Semantic Web Technologies I

Lehrveranstaltung im WS12/13

PD Dr. Sebastian Rudolph

Dr. Duc Thanh Tran

M.Sc. Anees ul Mehdi

Overview

- XML und URIs
- Einleitung in RDF
- RDF Schema
- Logik – Grundlagen
- Semantik von RDF(S)
- SPARQL – Syntax und Intuition
- Semantik von SPARQL
- **Linked Data**
- Semantic Search
- OWL – Syntax und Intuition I
- OWL – Syntax und Intuition II
- OWL – Semantik und Reasoning
- Konjunktive Anfragen und Regelsprachen
- Applications

Agenda

- Introduction
- Semantic Web data
  - The RDF data model
  - Publishing RDF
  - Last Lecture: crawling and indexing RDF data
- Query processing / matching
  - Last lecture: selected problems in structured query (SPARQL) processing
  - Here: big picture of querying with structured and keyword queries
- Ranking
- Result presentation

Why Semantic Search? I.

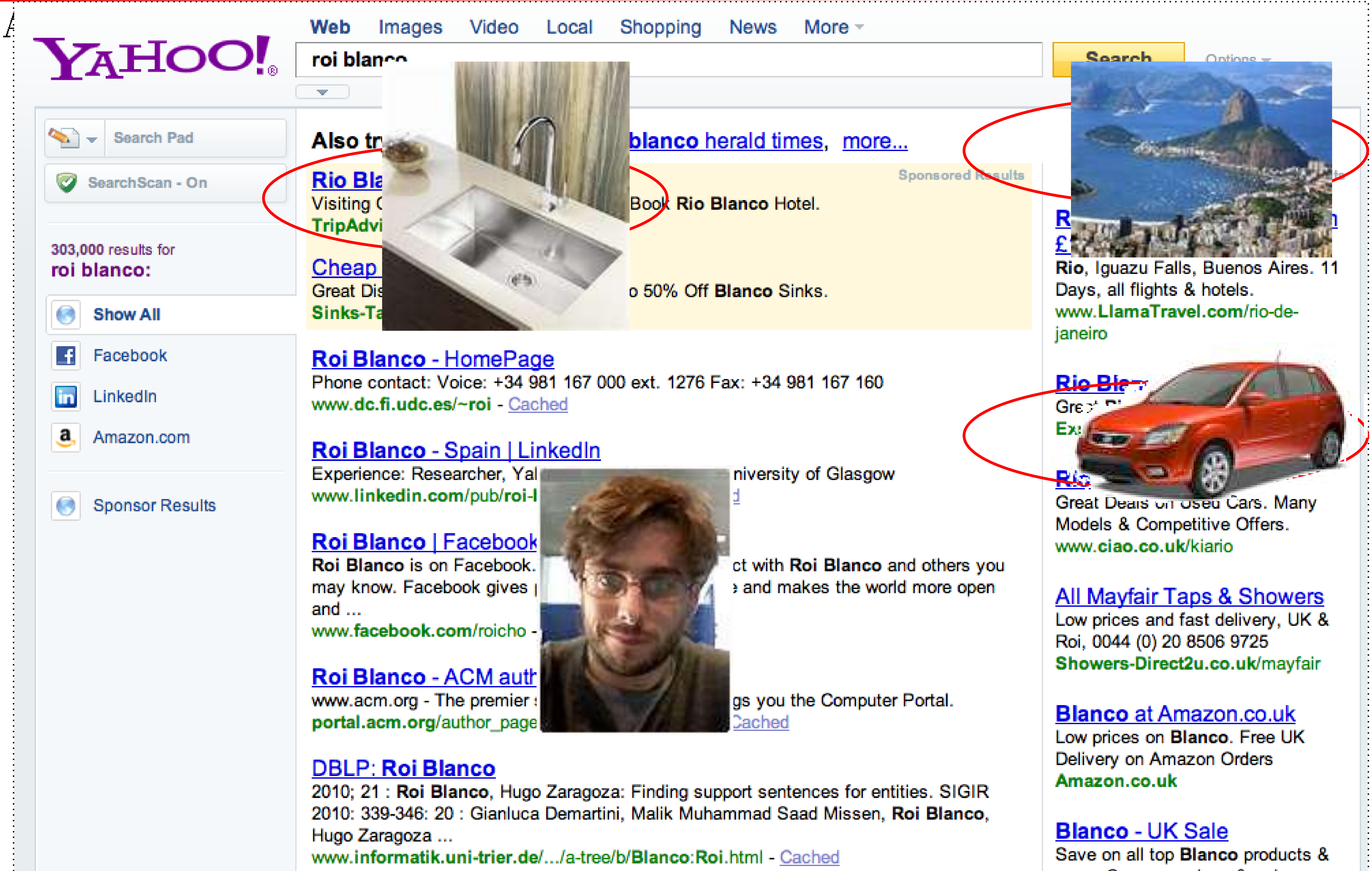
- “We are at the beginning of search.“ (Marissa Mayer)
  - Solved large classes of queries, e.g. navigational
  - Heavy investment in computational power
  - Remaining queries are hard, not solvable by brute force, and require a deep understanding of the world and human cognition
- Background knowledge and metadata can help to address poorly solved queries

Poorly solved information needs

- Ambiguous searches
  - paris hilton
- Long tail queries
  - george bush (and I mean the beer bre
- Multimedia search
  - paris hilton sexy
- Imprecise or overly precise searches
  - jim hendler
  - pictures of strong adventures people
- Precise searches for descriptions
  - countries in africa
  - 32 year old computer scientist living in barcelona
  - reliable digital camera under 300 dollars

Many of these queries would not be asked by users, who learned over time what search technology can and can not do.

Example: multiple interpretations



roi blanco

303,000 results for
roi blanco:

Also tr… blanco herald times, more...

Rio Bla…
Visiting … Book Rio Blanco Hotel.
TripAdvi…

Cheap …
Great Dis… o 50% Off Blanco Sinks.
Sinks-Ta…

Roi Blanco - HomePage
Phone contact: Voice: +34 981 167 000 ext. 1276 Fax: +34 981 167 160
www.dc.fi.udc.es/~roi - Cached

Roi Blanco - Spain | LinkedIn
Experience: Researcher, Yal… niversity of Glasgow
www.linkedin.com/pub/roi-l… d

Roi Blanco | Facebook
Roi Blanco is on Facebook. … ct with Roi Blanco and others you may know. Facebook gives … and makes the world more open and ...
www.facebook.com/roicho - …

Roi Blanco - ACM auth…
www.acm.org - The premier … gs you the Computer Portal.
portal.acm.org/author_page… Cached

DBLP: Roi Blanco
2010; 21 : Roi Blanco, Hugo Zaragoza: Finding support sentences for entities. SIGIR 2010: 339-346: 20 : Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, Hugo Zaragoza ...
www.informatik.uni-trier.de/…/a-tree/b/Blanco:Roi.html - Cached

R…
£…
Rio, Iguazu Falls, Buenos Aires. 11 Days, all flights & hotels.
www.LlamaTravel.com/rio-de-janeiro

Rio Bla…
Gre…
Ex…

Ric…
Great Deals on Used Cars. Many Models & Competitive Offers.
www.ciao.co.uk/kiario

All Mayfair Taps & Showers
Low prices and fast delivery, UK & Roi, 0044 (0) 20 8506 9725
Showers-Direct2u.co.uk/mayfair

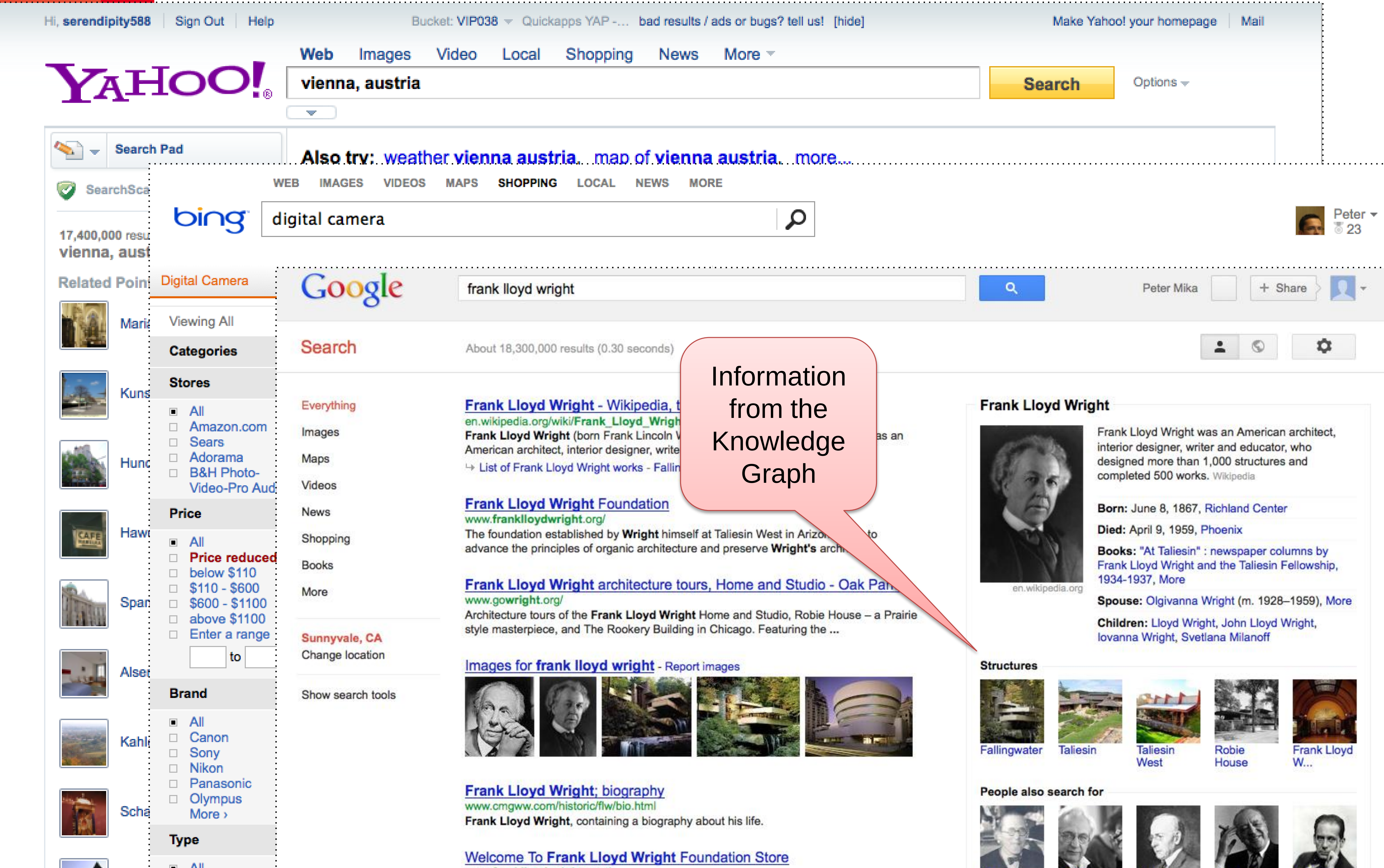
Blanco at Amazon.co.uk
Low prices on Blanco. Free UK Delivery on Amazon Orders
Amazon.co.uk

Blanco - UK Sale
Save on all top Blanco products &

Why Semantic Search? II.

- The Semantic Web is now a reality
  - Large amounts of data published in RDF
  - Heterogeneous data of varying quality
  - Users who are not skilled in writing complex queries (e.g. SPARQL) and may not be experts in the domain
- Searching data instead or in addition to searching documents
  - Direct answers
  - Novel search tasks

Example: direct answers in search



Hi, serendipity588 | Sign Out | Help   Bucket: VIP038 ▾ Quickapps YAP -... bad results / ads or bugs? tell us! [hide]   Make Yahoo! your homepage | Mail

Web Images Video Local Shopping News More ▾

YAHOO! vienna, austria [Search] Options ▾

Search Pad

Also try: weather vienna austria, map of vienna austria, more...

SearchSca

17,400,000 resu

vienna, aust

Related Poin

Mari
Kuns
Hund
Haw
Span
Alser
Kahl
Scha

WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE

bing digital camera

Peter ▾
23

Digital Camera

Viewing All

Categories

Stores

- ▣ All
- □ Amazon.com
- □ Sears
- □ Adorama
- □ B&H Photo-Video-Pro Aud

Price

- ▣ All
- □ Price reduced
- □ below $110
- □ $110 - $600
- □ $600 - $1100
- □ above $1100
- □ Enter a range
- [ ] to [ ]

Brand

- ▣ All
- □ Canon
- □ Sony
- □ Nikon
- □ Panasonic
- □ Olympus
- More ›

Type

- All

Google frank lloyd wright   Peter Mika + Share

Search   About 18,300,000 results (0.30 seconds)

Everything
Images
Maps
Videos
News
Shopping
Books
More

Sunnyvale, CA
Change location

Show search tools

Frank Lloyd Wright - Wikipedia, t
en.wikipedia.org/wiki/Frank_Lloyd_Wrigh
Frank Lloyd Wright (born Frank Lincoln W ... as an American architect, interior designer, write
↳ List of Frank Lloyd Wright works - Fallin

Frank Lloyd Wright Foundation
www.franklloydwright.org/
The foundation established by Wright himself at Taliesin West in Arizo... to advance the principles of organic architecture and preserve Wright's arch

Frank Lloyd Wright architecture tours, Home and Studio - Oak Par
www.gowright.org/
Architecture tours of the Frank Lloyd Wright Home and Studio, Robie House – a Prairie style masterpiece, and The Rookery Building in Chicago. Featuring the ...

Images for frank lloyd wright - Report images

Frank Lloyd Wright: biography
www.cmgww.com/historic/flw/bio.html
Frank Lloyd Wright, containing a biography about his life.

Welcome To Frank Lloyd Wright Foundation Store

Frank Lloyd Wright

Frank Lloyd Wright was an American architect, interior designer, writer and educator, who designed more than 1,000 structures and completed 500 works. Wikipedia

en.wikipedia.org

Born: June 8, 1867, Richland Center

Died: April 9, 1959, Phoenix

Books: "At Taliesin" : newspaper columns by Frank Lloyd Wright and the Taliesin Fellowship, 1934-1937, More

Spouse: Olgivanna Wright (m. 1928–1959), More

Children: Lloyd Wright, John Lloyd Wright, Iovanna Wright, Svetlana Milanoff

Structures

Fallingwater | Taliesin | Taliesin West | Robie House | Frank Lloyd W...

People also search for

Information from the Knowledge Graph

Document retrieval and data retrieval

- Information Retrieval (IR) support the retrieval of documents (document retrieval)
  - Representation based on lightweight syntax-centric models
  - Work well for topical search
  - Not so well for more complex information needs
  - Web scale
- Database (DB) and Knowledge-based Systems (KB) deliver more precise answers (data retrieval)
  - More expressive models
  - Allow for complex queries
  - Retrieve concrete answers that precisely match queries
  - Not just matching and filtering, but also joins
  - Limitations in scalability

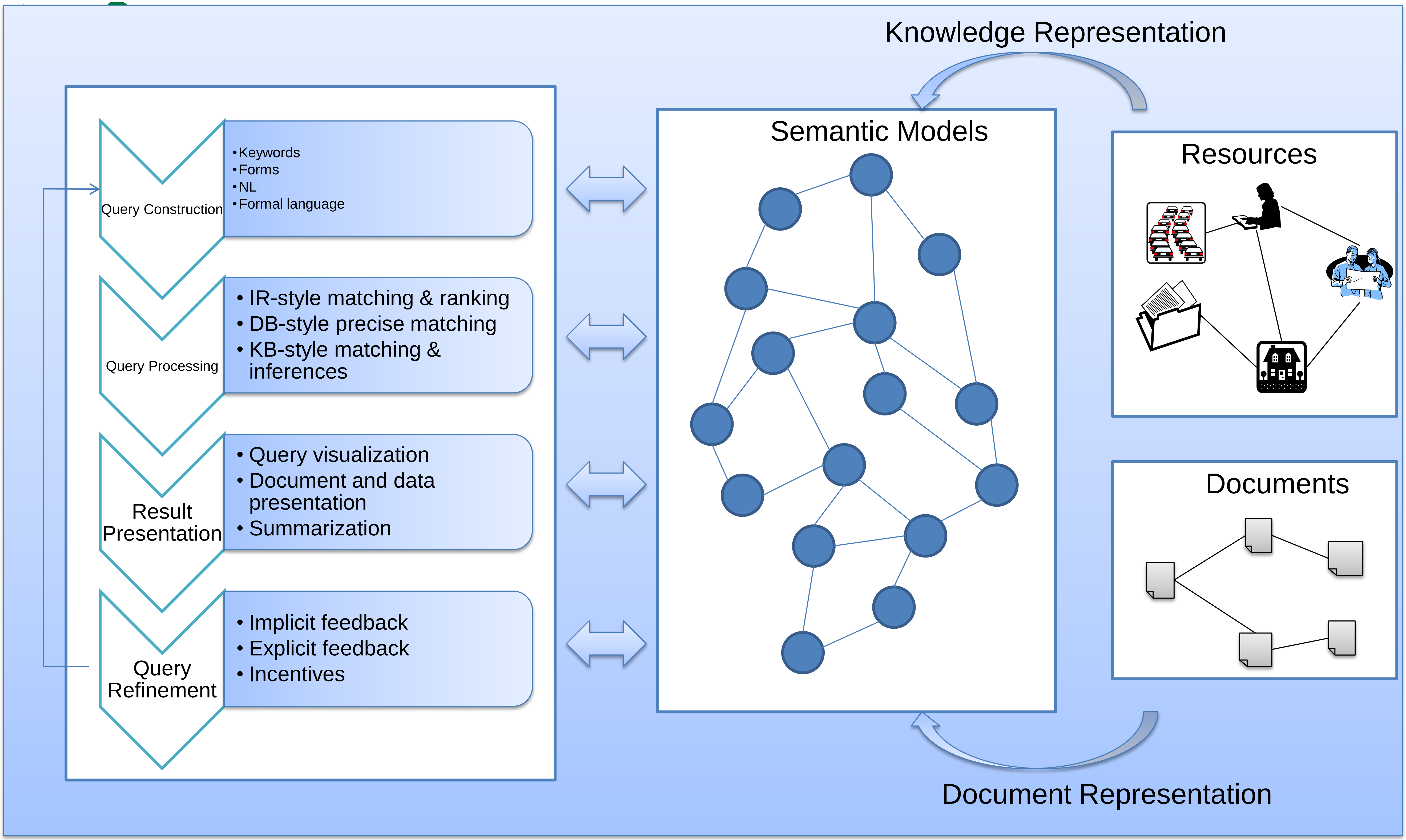
Combination of document and data retrieval

- Documents with metadata
    - Metadata may be embedded inside the document
    - *I'm looking for* ***documents*** *that mention countries in Africa.*
- Data retrieval
    - Structured data, but searchable text fields
    - *I'm looking for* ***directors****, who have directed movies where the synopsis mentions dinosaurs.*

Semantic Search

- Target (combination of) document and data retrieval
- Semantic search is a retrieval paradigm that
  - Exploits the structure/semantics of the data or explicit background knowledge to understand user intent and the meaning of content
  - Incorporates the intent of the query and the meaning of content into the search process (**semantic models**)
- Wide range of semantic search systems
  - Employ different semantic models**,** possibly at **different steps** of the search process and in order to support **different tasks**

Semantic Search – a process view



Knowledge Representation

Query Construction
- Keywords
- Forms
- NL
- Formal language

Query Processing
- IR-style matching & ranking
- DB-style precise matching
- KB-style matching & inferences

Result Presentation
- Query visualization
- Document and data presentation
- Summarization

Query Refinement
- Implicit feedback
- Explicit feedback
- Incentives

Semantic Models

Resources

Documents

Document Representation

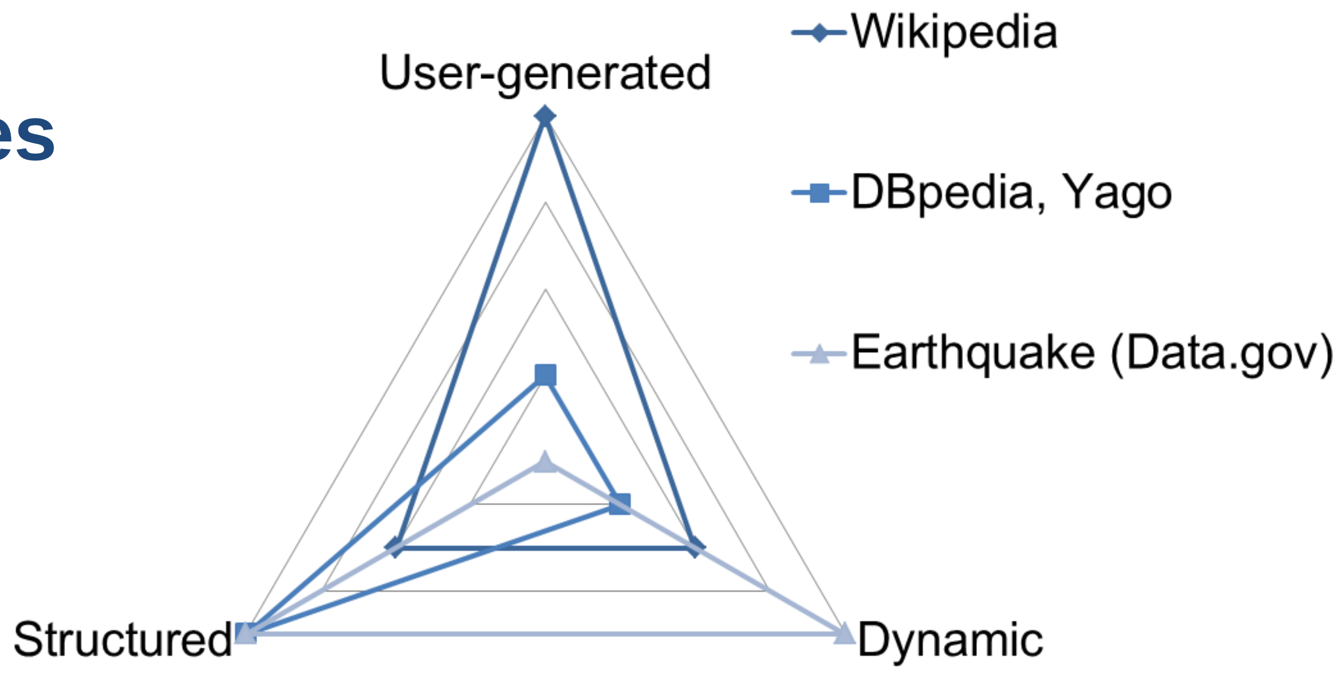
Semantic Search systems

For **data / document** retrieval, semantic search systems might combine a range of techniques, ranging from statistics-based **IR methods** for **ranking**, **database methods** for efficient **indexing** and **query processing**, up to complex **reasoning** techniques for making inferences!

