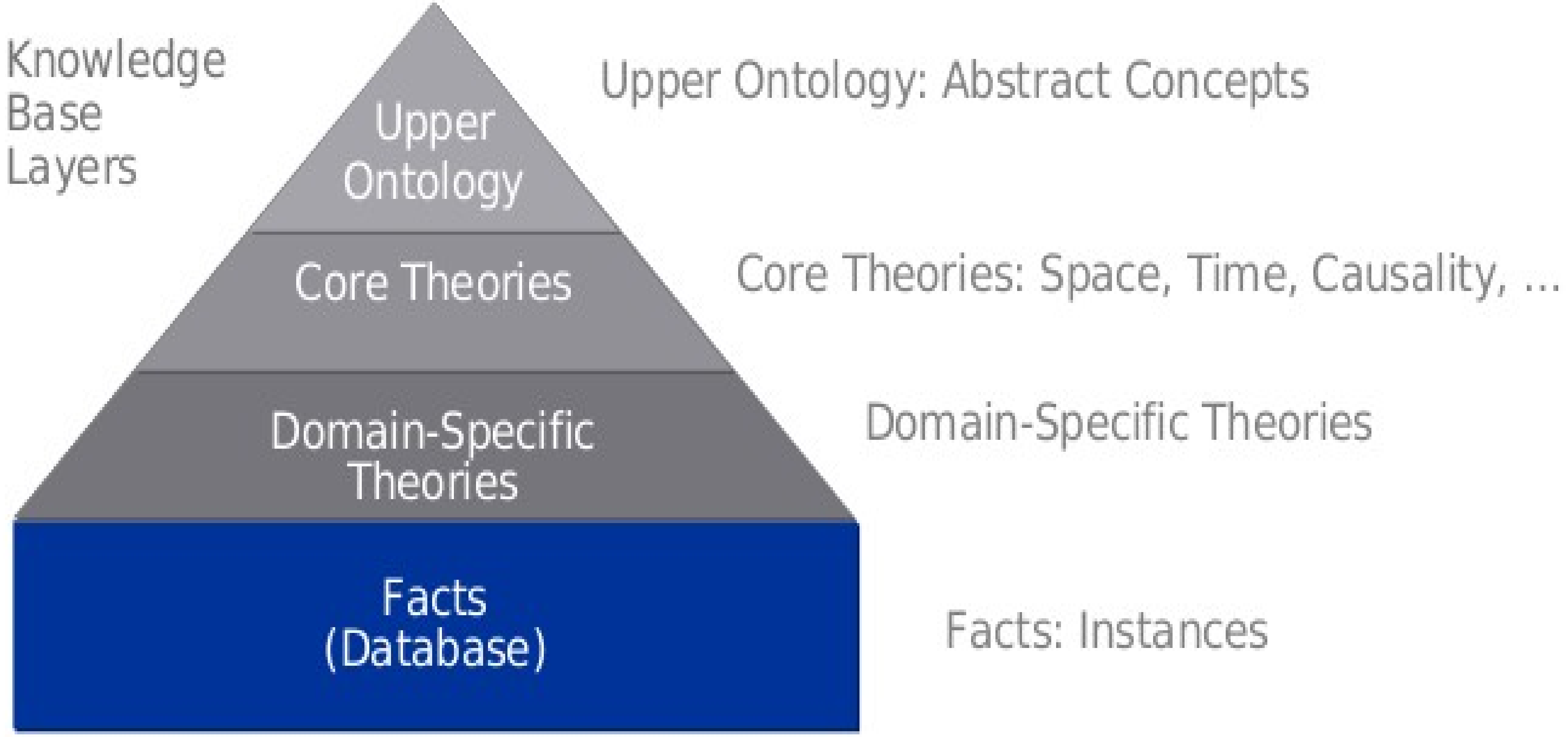
Cycorp®

Home of smarter solutions

Cyc - knowledge base

- Knowledge base
  - Doug Lenat
  - Conceptual networks (ontologies)
  - Higher ontology, basic theories, specific theories
  - Predefined semantic relationships
- Common sense reasoner
  - Based on predicate calculus
  - Rule-based reasoning

Cyc


Knowledge
Base
Layers
Upper
Ontology
Core Theories
Domain-Specific
Theories
Facts
(Database)
Upper Ontology: Abstract Concepts
Core Theories: Space, Time, Causality, ...
Domain-Specific Theories
Facts: Instances

Some conclusions

- There exist a variety of different dictionaries, properties, concepts, ...
  - Common definitions are not frequent
- There exist a variety of formats and models for knowledge and data representation
  - RDF is common data/knowledge model
- Senses of words are not represented