Repetition: Information Workbench

- Addressing the **lifecycle of interacting** with the Web of Data
  - Integration of data sources
  - Content generation by the end user
  - **Search and Exploration**
  - **Visualization**
  - Publishing

- Integrated management of **heterogeneous data sources**
  - Structured and unstructured
  - Published and user-generated
  - Static and dynamic
  - Open domain

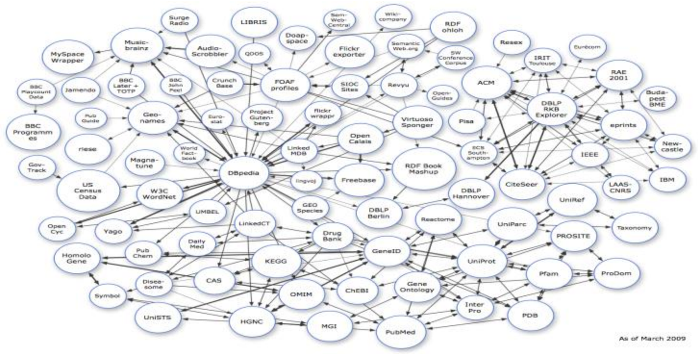


Data Sources in the Application

- Entire English Wikipedia

WIKIPEDIA
The Free Encyclopedia

- Data from Linked Open Data
  - DBpedia
  - YAGO
  - …



DBpedia

- Data from Data.gov (US Government)
  - E.g. live data about earthquakes

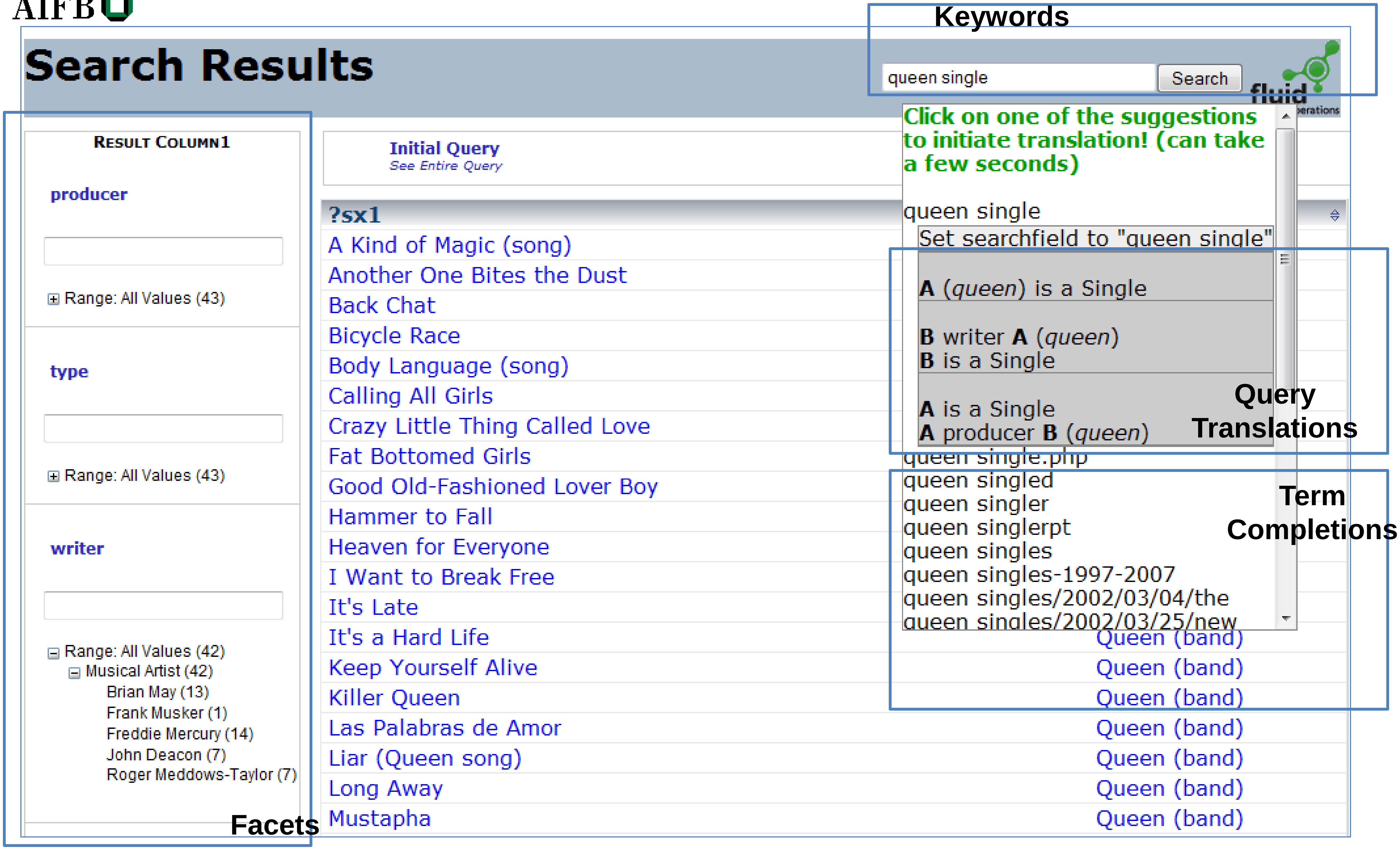
DATA-GOV
TWC

- Many more

Semantic Search

- **Hybrid Search**: Structured queries combined with keywords across structured and unstructured data sources
- **Query interpretation:** Translation of keywords into hybrid queries
- **Keyword search/query interpretation** combined with **faceted search**: iterative refinement process based on keywords and operations on facets

Search, Refinement and Navigation



Keywords

Search Results

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 Translations

queen single.php
queen singled
queen singler
queen singlerpt
queen singles
queen singles-1997-2007
queen singles/2002/03/04/the
queen singles/2002/03/25/new

Term Completions

Result Column1

producer

⊞ Range: All Values (43)

type

⊞ Range: All Values (43)

writer

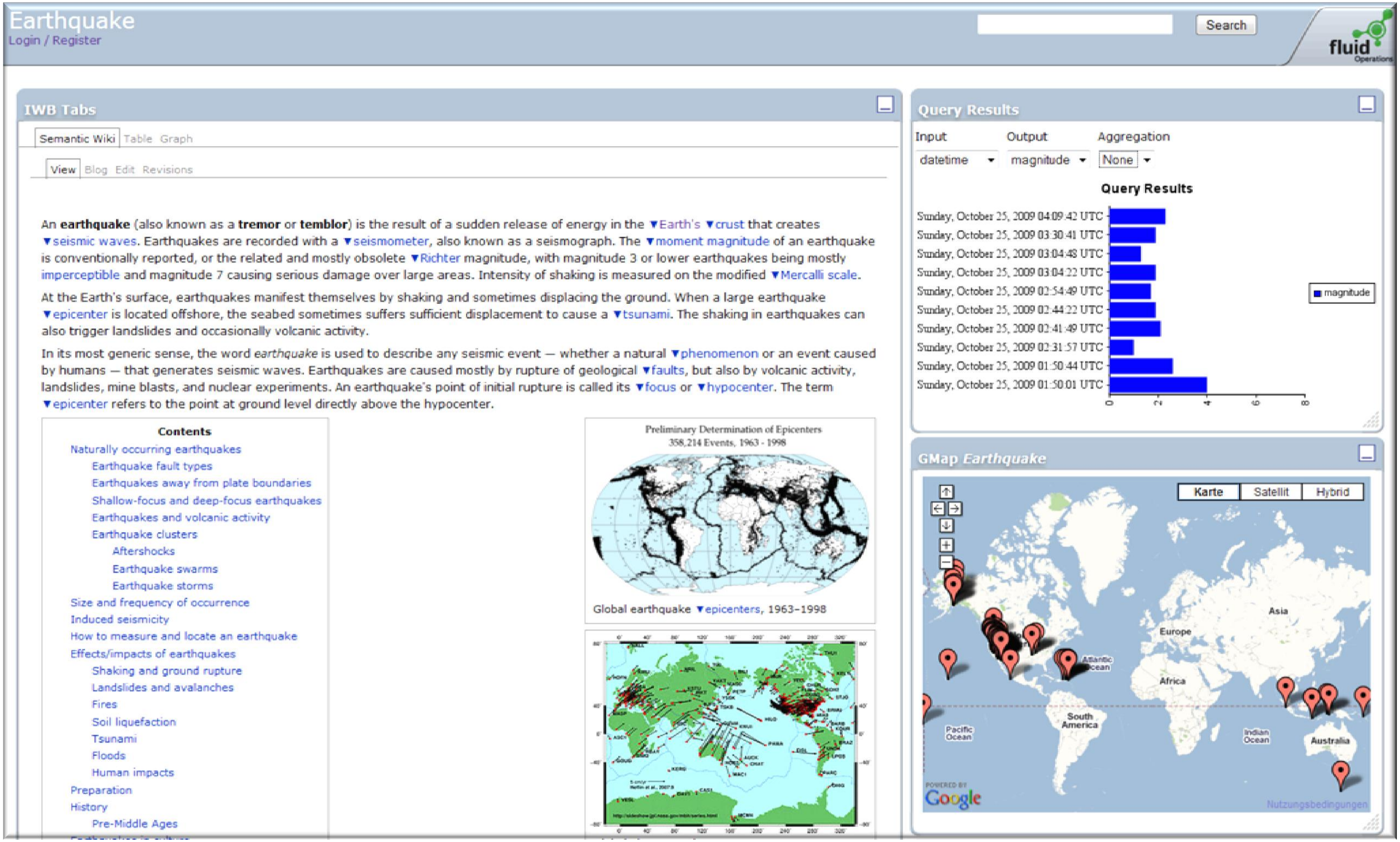
⊟ Range: All Values (42)
⊟ Musical Artist (42)
Brian May (13)
Frank Musker (1)
Freddie Mercury (14)
John Deacon (7)
Roger Meddows-Taylor (7)

Facets

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	Queen (band)
Keep Yourself Alive	Queen (band)
Killer Queen	Queen (band)
Las Palabras de Amor	Queen (band)
Liar (Queen song)	Queen (band)
Long Away	Queen (band)
Mustapha	Queen (band)

Result Inspection, Analysis and Browsing



Earthquake
Login / Register

Search

IWB Tabs

Semantic Wiki Table Graph

View Blog Edit Revisions

An **earthquake** (also known as a **tremor** or **temblor**) 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

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

Global earthquake ▼epicenters, 1963–1998

Query Results

Input: datetime
Output: magnitude
Aggregation: None

Query Results

Sunday, October 25, 2009 04:09:42 UTC
Sunday, October 25, 2009 03:30:41 UTC
Sunday, October 25, 2009 03:04:48 UTC
Sunday, October 25, 2009 03:04:22 UTC
Sunday, October 25, 2009 02:54:49 UTC
Sunday, October 25, 2009 02:44:22 UTC
Sunday, October 25, 2009 02:41:49 UTC
Sunday, October 25, 2009 02:31:57 UTC
Sunday, October 25, 2009 01:50:44 UTC
Sunday, October 25, 2009 01:50:01 UTC
0
2
4
6
8
■ magnitude

GMap *Earthquake*

Semantic Web data

Data on the Web

- Data on the Web is not directly accessible
  - Most web pages are generated from databases, but formatted for human consumption
  - APIs offer limited views over data
- Two solutions
  - Extraction using **Information Extraction** (IE) techniques
    - Out of scope for this tutorial
  - Relying on publishers to expose structured data using standard **Semantic Web** formats

Semantic Web

- Sharing data across the Web
  - Standard data model
    - ✦ RDF
  - A number of syntaxes (file formats)
    - ✦ RDF/XML, RDFa
  - Powerful, logic-based languages for schemas
    - ✦ OWL, RIF
  - Query languages and protocols
    - ✦ HTTP, SPARQL

Publishing RDF

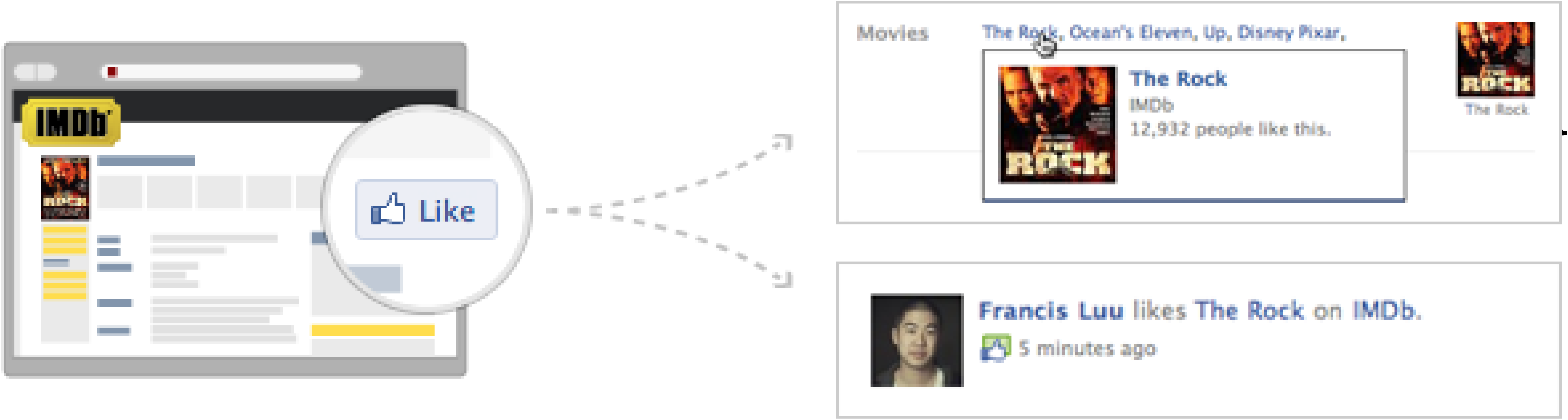
- Interlinked RDF documents (Linked Data)
  - Each document describes a single resource with URIs pointing to related resources
  - Common RDF file formats are RDF/XML and Turtle
  - Mostly implemented as a wrapper around a database or Web service
- Embedding RDF inside HTML
  - RDFa, microdata
- SPARQL endpoints
  - Triple stores are databases for managing RDF data
  - SPARQL is a standard protocol and query language for accessing triple stores using HTTP

Example ontologies: schema.org

- Agreement on a shared set of schemas for common types of web content
  - Bing, Google, and Yahoo! as initial supporters
  - Similar in intent to sitemaps.org (2006)
    - Use a single format to communicate the same information to all three search engines
- Support for microdata
- schema.org covers areas of interest to all search engines
  - Business listings (local), creative works (video), recipes, reviews
  - User defined extensions
- Each search engine continues to develop its products

Example: Facebook's Open Graph Protocol

- The 'Like' button provides publishers with a way to promote their content on Facebook and build communities
  - Shows up in profiles and news feed
  - Site owners can later reach users who have liked an object
  - Facebook Graph API allows 3rd party developers to



12,932 people like this.

Francis Luu likes The Rock on IMDb.

5 minutes ago

- t

Example: Facebook’s Open Graph Protocol

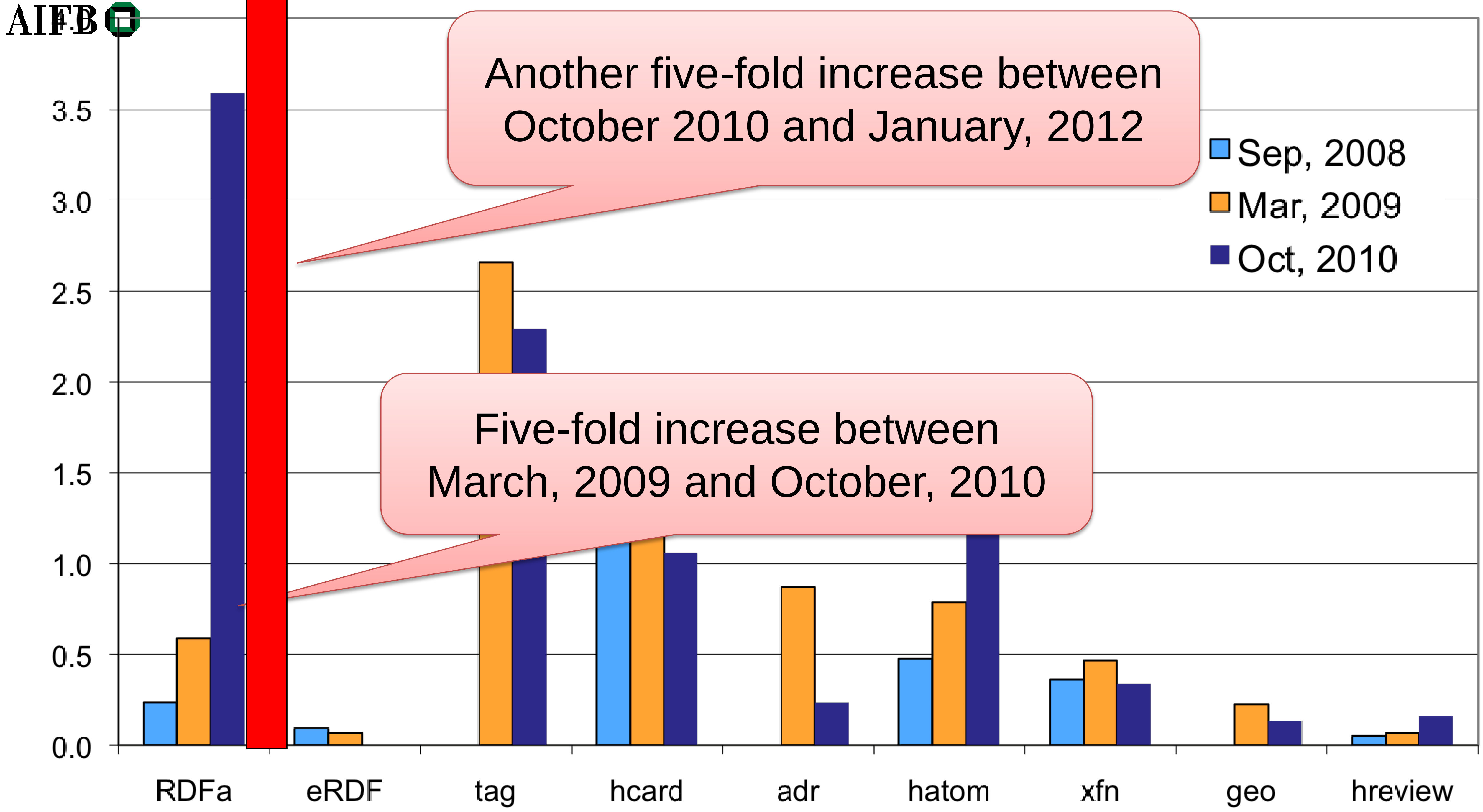
- RDF vocabulary to be used in conjunction with RDFa
  - Simplify the work of developers by restricting the freedom in RDFa
- Activities, Businesses, Groups, Organizations, People, Places, Products and Entertainment
- Only HTML <head> accepted

```
<html xmlns:og="http://opengraphprotocol.org/schema/">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="movie" />
<meta property="og:url"
content="http://www.imdb.com/title/tt0117500/" />
<meta property="og:image" content="http://ia.media-
imdb.com/images/rock.jpg" /> …
</head> ...
```

Current state of metadata on the Web

- 31% of webpages, 5% of domains contain some metadata
  - Analysis of the Bing Crawl (US crawl, January, 2012)
  - RDFa is most common format
    - By URL: 25% RDFa, 7% microdata, 9% microformat
    - By eTLD (PLD): 4% RDFa, 0.3% microdata, 5.4% microformat
  - Adoption is stronger among large publishers
    - Especially for RDFa and microdata
- See also
  - P. Mika, T. Potter. Metadata Statistics for a Large Web Corpus, LDOW 2012
  - H.Mühleisen, C.Bizer.Web Data Commons - Extracting Structured Data from Two Large Web Corpora, LDOW 2012

Exponential growth in RDFa data



Another five-fold increase between October 2010 and January, 2012

Five-fold increase between March, 2009 and October, 2010

Percentage of URLs with embedded metadata in various formats

Query Processing / Matching

Structure

- Taxonomy of search approaches
- Query processing / matching techniques for Semantic Search
- Types of semantic data
- Formalisms for querying semantic data
- Approaches
  - General task: hybrid graph pattern matching
  - Matching keyword query against text
  - Matching structured query against structured data
  - Matching keyword query against structured data
  - Matching structured query against text (a hybrid case)
- Main tasks, challenges and opportunities

Taxonomy of search approaches

- The search problem
  - A collection of resources, called *data*
  - Information needs expressed as *queries*
  - Search is the task of **efficiently computing results** from data that are **relevant** to queries
- **Document** data retrieval vs. **structured data** retrieval
  - Differences in query and data representation and matching
  - Efficiently retrieve structured data that exactly match formal information needs expressed as structured queries
  - Effectively rank textual results that match ambiguous NL / keyword queries to a certain degree (notions of relevance)
- Semantic search: **ranked** retrieval of document and structured data (given **ambiguous** queries / data)

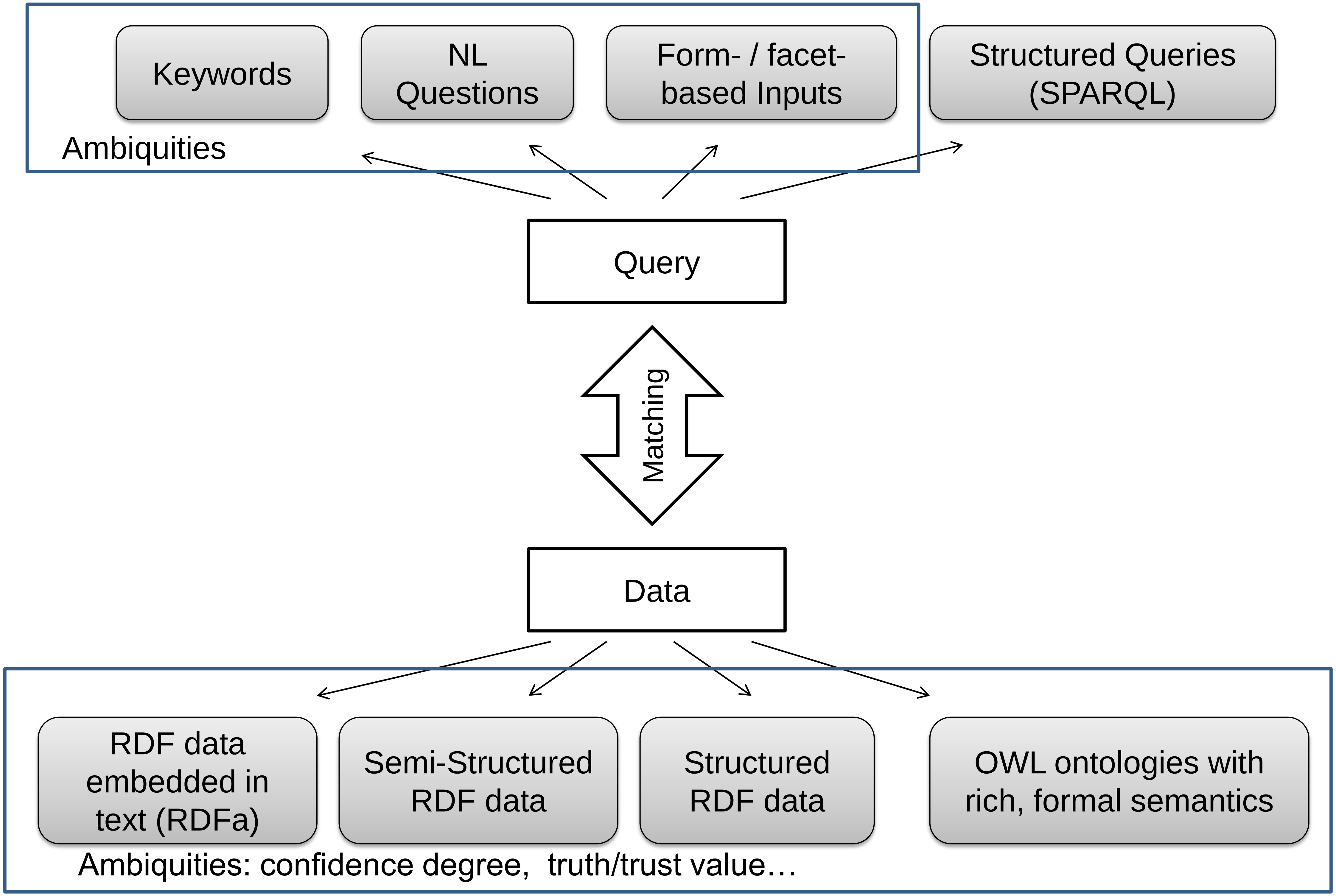
Query processing for Semantic Search (1)

- Resources represented by semantic data ranging from
  - Structured data with well defined schemas
  - Semi-structured data with incomplete or no schemas
  - Data that largely comprise text
  - Hybrid / embedded data
- Information needs of varying complexity, captured using different formalisms and querying paradigms
  - Natural language texts and keywords
  - Form-based inputs
  - Formal structured queries

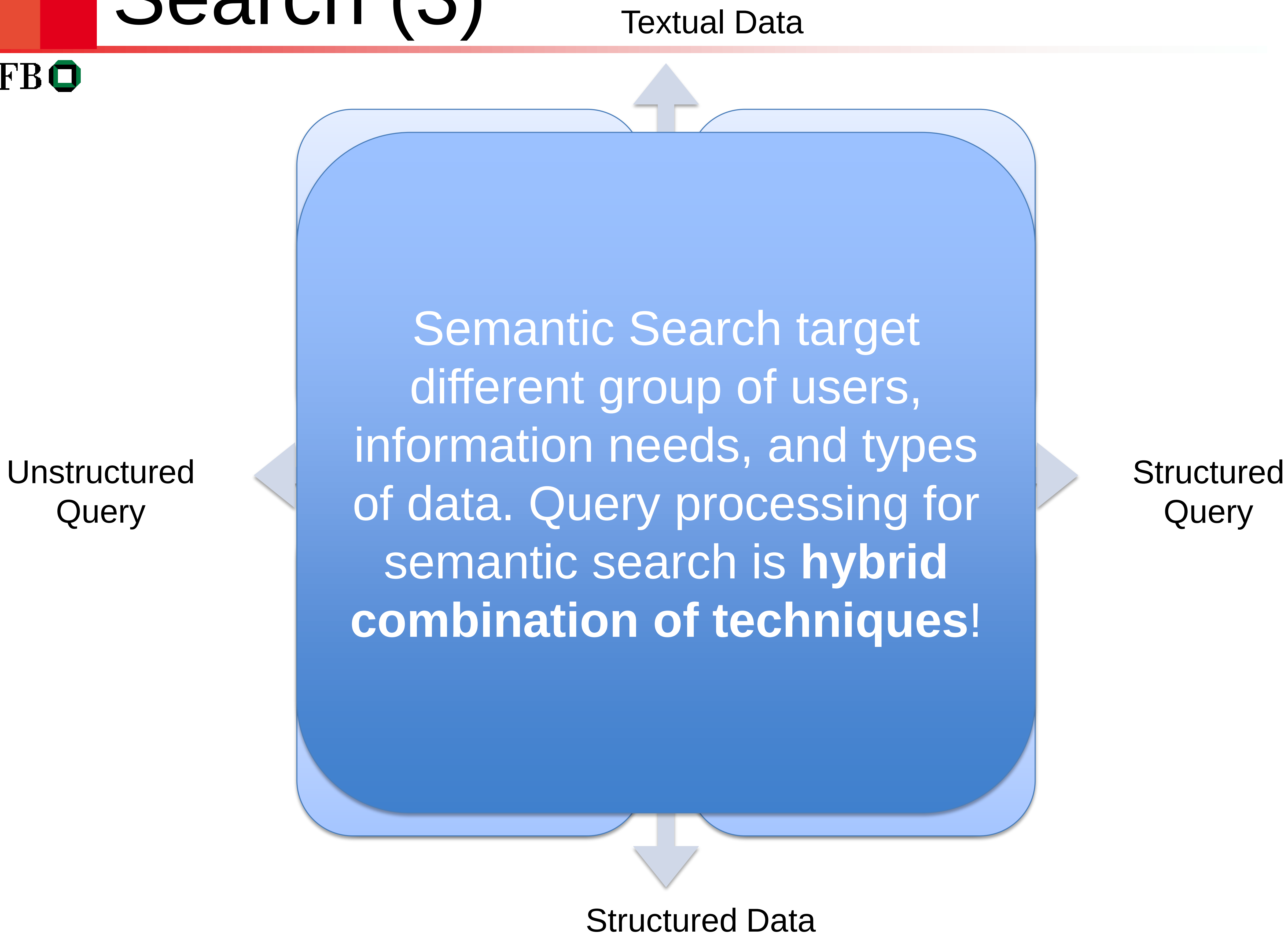
(*Search is end-user oriented paradigm, requires "natural", intuitive querying interfaces)*

- Semantic search: efficiently computing results (**query processing)** from data that are relevant to queries (**ranking**)

Query processing for Semantic Search (2)


Keywords
NL Questions
Form- / facet-based Inputs
Structured Queries (SPARQL)
Ambiquities
Query
Matching
Data
RDF data embedded in text (RDFa)
Semi-Structured RDF data
Structured RDF data
OWL ontologies with rich, formal semantics
Ambiquities: confidence degree, truth/trust value…

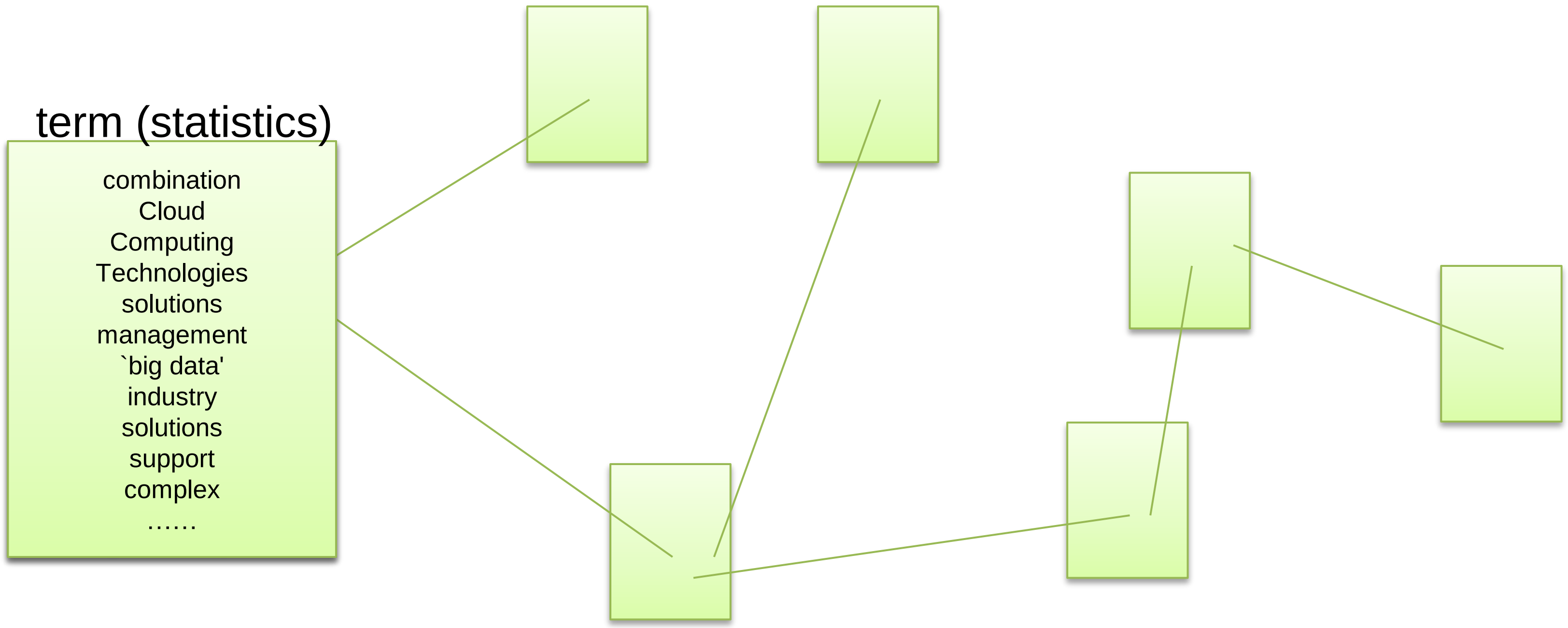
Query processing for Semantic Search (3)



Semantic Search target different group of users, information needs, and types of data. Query processing for semantic search is **hybrid combination of techniques**!

Types of data models (1)

- Textual
  - **Bag-of-words**
  - Represent documents, text in structured data,…, real-world objects (captured as structured data)
  - Lacks “structure”
    - in text, e.g. linguistic structure, hyperlinks, (positional information)
    - Structure in structured data representation

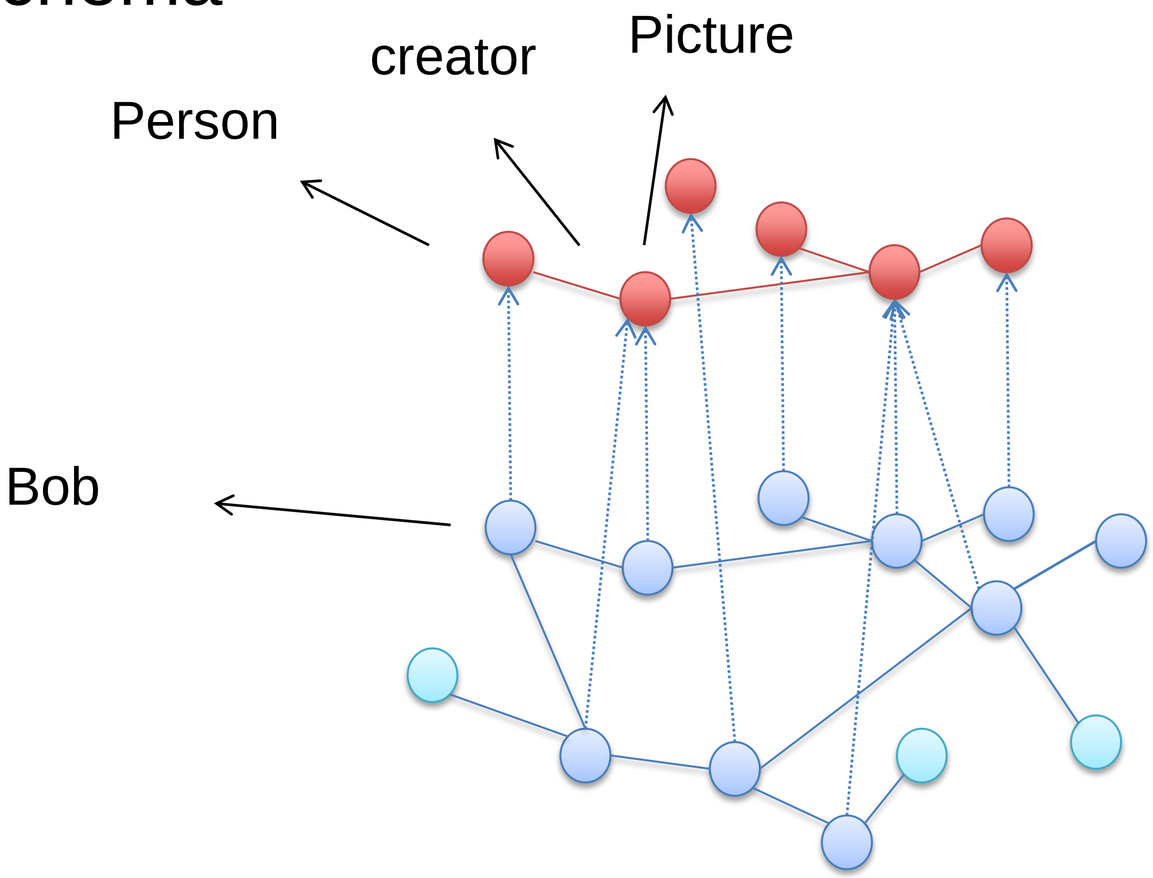


term (statistics)

combination
Cloud
Computing
Technologies
solutions
management
`big data'
industry
solutions
support
complex
……

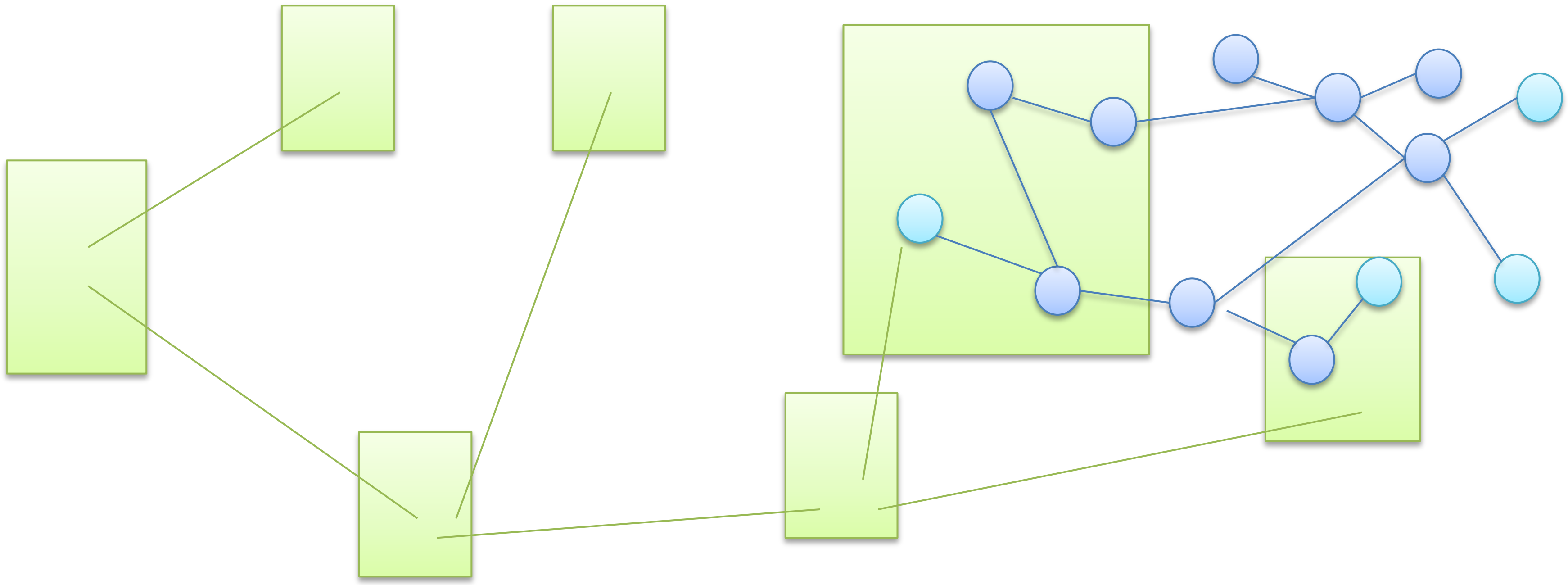
Types of data models (2)

- Textual
- Structured
  - **Resource Description Framework (RDF)**
  - Represent real-world objects, services, applications, …. documents
  - Resource attribute values and relationships between resources
  - Schema



Types of data models (3)

- Textual
- Structured
- **Hybrid**
  - RDF data embedded in text (RDFa)



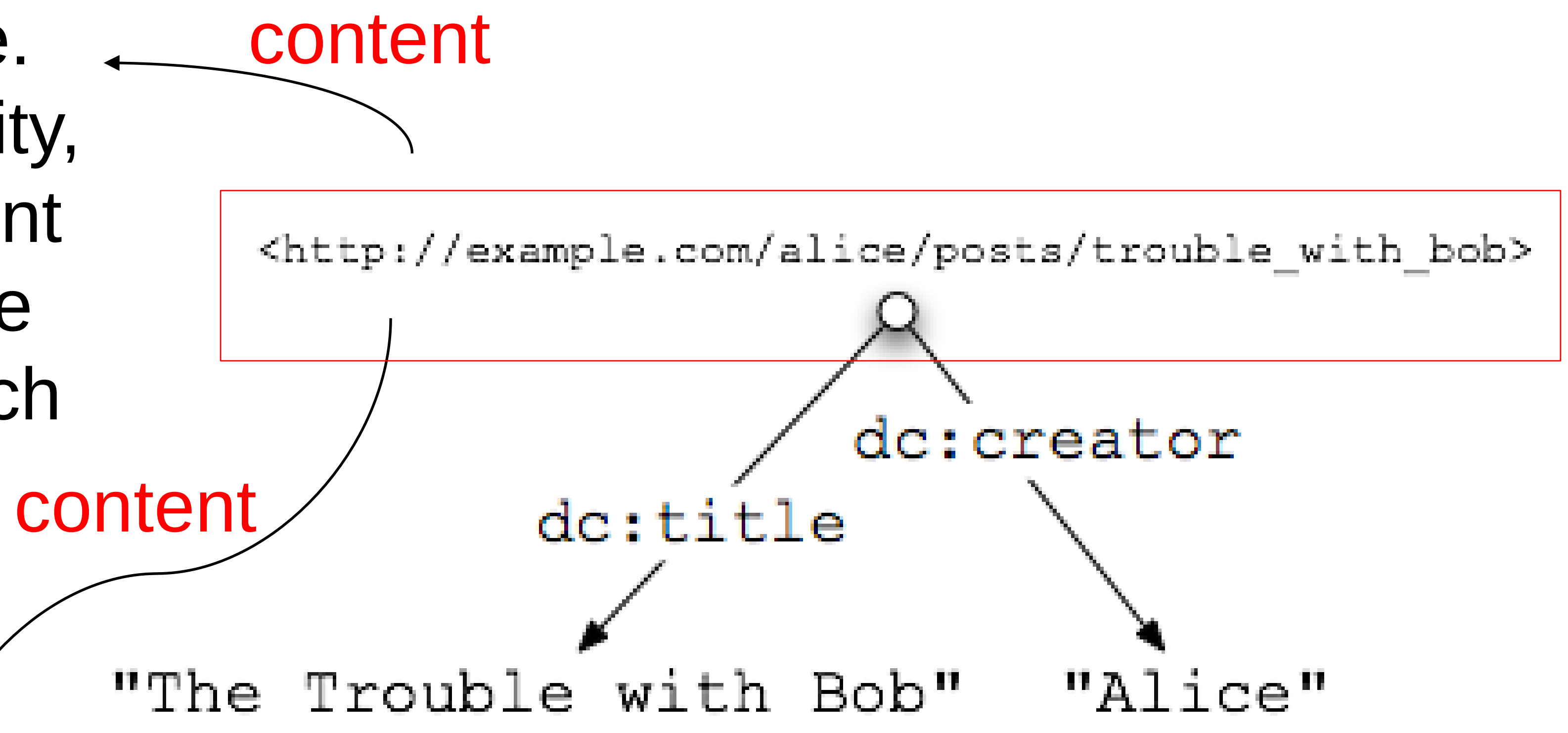
Types of data models – RDFa (1)

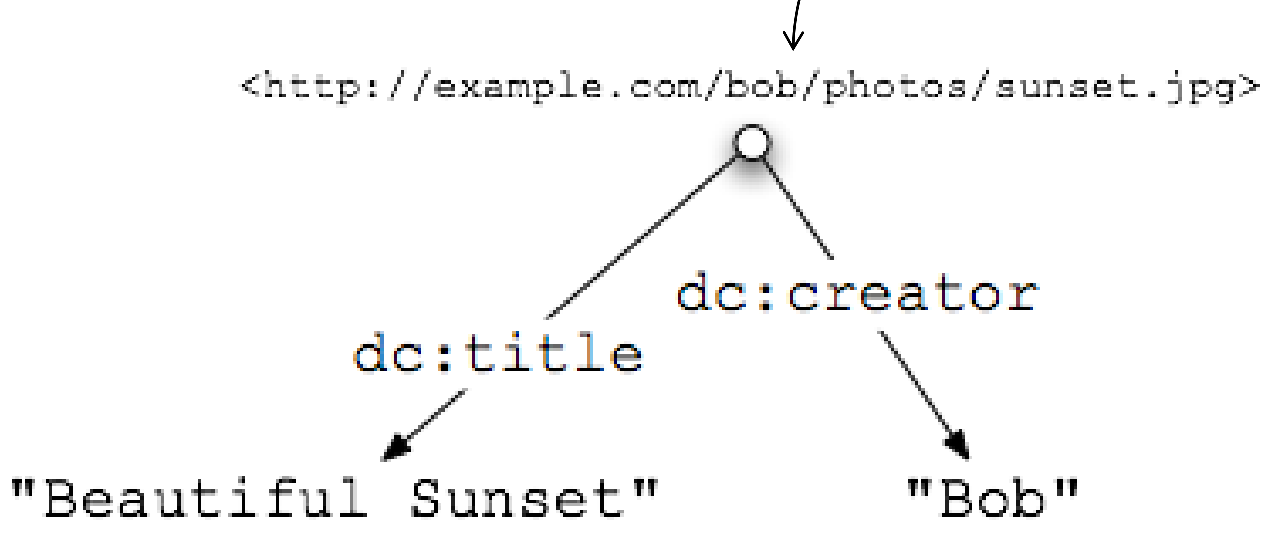
```
…
<div about="/alice/posts/trouble_with_bob">
      <h2 property="dc:title">The trouble with Bob</h2>
      <h3 property="dc:creator">Alice</h3>

             Bob is a good friend of mine. We went to the same university, and
             also shared an apartment in Berlin in 2008. The trouble with Bob is           that he takes
much better photos than I do:

      <div about="http://example.com/bob/photos/sunset.jpg">
        <img src="http://example.com/bob/photos/sunset.jpg" />
        <span property="dc:title">Beautiful Sunset</span>
        by <span property="dc:creator">Bob</span>.
      </div>
</div>
…
```

adopted from : http://www.w3.org/TR/xhtml-rdfa-primer/

Types of semantic data – RDFa (2)

Bob is a good friend of mine. We went to the same university, and also shared an apartment in Berlin in 2008. The trouble with Bob is that he takes much better photos than I do:





adopted from : http://www.w3.org/TR/xhtml-rdfa-primer/

Types of semantic data - conclusion

Semantic data in general can be conceived as a **graph** with **text** and **structured data** items as nodes, and edges represent different types of relationships including explicit **semantic relationships** and vaguely specified ones such as **hyperlinks**!

Formalisms for querying semantic data (1)

Example information need

“Information about a friend of Alice, who shared an apartment with her in Berlin and knows someone working at KIT.”

- Unstructured queries
- Fully-structured queries
- Hybrid queries: unstructured + structured

Formalisms for querying semantic data (2)

Example information need

“Information about a friend of ***Alice****, who* ***shared an apartment with her in Berlin*** *and knows someone working at KIT.”*

- Unstructured
  - NL
  - **Keywords**

shared | apartment | Berlin | Alice

Formalisms for querying semantic data (3)

AI

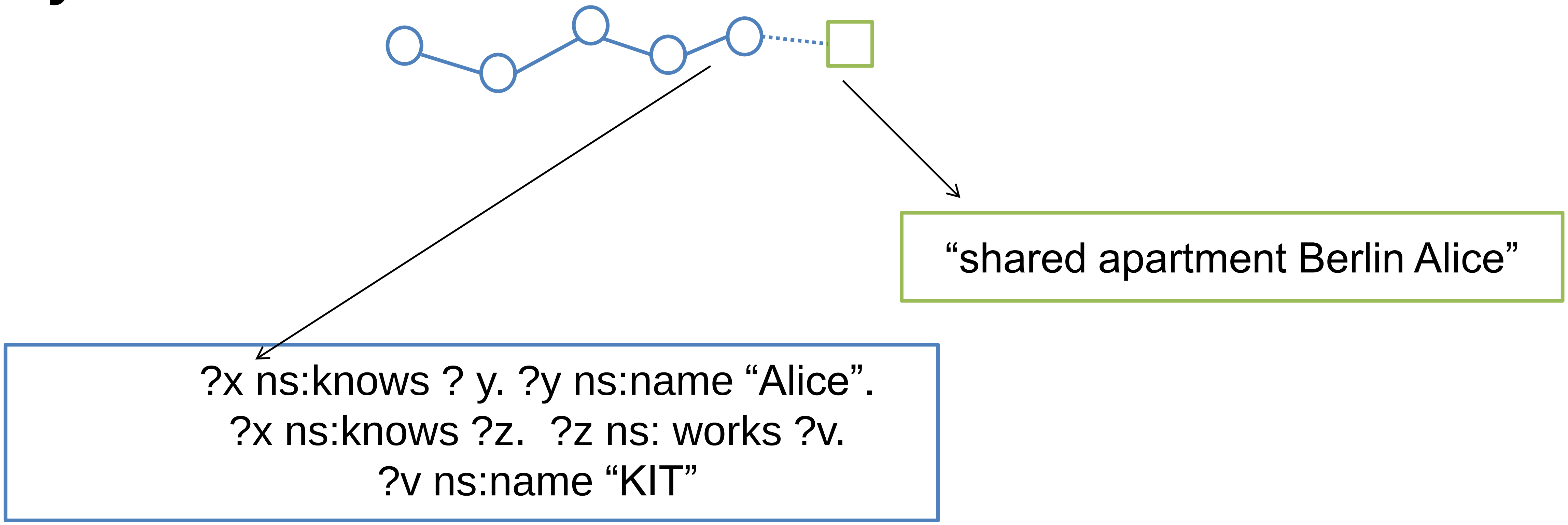
Example information need

*"Information about **a friend of Alice**, who shared an apartment with her in Berlin and **knows someone working at KIT**."*

- Unstructured
- Fully-structured
  - SPARQL: **BGP**, filter, optional, union, select, construct, ask, describe
    - ✦ PREFIX ns: <http://example.org/ns#>
      SELECT ?x
      WHERE { ?x ns:knows ? y. ?y ns:name "Alice".
          ?x ns:knows ?z. ?z ns: works ?v. ?v ns:name "KIT" }

Formalisms for querying semantic data (4)

- Fully-structured
- Unstructured
- Hybrid: content and structure constraints

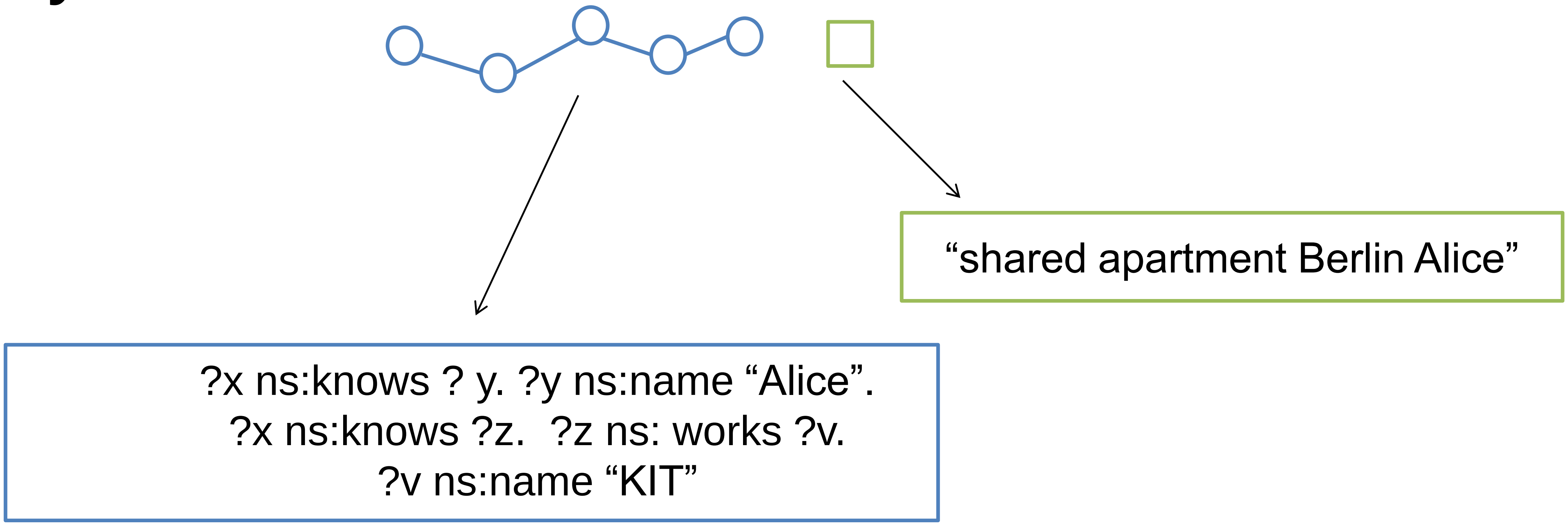


“shared apartment Berlin Alice”

?x ns:knows ? y. ?y ns:name “Alice”.
?x ns:knows ?z. ?z ns: works ?v.
?v ns:name “KIT”

Formalisms for querying semantic data (5)

- Fully-structured
- Unstructured
- Hybrid: content and structure constraints



“shared apartment Berlin Alice”

?x ns:knows ? y. ?y ns:name “Alice”.
?x ns:knows ?z. ?z ns: works ?v.
?v ns:name “KIT”

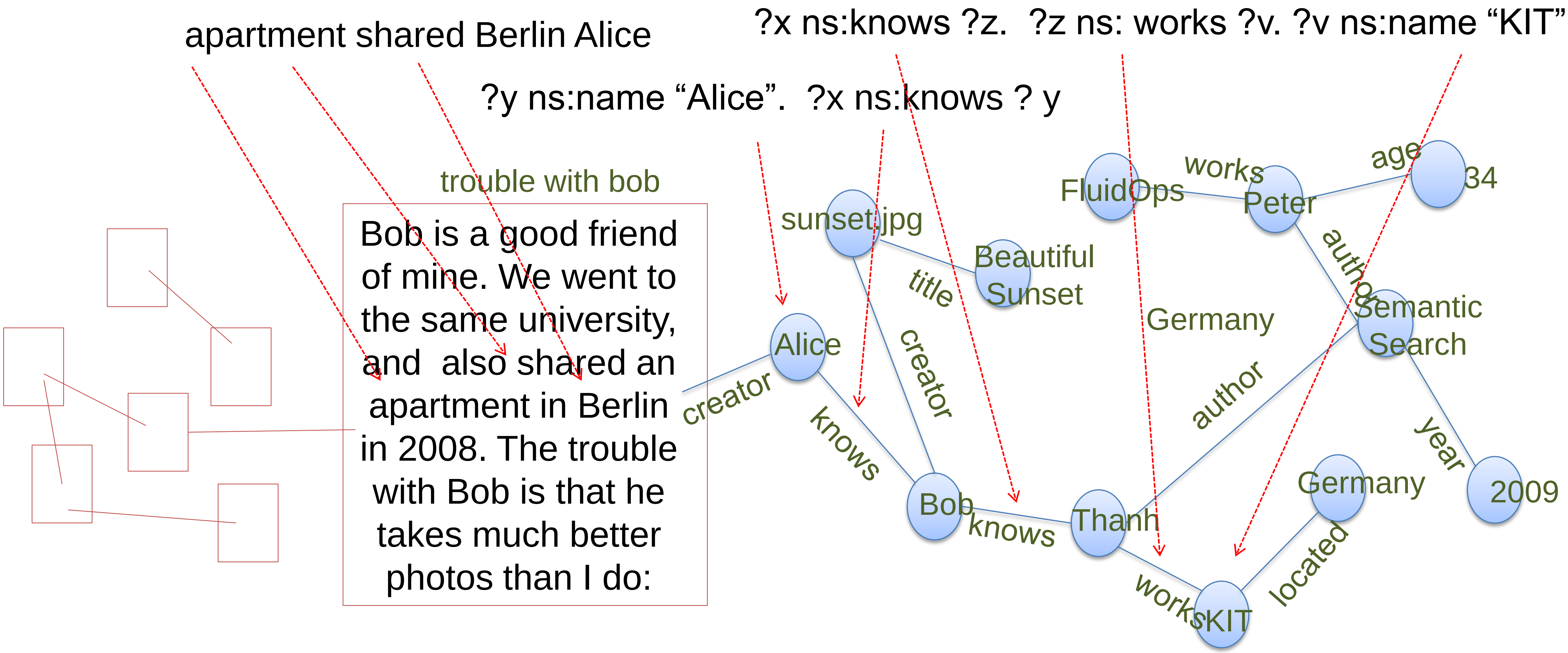
Formalisms for querying semantic data - conclusion

Semantic search queries can be conceived as **graph patterns** with nodes referring to **text** and **structured data** items, and edges referring to relationships between these items!

Processing hybrid graph patterns (1)

Example information need

*“Information about a friend of **Alice**, who **shared an apartment with her in Berlin** and knows someone working at KIT.”*



apartment shared Berlin Alice

?x ns:knows ?z. ?z ns: works ?v. ?v ns:name “KIT”

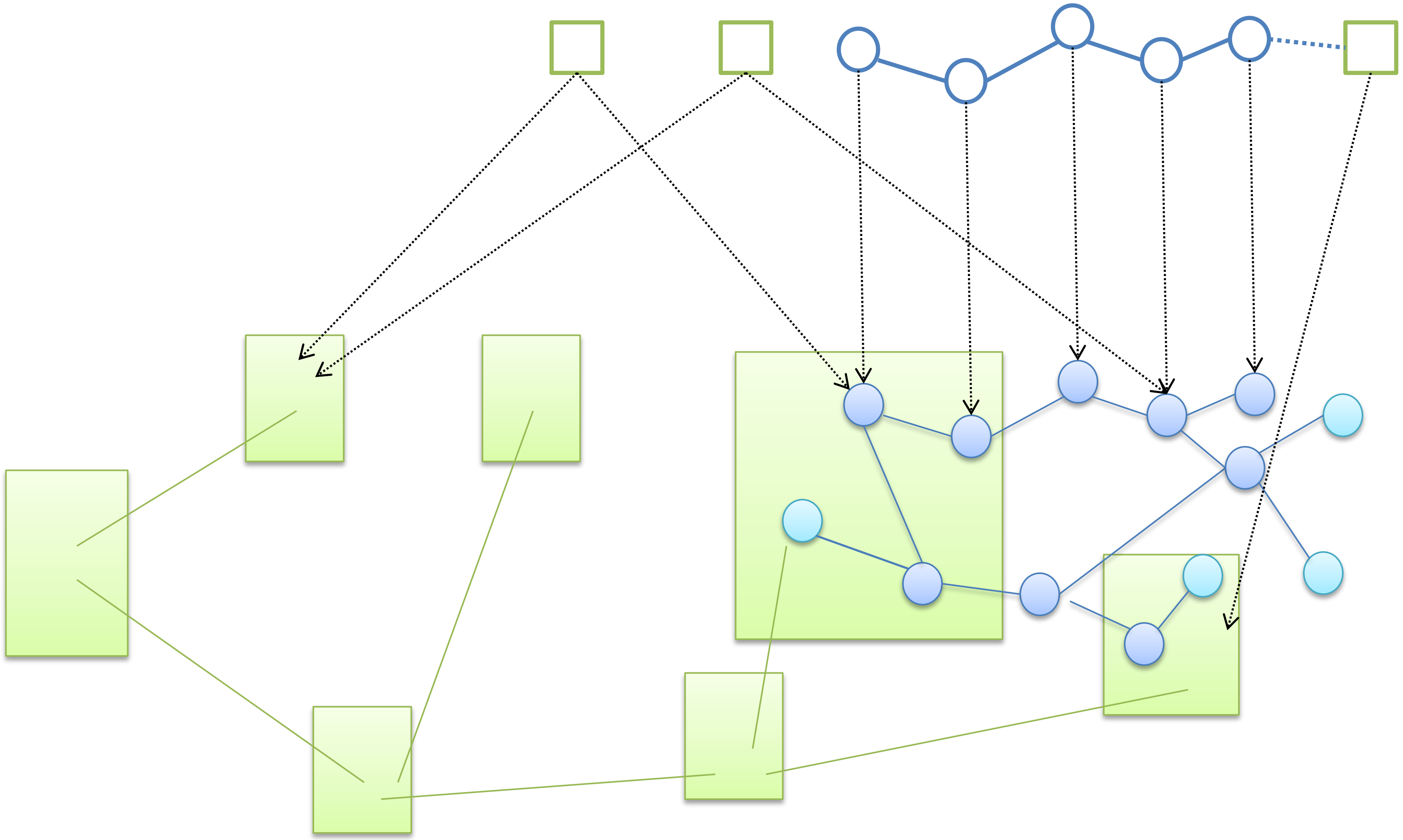
?y ns:name “Alice”. ?x ns:knows ? y

trouble with bob

Bob is a good friend of mine. We went to the same university, and also shared an apartment in Berlin in 2008. The trouble with Bob is that he takes much better photos than I do:

Processing hybrid graph patterns (2)

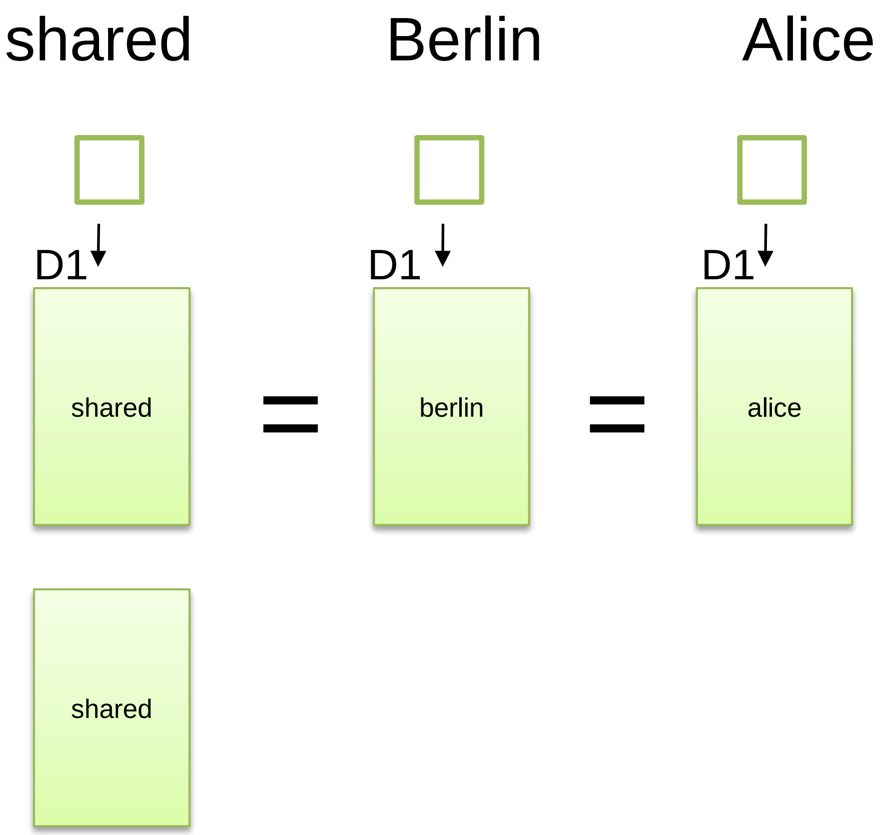
- **Matching** hybrid graph patterns against data

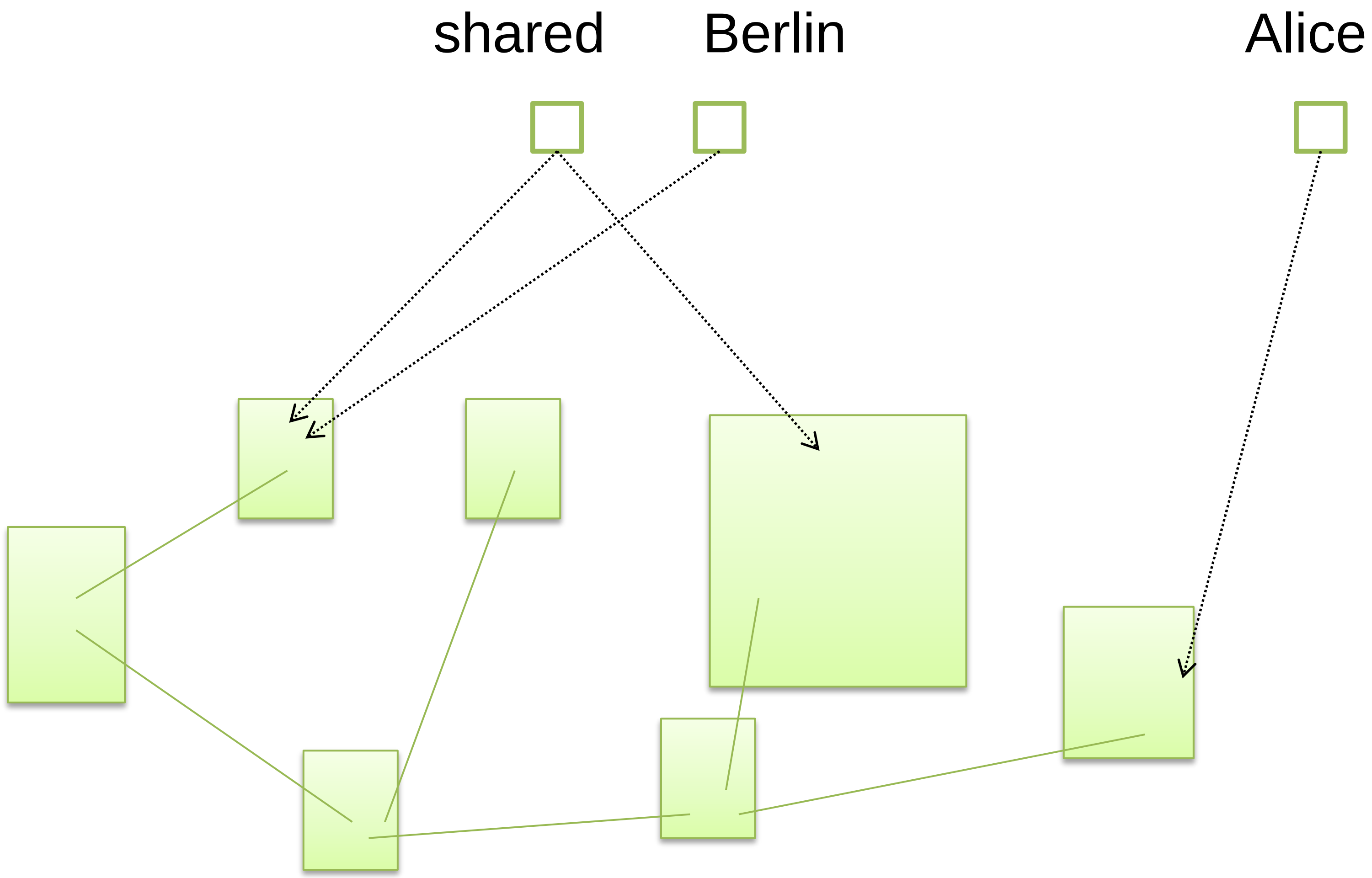


Matching keyword query against text

- Retrieve documents
  - Inverted list (inverted index)

    keyword → {<doc1, pos, score, ...>,
    <doc2, pos, score, ...>, ...}
- AND-semantics: top-k join



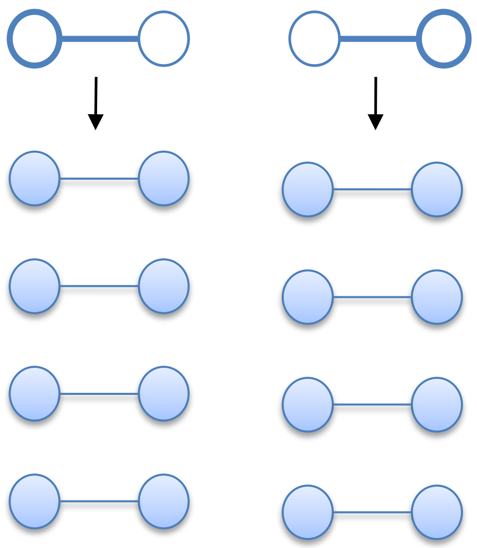


Matching structured query against structured data

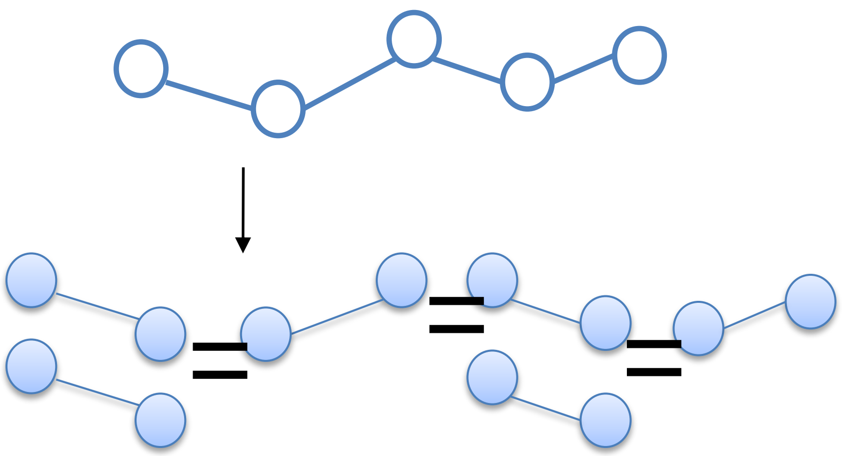
- Retrieve data for triple patterns
  - Index on tables
  - Multiple “redundant” indexes to cover different access patterns
- Join (conjunction of triples)
  - Blocking, e.g. linear merge join (required sorted input)
  - Non-blocking, e.g. symmetric hash-join
  - Materialized join indexes

Per1 ns:works **?v** **?v** ns:name “KIT”

SP-index PO-index

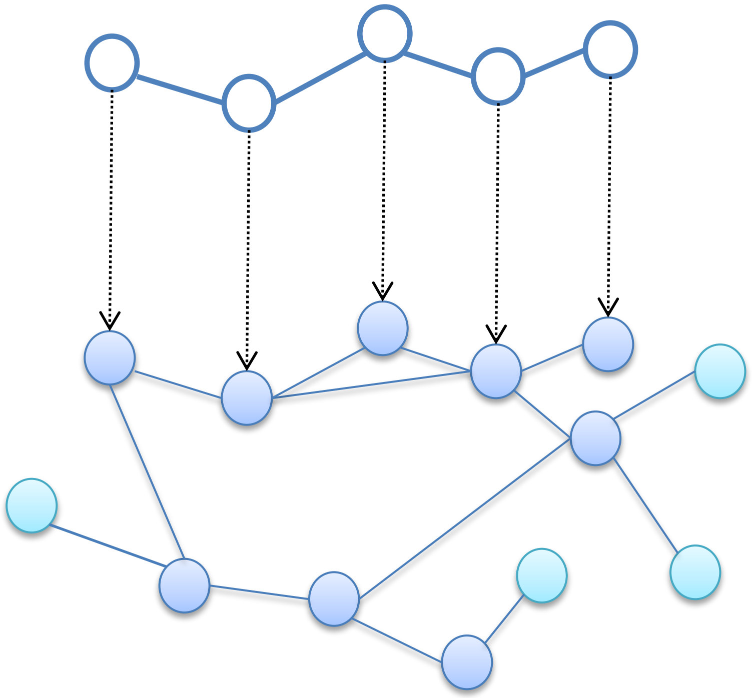


Per1 ns:works Ins1 Ins1 ns:name KIT



Per1 ns:works **Ins1** **Ins1** ns:name KIT

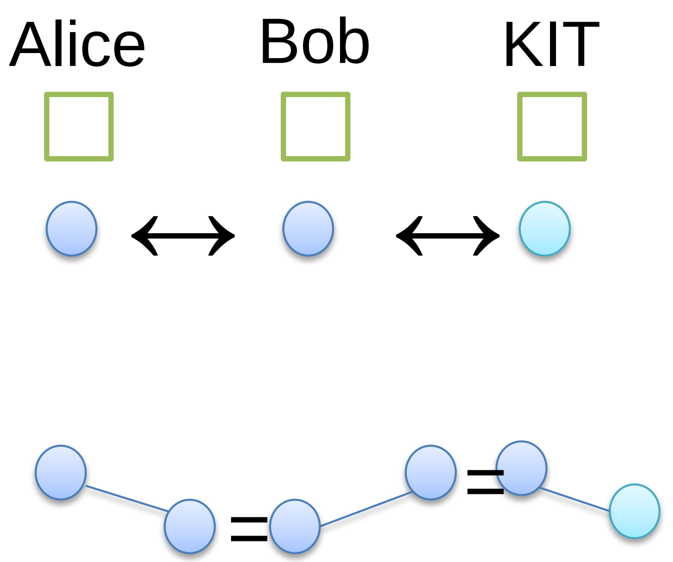
?x ns:knows ?y. ?x ns:knows ?z.
?z ns: works ?v. ?v ns:name “KIT”



Matching keyword query against structured data

- Retrieve keyword elements
  - Using inverted index

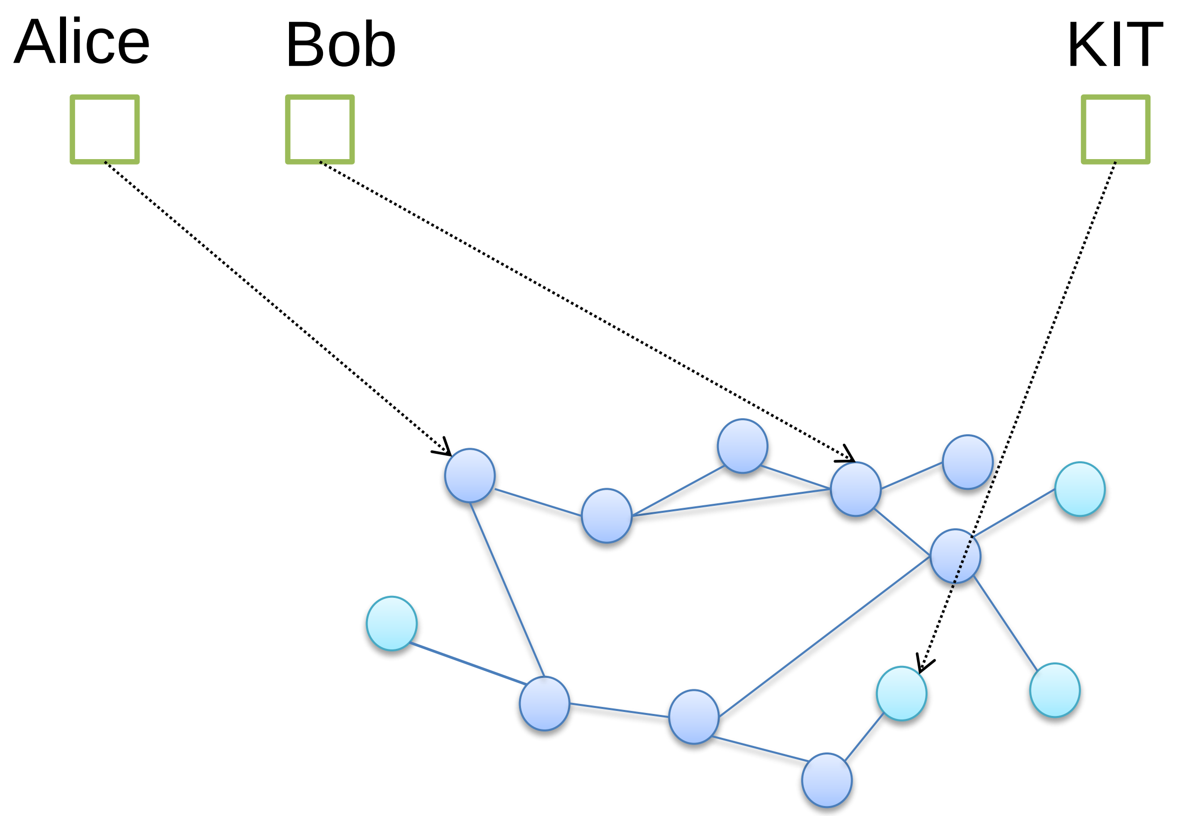
    keyword → {<el1, score, ...>, <el2, score, ...>,…}
- Exploration / “Join”
  - Data indexes for triple lookup
  - Materialized index (paths up to graphs)
  - Top-k Steiner tree search, top-k subgraph exploration



Alice ns:knows **Bob**

Bob ns:works **Inst1**

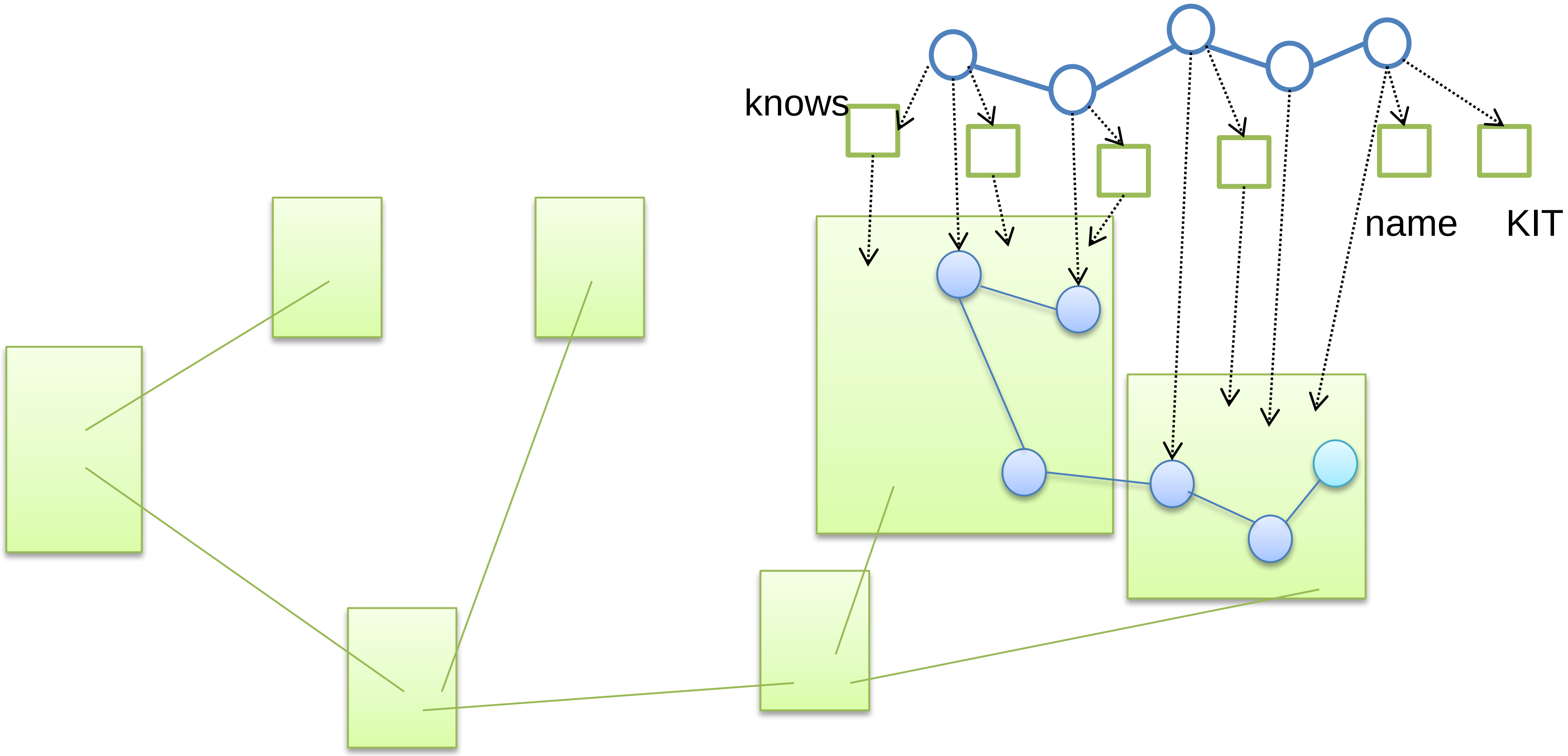
Inst1 ns:name KIT



Matching structured query against text

- Based on offline IE (offline see Peter's slides)
- Based on online IE, i.e., "retrieve " is as follows
  - Derive keywords to retrieve relevant documents
  - On-the-fly information extraction, i.e., phrase pattern matching "X name Y"
  - Retrieve extracted data for structured part
  - Retrieve documents for derived text patterns, e.g. sequence, windows, reg. exp.

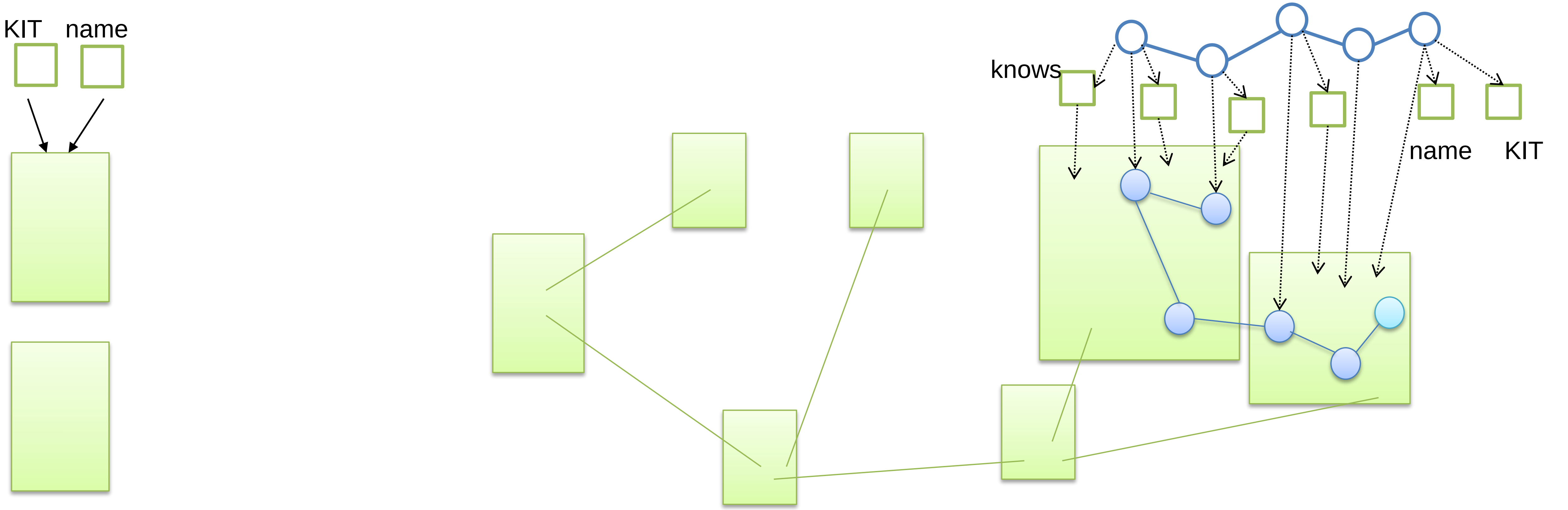
?x ns:knows ?y. ?x ns:knows ?z.
?z ns: works ?v. ?v ns:name "KIT"



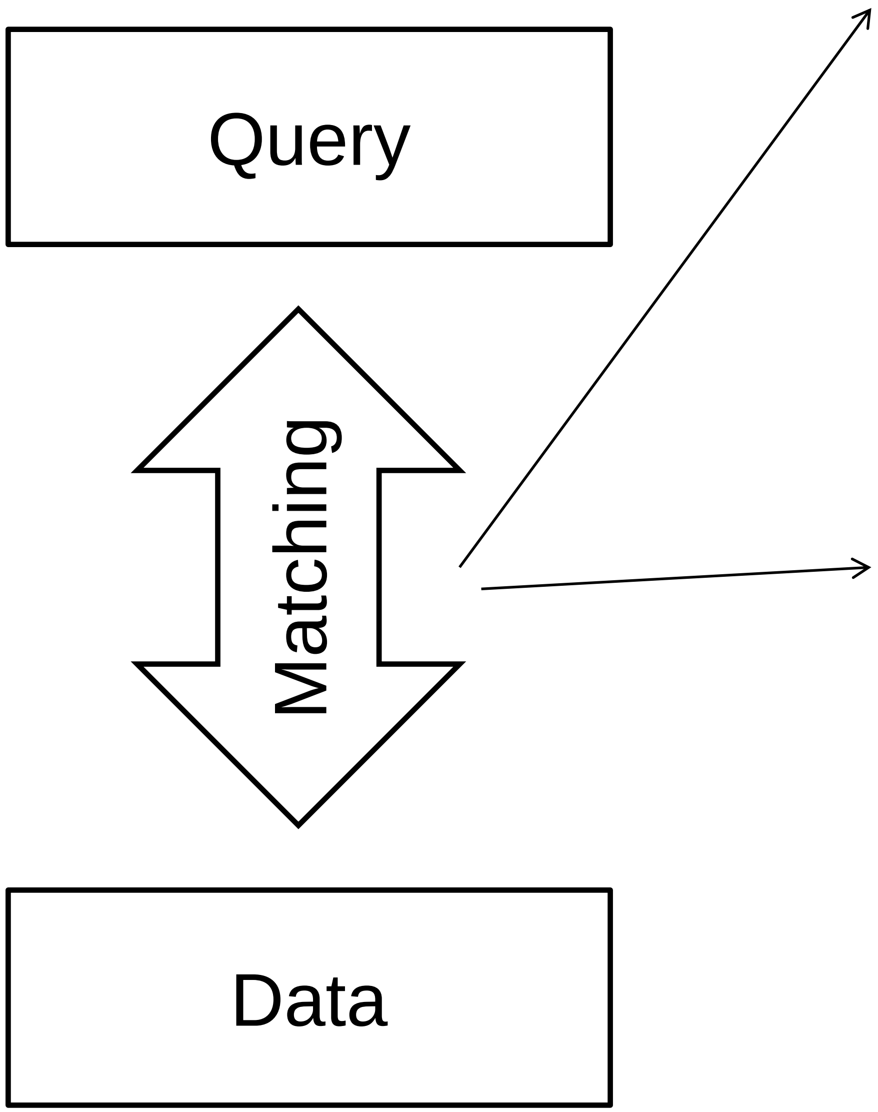
Matching structured query against text

- Index
  - Inverted index for document retrieval and pattern matching
  - Join index → inverted index for storing materialized joins between keywords
  - Neighborhood indexes for phrase patterns

?x ns:knows ?y. ?x ns:knows ?z.
?z ns: works ?v. ?v ns:name “KIT”



Query processing – main tasks



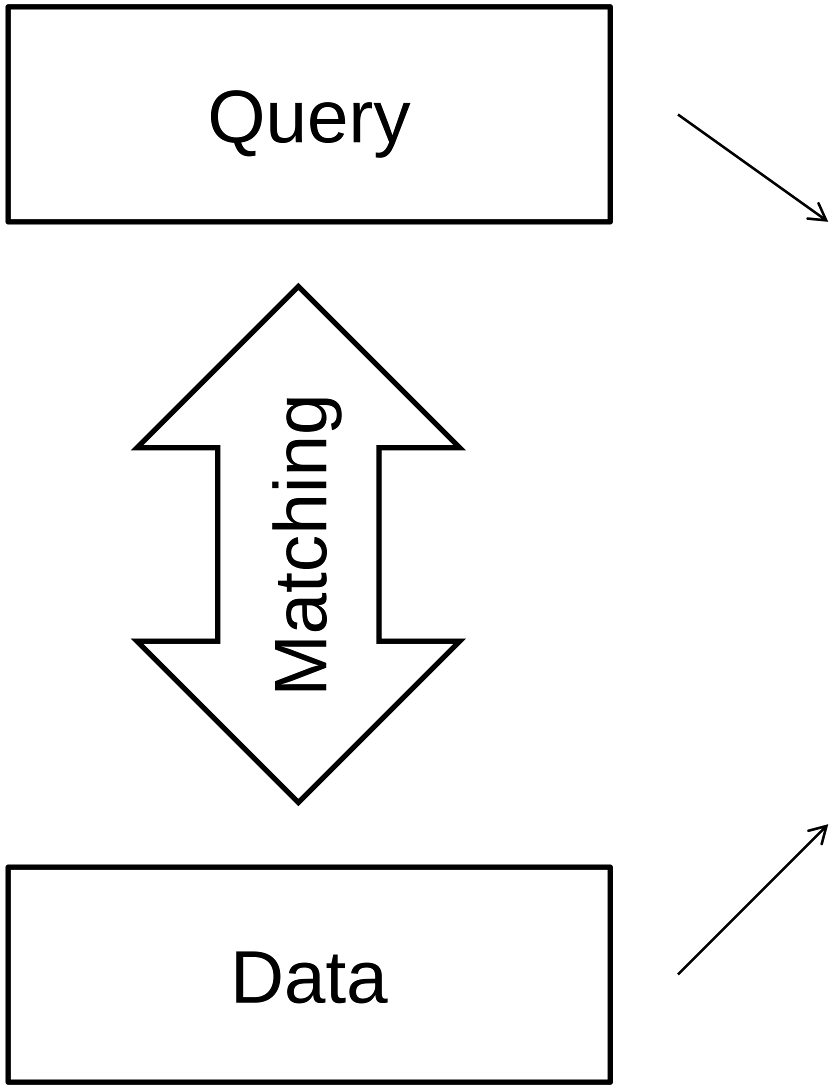
- Retrieval
  - Documents , data elements, triples, paths, graphs
  - Inverted index,…, but also other (B+ tree)
  - Index documents, triples, materialized paths
- Join
  - Different join implementations, efficiency depends on availability of indexes
  - Non-blocking join good for early result reporting and for “unpredictable” Linked Data / data streams scenario

Ranking

Structure

- Problem definition
- Types of ambiguities
- Ranking paradigms
- Model construction
  - Content-based
  - Structure-based

Ranking – problem definition



- Ambiguities arise when representation is incomplete / imprecise
- Ambiguities at the level of
  - elements (**content ambiguity**)
  - structure between elements (**structure ambiguity**)

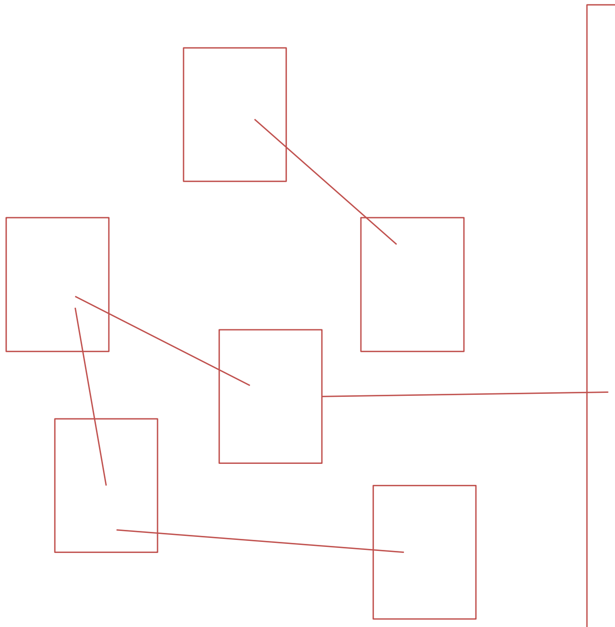
Due to ambiguities in the representation of the information needs and the underlying resources, the results cannot be guaranteed to exactly match the query. Ranking is the problem of determining the **degree of matching** using some notions of **relevance**.

Content ambiguity

apartment shared Berlin Alice

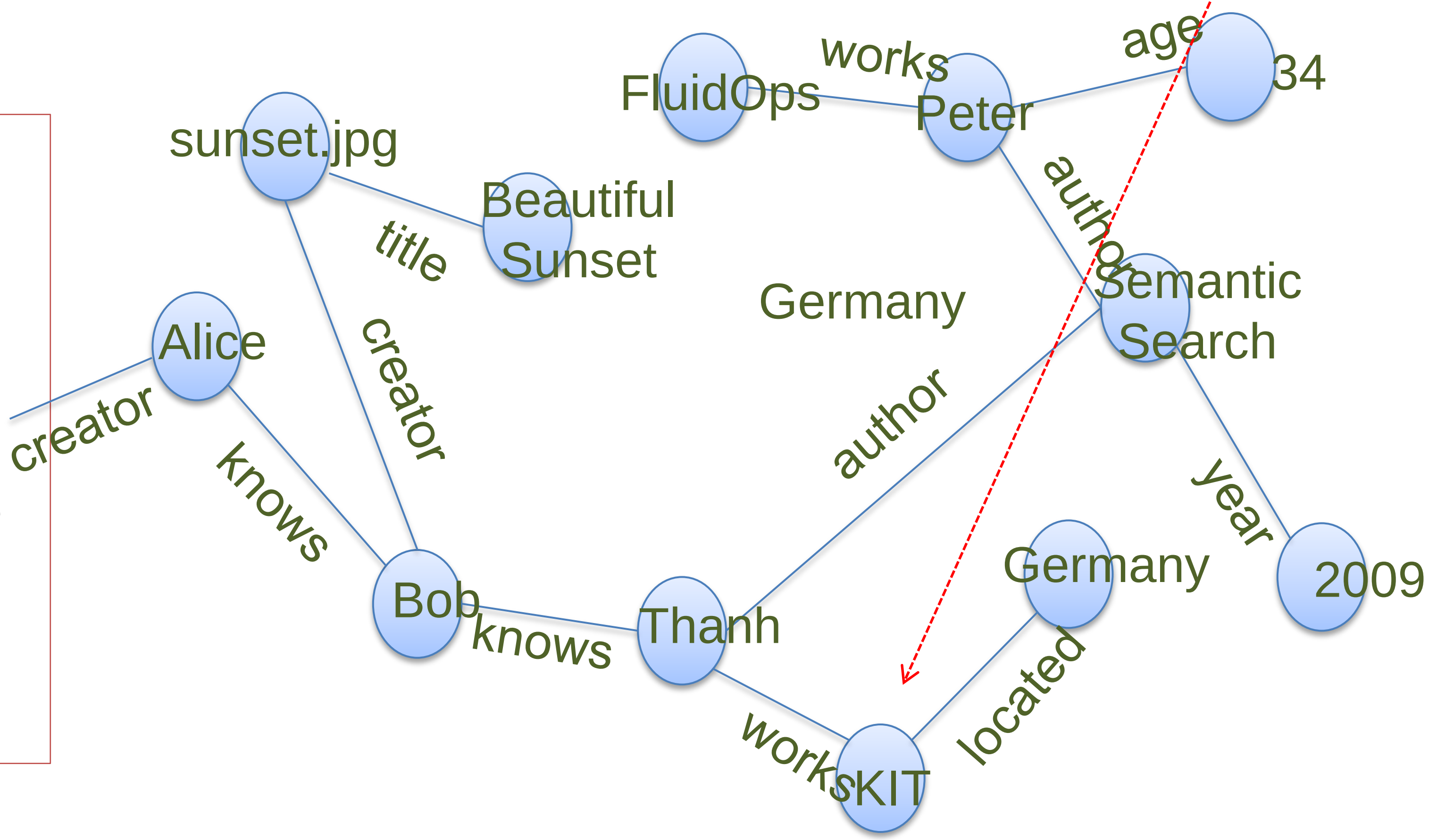
?x ns:knows ?z. ?z ns: works ?v. ?v ns:name “KIT”

?y ns:name “Alice”. ?x ns:knows ? y



trouble with bob

Bob is a good friend of mine. We went to the same university, and also shared an apartment in Berlin in 2008. The trouble with Bob is that he takes much better photos than I do:



What is meant by “Berlin” in the query?
What is meant by “Berlin” in the data?
A city with the name Berlin? a person?

What is meant by “KIT” in the query?
What is meant by “KIT” in the data?
A research group? a university? a location?

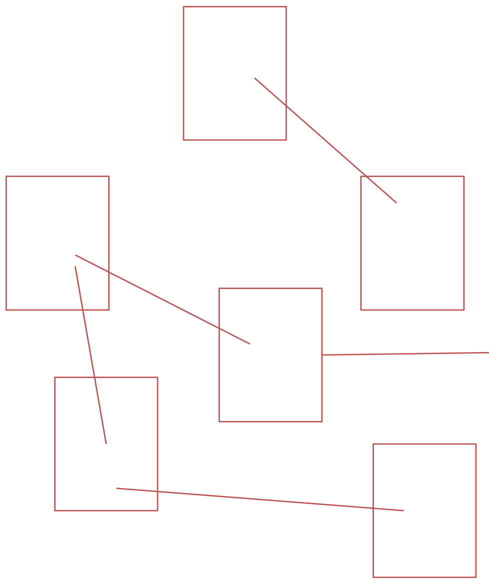
Structure ambiguity

apartment shared Berlin Alice

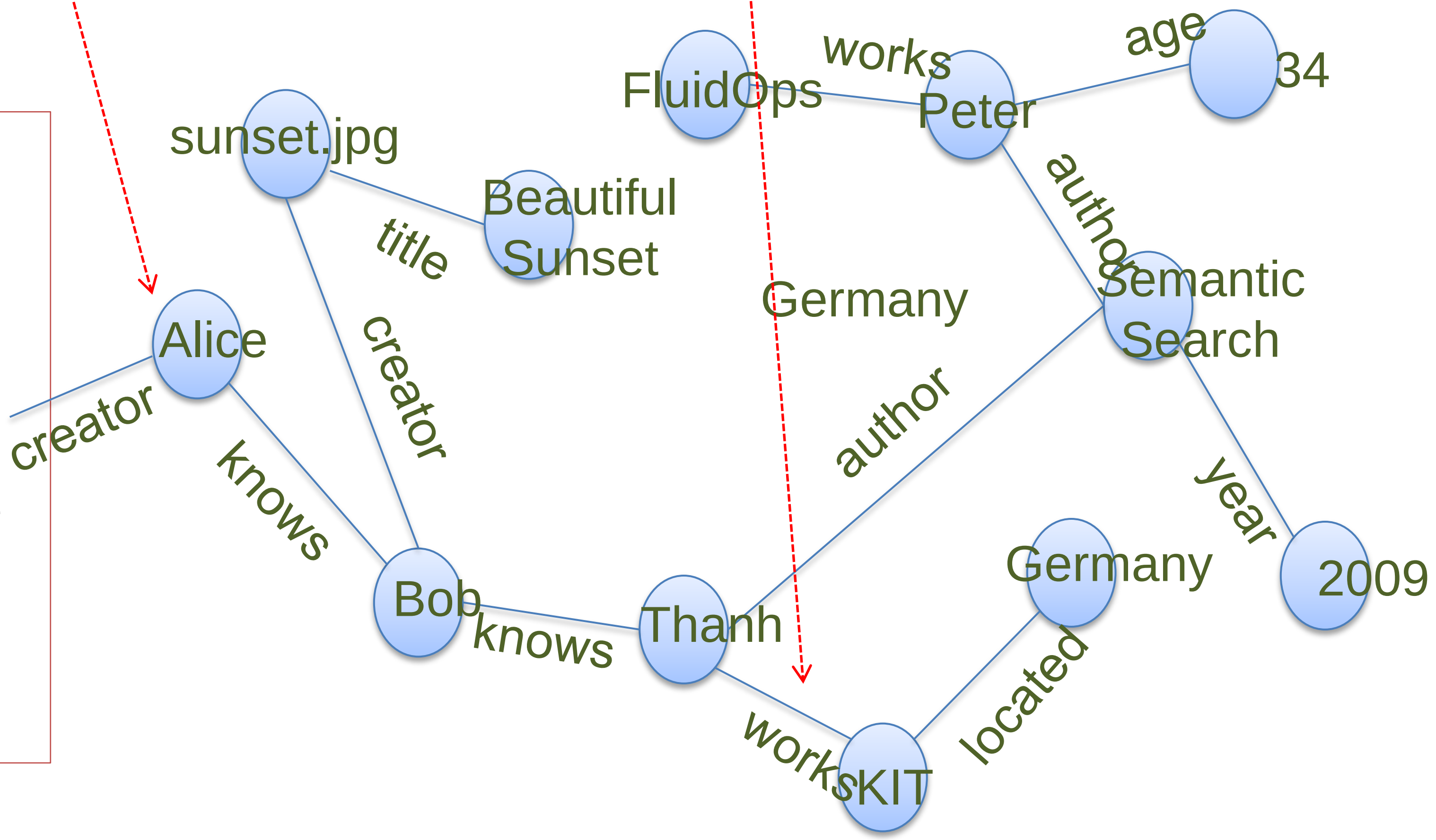
?x ns:knows ?z. ?z ns: works ?v. ?v ns:name “KIT”

?y ns:name “Alice”. ?x ns:knows ? y

trouble with bob



Bob is a good friend of mine. We went to the same university, and also shared an apartment in Berlin in 2008. The trouble with Bob is that he takes much better photos than I do:



What is the connection between “Berlin” and “Alice”?
Friend? Co-worker?

What is meant by “works”?
Works at? employed?

Ambiguity

- Recall: query processing is matching at the level of syntax and semantics
- Ambiguities arise when data or query allow for **multiple interpretations**, i.e. multiple matches
  - **Syntactic**, e.g. works vs. works at
  - **Semantic**, e.g. works vs. employ
- “**Aboutness**”, i.e., contain some elements which represent the correct interpretation
  - Ambiguities arise when matching elements of **different granularities**
  - Does *i* contains the interpretation for *j,* given some part(s) of *i* (syntactically/semantically) match *j*
  - E.g. Berlin vs. “…we went to the same university, and also, we shared an apartment in Berlin in 2008…”
- Strictly speaking, ranking is performed after syntactic / semantic matching is done!

Features: What to use to deal with ambiguities?

What is meant by “Berlin”? What is the connection between “Berlin” and “Alice”?

- Content features
  - **Frequencies** of terms: *d* more likely to be “about” a query term *k* when *d* more often, mentions *k* *(probabilistic IR)*
  - **Co-occurrences**: *terms K that often co-occur form a contextual interpretation, i.e., topics (cluster hypothesis)*
- Structure features
  - Consider relevance at level of fields
  - Linked-based popularity

Ranking paradigms

- Explicit relevance model
  - Foundation: **probability ranking principle**
  - Ranking results by the posterior probability (odds) of being observed in the relevant class:
  - P(w|R) varies in different approaches, e.g., binary independence model, 2-poisson model, **relevance model**

$$\frac{P(D|R)}{P(D/N)}$$

$$P(D|R) = \prod_{w \in D} P(w|R) \prod_{w \notin D} (1 - P(w|N))$$

$$P(w|R) \approx P(w|q_1,\ldots,q_k) = \sum_{m \in M} P(m)P(w|m) \sum_{i=1}^{k} P(q_k|m)$$

Ranking paradigms

- No explicit notion of relevance: similarity between the query and the document model
  - Vector space model (cosine similarity)
  - Language models (KL divergence)

$$Sim(q,d) = Cos((w_{1,d},\ldots,w_{t,d}),(w_{1,q},\ldots,w_{k,q}))$$

$$Sim(q,d) = -KL(\theta_q \| \theta_d) = -\sum_{t \in V} P(t|\theta_q)\log(\frac{P(t|\theta_q)}{P(t|\theta_d)})$$

Model construction

- How to obtain
  - Relevance models?
  - Weights for query / document terms?
  - Language models for document / queries?

Content-based model construction

- Document statistics, e.g.
  - Term frequency
  - Document length
- Collection statistics, e.g.
  - Inverse document frequency
  - Background language models

$$w_{t,d} = \frac{tf}{|d|} * idf$$

$$P(t \mid \theta_d) = \lambda \frac{tf}{|d|} + (1 - \lambda) P(t \mid C)$$

- An object is more likely about “Berlin”?
  - When it contains a **relatively** high number of **mentions** of the term “Berlin”
  - When the number of mentions of this term in the overall collection is relatively low

Structure-based model construction

- Consider structure of objects during content-based modeling, i.e., to obtain structured content-based model
  - Content-based model for structured objects, documents and for general tuples

$$P(t \mid \theta_d) = \sum_{f \in F_d} \alpha_f P(t \mid \theta_f)$$

- An object is more likely about “Berlin”?
  - When one of its (important) **fields** contains a relatively high number of mentions of the term “Berlin”

Structure-based model construction

- PageRank
  - Link analysis algorithm
  - Measuring relative importance of nodes
  - Link counts as a vote of support
  - The PageRank of a node recursively depends on the number and PageRank of all nodes that link to it (incoming links)
- ObjectRank
  - Types and semantics of links vary in structured data setting
  - Authority transfer schema graph specifies connection strengths
  - Recursively compute authority transfer data graph

- An object about “Berlin” is more important than one another?
  - When a relatively large number of objects are linked to it

Taxonomy of ranking approaches

- Explicitly vs. non-explicitly relevance-based
- Content-based ranking
- Structure-based ranking
- Content- and-structure-based ranking

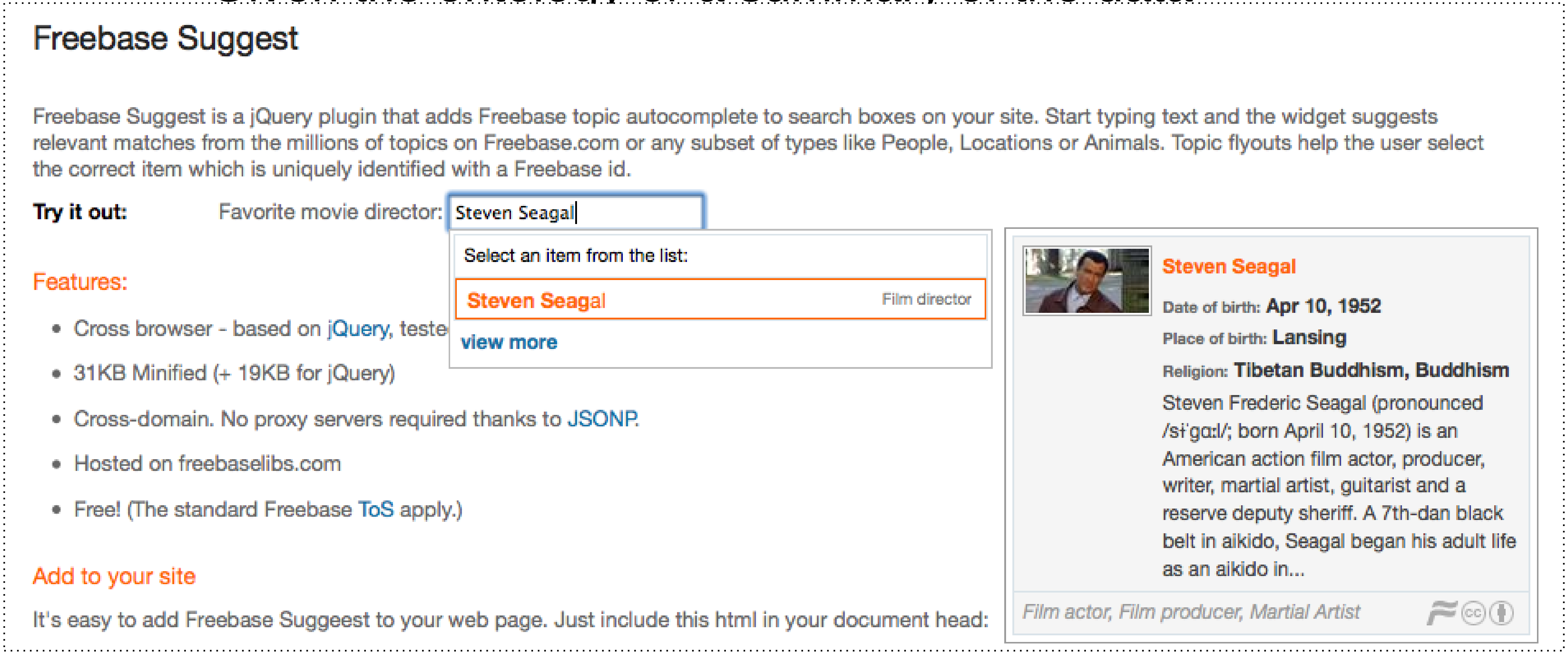
Result Presentation

Search interface

- Input and output functionality
  - helping the user to formulate complex queries
  - presenting the results in an intelligent manner
- Semantic Search brings improvements in
  - Query formulation
  - Snippet generation
  - Suggesting related entities
  - Adaptive and interactive presentation
    - Presentation adapts to the kind of query and results presented
    - Object results can be actionable, e.g. buy this product
  - Aggregated search
    - Grouping similar items, summarizing results in various ways
    - Filtering (facets), possibly across different dimensions
  - Task completion
    - Help the user to fulfill the task by placing the query in a task context

Query formulation

- “Snap-to-grid”: suggest the most likely interpretation of the query
  - Given the ontology or a summary of the data



Freebase Suggest

Freebase Suggest is a jQuery plugin that adds Freebase topic autocomplete to search boxes on your site. Start typing text and the widget suggests relevant matches from the millions of topics on Freebase.com or any subset of types like People, Locations or Animals. Topic flyouts help the user select the correct item which is uniquely identified with a Freebase id.

Try it out: Favorite movie director: Steven Seagal

Select an item from the list:

Steven Seagal — Film director

view more

Features:

- Cross browser - based on jQuery, teste
- 31KB Minified (+ 19KB for jQuery)
- Cross-domain. No proxy servers required thanks to JSONP.
- Hosted on freebaselibs.com
- Free! (The standard Freebase ToS apply.)

Add to your site

It's easy to add Freebase Suggeest to your web page. Just include this html in your document head:

Steven Seagal

Date of birth: **Apr 10, 1952**

Place of birth: **Lansing**

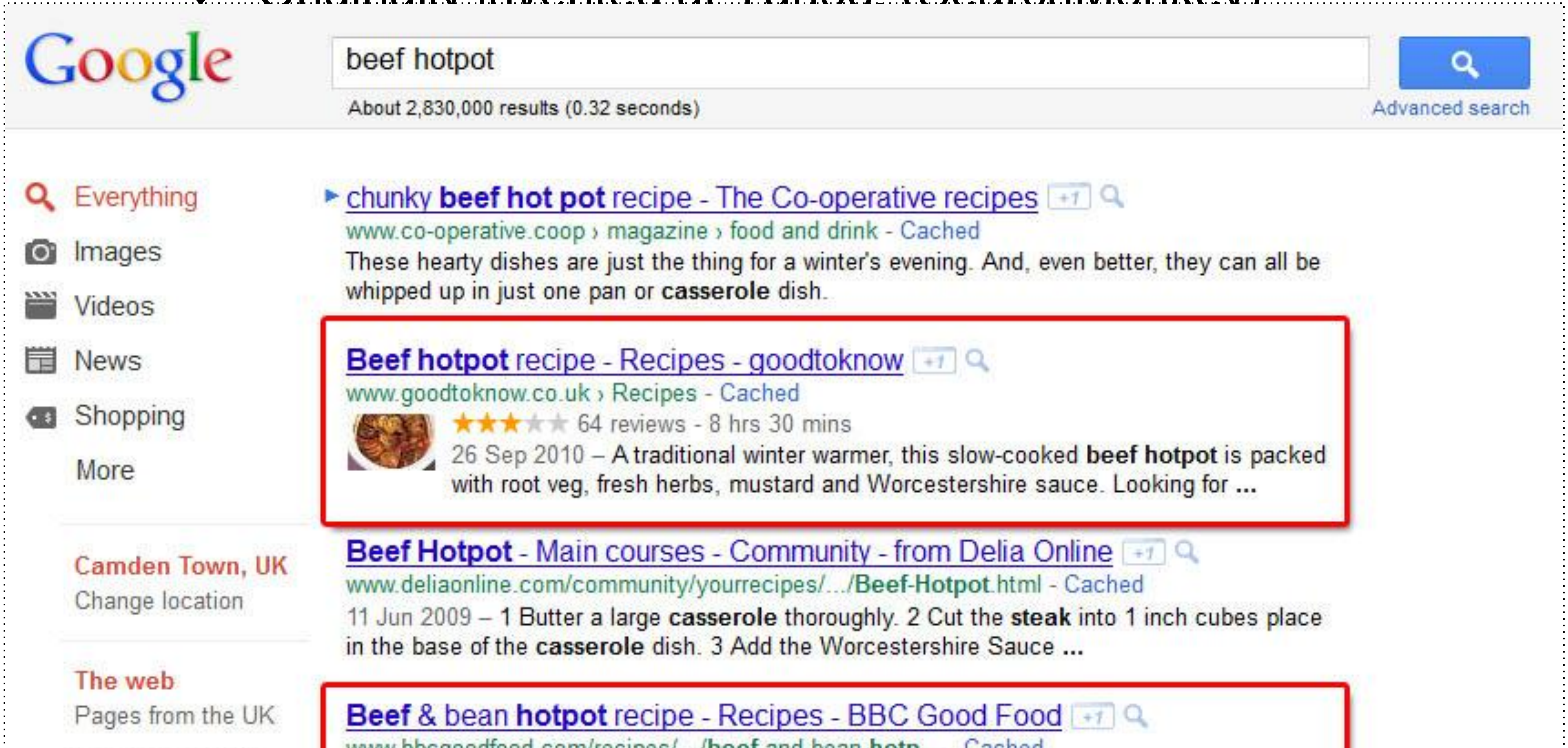
Religion: **Tibetan Buddhism, Buddhism**

Steven Frederic Seagal (pronounced /sɪˈɡɑːl/; born April 10, 1952) is an American action film actor, producer, writer, martial artist, guitarist and a reserve deputy sheriff. A 7th-dan black belt in aikido, Seagal began his adult life as an aikido in...

Film actor, Film producer, Martial Artist

Enhanced results/Rich Snippets

- Use mark-up from the webpage to generate search snippets
  - Originally invented at Yahoo! (SearchMonkey)



beef hotpot

About 2,830,000 results (0.32 seconds)

chunky **beef hot pot** recipe - The Co-operative recipes
www.co-operative.coop › magazine › food and drink - Cached
These hearty dishes are just the thing for a winter's evening. And, even better, they can all be whipped up in just one pan or **casserole** dish.

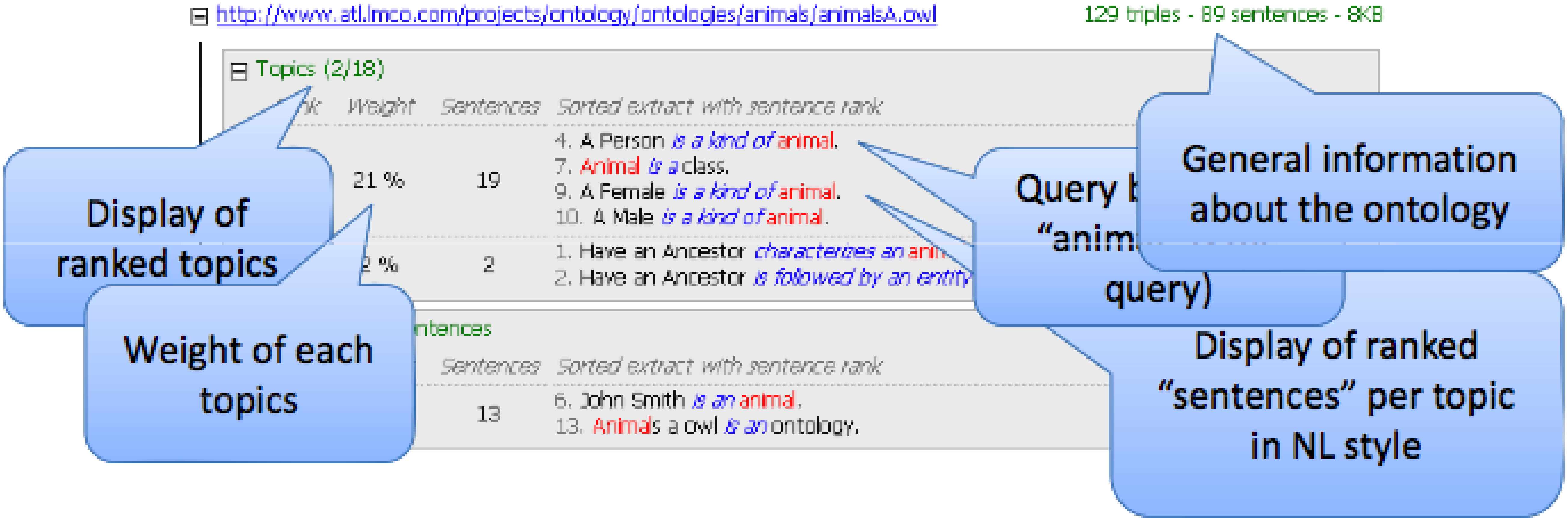
Beef hotpot recipe - Recipes - goodtoknow
www.goodtoknow.co.uk › Recipes - Cached
★★★☆☆ 64 reviews - 8 hrs 30 mins
26 Sep 2010 – A traditional winter warmer, this slow-cooked **beef hotpot** is packed with root veg, fresh herbs, mustard and Worcestershire sauce. Looking for ...

Beef Hotpot - Main courses - Community - from Delia Online
www.deliaonline.com/community/yourrecipes/.../**Beef-Hotpot**.html - Cached
11 Jun 2009 – 1 Butter a large **casserole** thoroughly. 2 Cut the **steak** into 1 inch cubes place in the base of the **casserole** dish. 3 Add the Worcestershire Sauce ...

Beef & bean **hotpot** recipe - Recipes - BBC Good Food
www.bbcgoodfood.com/recipes/.../**beef-and-bean-hotp**... - Cached

Other result presentation tasks

- Select the most relevant resources within an RDF document
  - Penin et al. Snippet Generation for Semantic Web Search Engines, ASWC 2010
- For each resource, rank the properties to be



⊟ http://www.atl.lmco.com/projects/ontology/ontologies/animals/animalsA.owl 129 triples - 89 sentences - 8KB

⊟ Topics (2/18)

nk	Weight	Sentences	Sorted extract with sentence rank
	21 %	19	4. A Person *is a kind of* animal. 7. Animal *is a* class. 9. A Female *is a kind of* animal. 10. A Male *is a kind of* animal.
	2 %	2	1. Have an Ancestor *characterizes an ani* 2. Have an Ancestor *is followed by an entity*

ntences

	Sentences	Sorted extract with sentence rank
	13	6. John Smith *is an* animal. 13. Animals a owl *is an* ontology.

Display of ranked topics

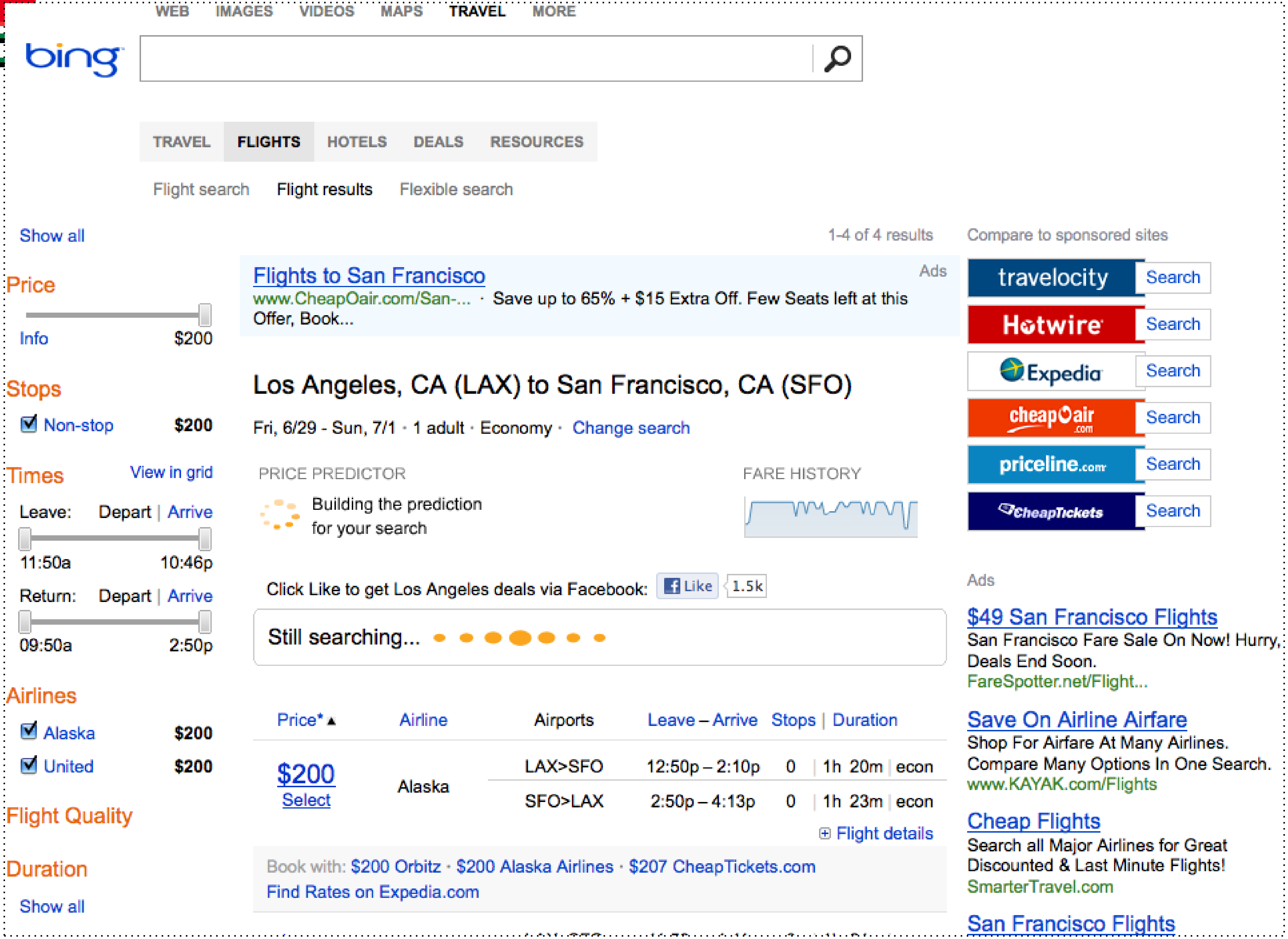
Weight of each topics

Query b “anim query)

General information about the ontology

Display of ranked “sentences” per topic in NL style

Aggregated search: facets



WEB IMAGES VIDEOS MAPS TRAVEL MORE
bing
TRAVEL FLIGHTS HOTELS DEALS RESOURCES
Flight search Flight results Flexible search

Show all

Price

Info $200

Stops

☑ Non-stop $200

Times View in grid

Leave: Depart | Arrive

11:50a 10:46p

Return: Depart | Arrive

09:50a 2:50p

Airlines

☑ Alaska $200

☑ United $200

Flight Quality

Duration

Show all

1-4 of 4 results

Flights to San Francisco — Ads

www.CheapOair.com/San-... · Save up to 65% + $15 Extra Off. Few Seats left at this Offer, Book...

Los Angeles, CA (LAX) to San Francisco, CA (SFO)

Fri, 6/29 - Sun, 7/1 · 1 adult · Economy · Change search

PRICE PREDICTOR

Building the prediction for your search

FARE HISTORY

Click Like to get Los Angeles deals via Facebook: Like 1.5k

Still searching...

Price* ▲	Airline	Airports	Leave – Arrive	Stops	Duration	
$200 Select	Alaska	LAX>SFO	12:50p – 2:10p	0	1h 20m	econ
		SFO>LAX	2:50p – 4:13p	0	1h 23m	econ

⊕ Flight details

Book with: $200 Orbitz · $200 Alaska Airlines · $207 CheapTickets.com

Find Rates on Expedia.com

Compare to sponsored sites

travelocity	Search
Hotwire	Search
Expedia	Search
cheapOair.com	Search
priceline.com	Search
CheapTickets	Search

Ads

$49 San Francisco Flights

San Francisco Fare Sale On Now! Hurry, Deals End Soon.

FareSpotter.net/Flight...

Save On Airline Airfare

Shop For Airfare At Many Airlines. Compare Many Options In One Search.

www.KAYAK.com/Flights

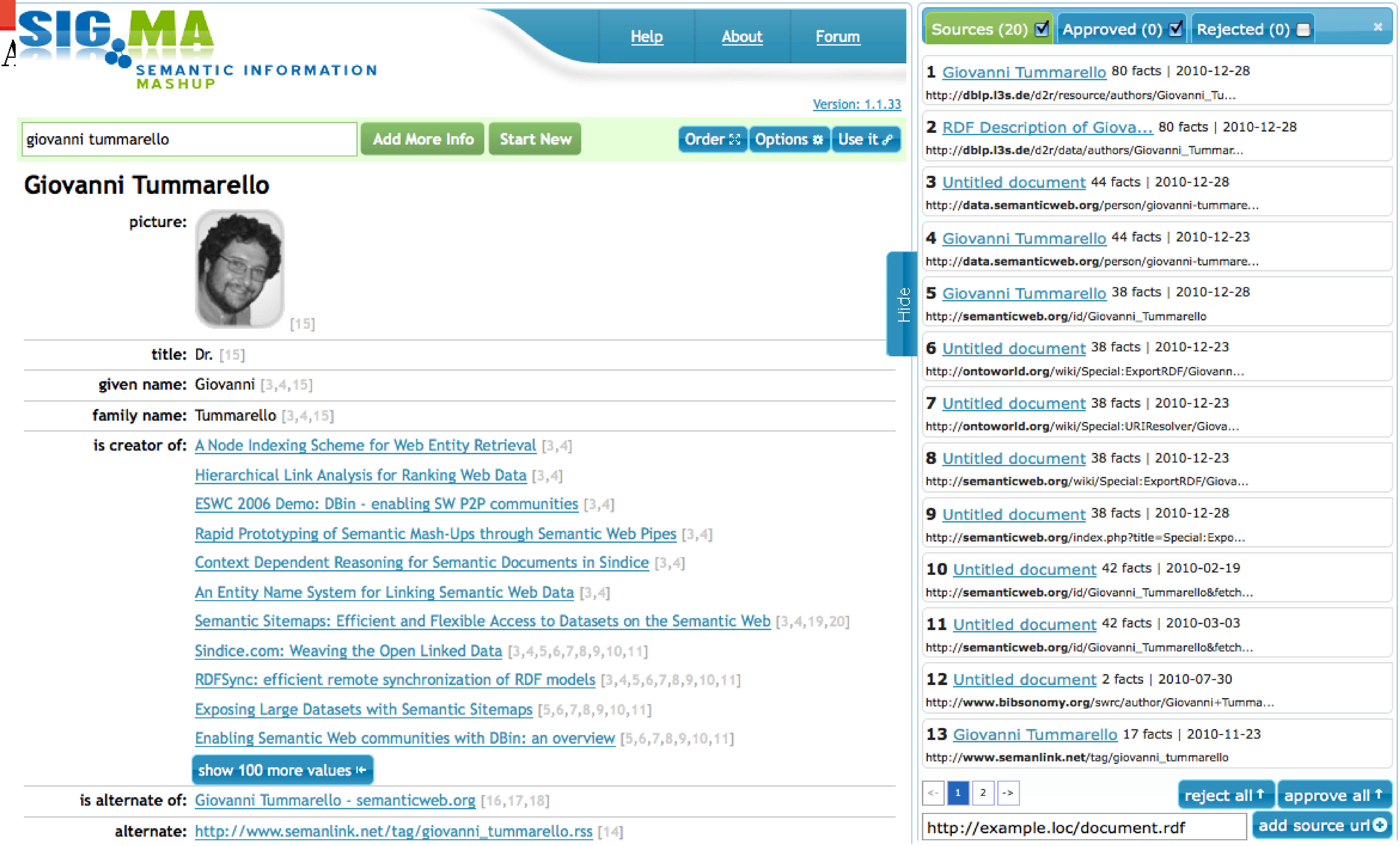
Cheap Flights

Search all Major Airlines for Great Discounted & Last Minute Flights!

SmarterTravel.com

San Francisco Flights

Aggregated search: Sig.ma



giovanni tummarello

Giovanni Tummarello

picture: [15]

title: Dr. [15]

given name: Giovanni [3,4,15]

family name: Tummarello [3,4,15]

is creator of: A Node Indexing Scheme for Web Entity Retrieval [3,4]
Hierarchical Link Analysis for Ranking Web Data [3,4]
ESWC 2006 Demo: DBin - enabling SW P2P communities [3,4]
Rapid Prototyping of Semantic Mash-Ups through Semantic Web Pipes [3,4]
Context Dependent Reasoning for Semantic Documents in Sindice [3,4]
An Entity Name System for Linking Semantic Web Data [3,4]
Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web [3,4,19,20]
Sindice.com: Weaving the Open Linked Data [3,4,5,6,7,8,9,10,11]
RDFSync: efficient remote synchronization of RDF models [3,4,5,6,7,8,9,10,11]
Exposing Large Datasets with Semantic Sitemaps [5,6,7,8,9,10,11]
Enabling Semantic Web communities with DBin: an overview [5,6,7,8,9,10,11]

show 100 more values

is alternate of: Giovanni Tummarello - semanticweb.org [16,17,18]

alternate: http://www.semanlink.net/tag/giovanni_tummarello.rss [14]

1 Giovanni Tummarello 80 facts | 2010-12-28
http://dblp.l3s.de/d2r/resource/authors/Giovanni_Tu...

2 RDF Description of Giova... 80 facts | 2010-12-28
http://dblp.l3s.de/d2r/data/authors/Giovanni_Tummar...

3 Untitled document 44 facts | 2010-12-28
http://data.semanticweb.org/person/giovanni-tummare...

4 Giovanni Tummarello 44 facts | 2010-12-23
http://data.semanticweb.org/person/giovanni-tummare...

5 Giovanni Tummarello 38 facts | 2010-12-28
http://semanticweb.org/id/Giovanni_Tummarello

6 Untitled document 38 facts | 2010-12-23
http://ontoworld.org/wiki/Special:ExportRDF/Giovann...

7 Untitled document 38 facts | 2010-12-23
http://ontoworld.org/wiki/Special:URIResolver/Giova...

8 Untitled document 38 facts | 2010-12-23
http://semanticweb.org/wiki/Special:ExportRDF/Giova...

9 Untitled document 38 facts | 2010-12-28
http://semanticweb.org/index.php?title=Special:Expo...

10 Untitled document 42 facts | 2010-02-19
http://semanticweb.org/id/Giovanni_Tummarello&fetch...

11 Untitled document 42 facts | 2010-03-03
http://semanticweb.org/id/Giovanni_Tummarello&fetch...

12 Untitled document 2 facts | 2010-07-30
http://www.bibsonomy.org/swrc/author/Giovanni+Tumma...

13 Giovanni Tummarello 17 facts | 2010-11-23
http://www.semanlink.net/tag/giovanni_tummarello

<- 1 2 ->

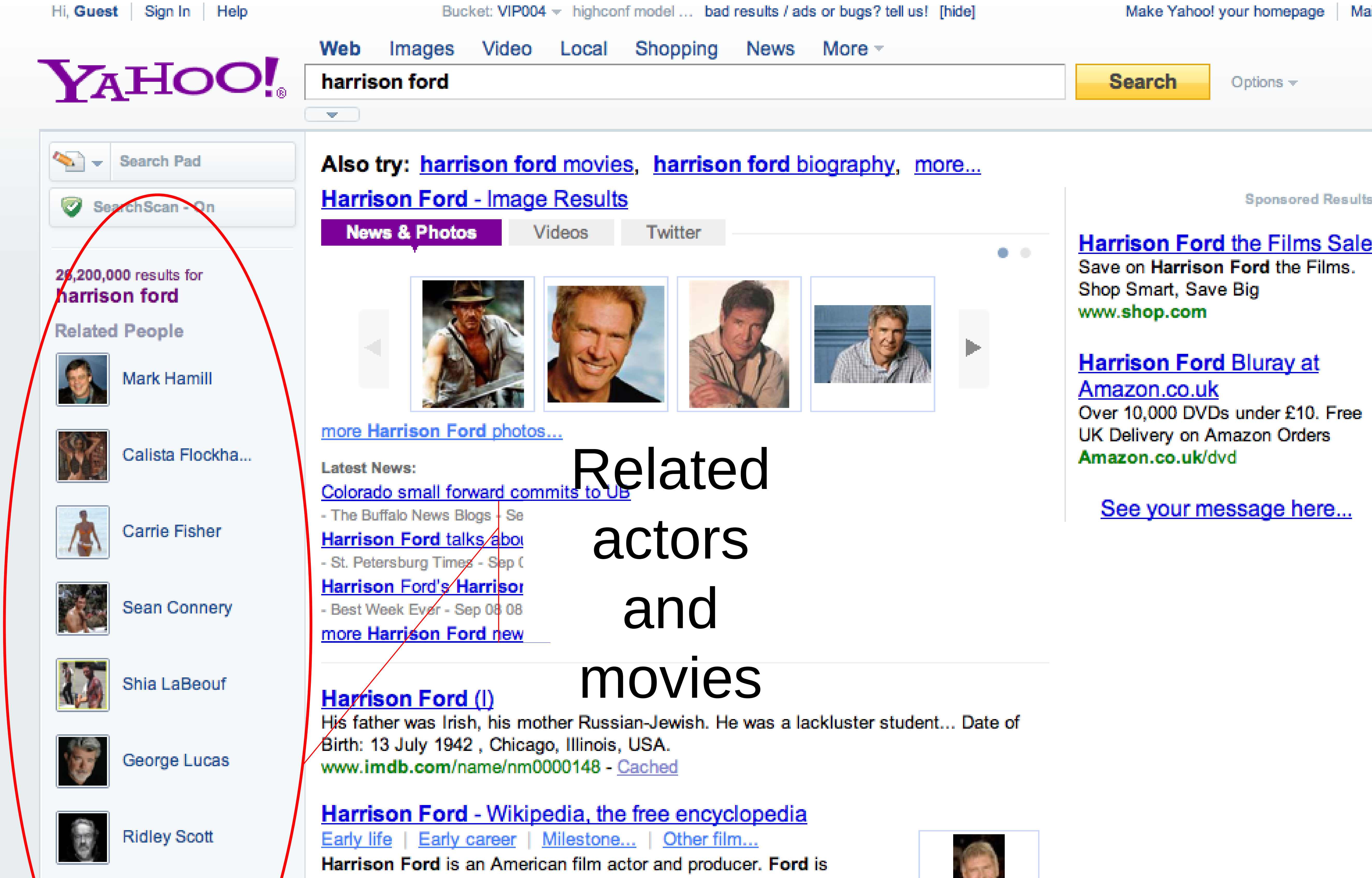
reject all

approve all

http://example.loc/document.rdf

add source url

Related entities



Hi, Guest | Sign In | Help

Bucket: VIP004 ▾ highconf model ... bad results / ads or bugs? tell us! [hide]

Make Yahoo! your homepage | Ma

harrison ford

28,200,000 results for **harrison ford**

Related People

- Mark Hamill
- Calista Flockha...
- Carrie Fisher
- Sean Connery
- Shia LaBeouf
- George Lucas
- Ridley Scott

Also try: harrison ford movies, harrison ford biography, more...

Harrison Ford - Image Results

more Harrison Ford photos...

Latest News:

Colorado small forward commits to UB

- The Buffalo News Blogs - Se

Harrison Ford talks abou

- St. Petersburg Times - Sep (

Harrison Ford's Harrisor

- Best Week Ever - Sep 08 08

more Harrison Ford new

Related actors and movies

Harrison Ford (I)

His father was Irish, his mother Russian-Jewish. He was a lackluster student... Date of Birth: 13 July 1942 , Chicago, Illinois, USA.

www.imdb.com/name/nm0000148 - Cached

Harrison Ford - Wikipedia, the free encyclopedia

Early life | Early career | Milestone... | Other film...

Harrison Ford is an American film actor and producer. **Ford** is

Sponsored Results

Harrison Ford the Films Sale

Save on **Harrison Ford** the Films. Shop Smart, Save Big

www.shop.com

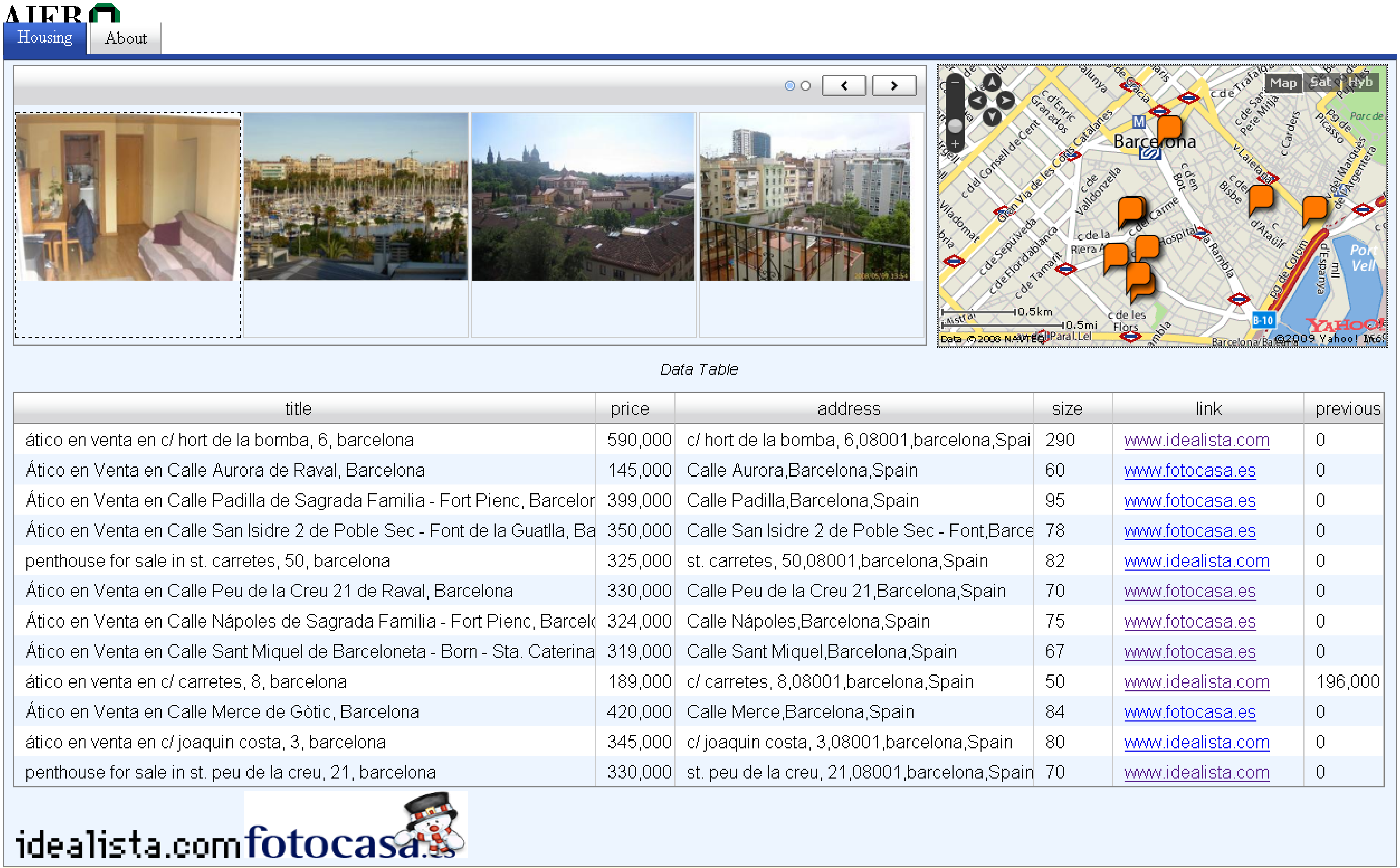
Harrison Ford Bluray at Amazon.co.uk

Over 10,000 DVDs under £10. Free UK Delivery on Amazon Orders

Amazon.co.uk/dvd

See your message here...

Adaptive presentation: housing search



Data Table

title	price	address	size	link	previous
ático en venta en c/ hort de la bomba, 6, barcelona	590,000	c/ hort de la bomba, 6,08001,barcelona,Spai	290	www.idealista.com	0
Ático en Venta en Calle Aurora de Raval, Barcelona	145,000	Calle Aurora,Barcelona,Spain	60	www.fotocasa.es	0
Ático en Venta en Calle Padilla de Sagrada Familia - Fort Pienc, Barcelor	399,000	Calle Padilla,Barcelona,Spain	95	www.fotocasa.es	0
Ático en Venta en Calle San Isidre 2 de Poble Sec - Font de la Guatlla, Ba	350,000	Calle San Isidre 2 de Poble Sec - Font,Barce	78	www.fotocasa.es	0
penthouse for sale in st. carretes, 50, barcelona	325,000	st. carretes, 50,08001,barcelona,Spain	82	www.idealista.com	0
Ático en Venta en Calle Peu de la Creu 21 de Raval, Barcelona	330,000	Calle Peu de la Creu 21,Barcelona,Spain	70	www.fotocasa.es	0
Ático en Venta en Calle Nápoles de Sagrada Familia - Fort Pienc, Barcelc	324,000	Calle Nápoles,Barcelona,Spain	75	www.fotocasa.es	0
Ático en Venta en Calle Sant Miquel de Barceloneta - Born - Sta. Caterina	319,000	Calle Sant Miquel,Barcelona,Spain	67	www.fotocasa.es	0
ático en venta en c/ carretes, 8, barcelona	189,000	c/ carretes, 8,08001,barcelona,Spain	50	www.idealista.com	196,000
Ático en Venta en Calle Merce de Gòtic, Barcelona	420,000	Calle Merce,Barcelona,Spain	84	www.fotocasa.es	0
ático en venta en c/ joaquin costa, 3, barcelona	345,000	c/ joaquin costa, 3,08001,barcelona,Spain	80	www.idealista.com	0
penthouse for sale in st. peu de la creu, 21, barcelona	330,000	st. peu de la creu, 21,08001,barcelona,Spain	70	www.idealista.com	0

Resources

- Books
  - Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press. 2011
- Survey papers
  - Thanh Tran, Peter Mika. Survey of Semantic Search Approaches. Under submission, 2012.
- Conferences and workshops
  - ISWC, ESWC, WWW, SIGIR, CIKM, SemTech
  - Semantic Search workshop series
  - Exploiting Semantic Annotations in Information Retrieval (ESAIR)
  - Entity-oriented Search (EOS) workshop

Plan

- XML und URIs
- Einleitung in RDF
- RDF Schema
- Logik – Grundlagen
- Semantik von RDF(S)
- SPARQL – Syntax und Intuition
- Semantik von SPARQL
- Linked Data
- Semantic Search
- **OWL – Syntax und Intuition I**
- OWL – Syntax und Intuition II
- OWL – Semantik und Reasoning
- Konjunktive Anfragen und Regelsprachen
- Applications

Attribution

- Slides erstellt von Thanh Tran, Peter Mika für das Tutorial “**Semantic Search**”
  - https://sites.google.com/site/kimducthanh/activity