Semantic Web Technologies II

SS 2009

03.06.2009

Semantic Search and Information Integration

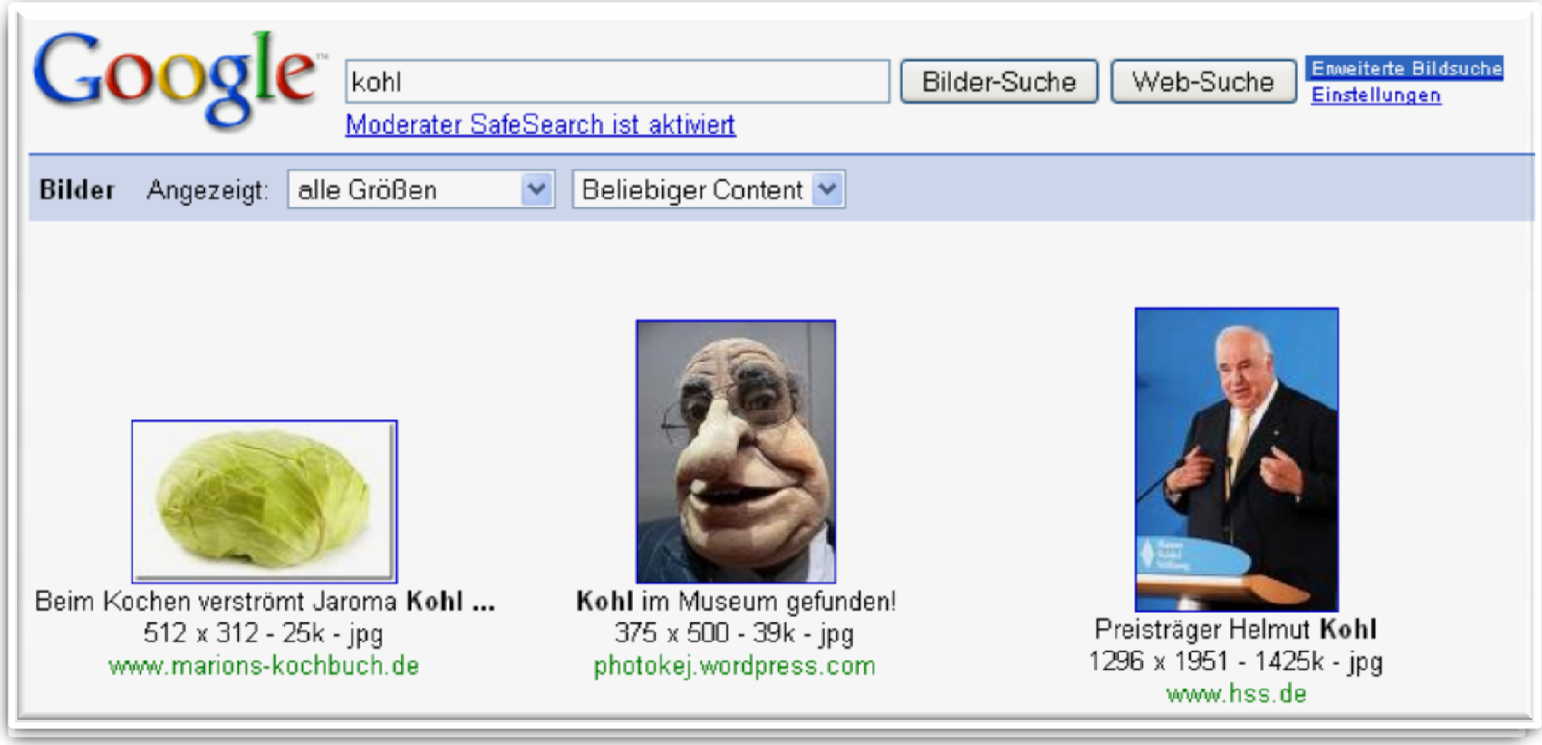
Dr. Sudhir Agarwal
Dr. Stephan Grimm
Dr. Peter Haase
PD Dr. Pascal Hitzler
Denny Vrandecic

Topics

- **Semantic Search**
  - **Overview**
  - Examples of Semantic Search Systems
  - Details of the Semantic Search Process
  - Ontology-based Interpretation of Keyword Queries
  - Natural Language Interfaces
- Information Integration
  - Ontology Mapping
  - Automated Mapping Discovery

“Traditional Search”

- Keyword-based
- “Meaning” of the information not interpreted by the machine



kohl

Moderater SafeSearch ist aktiviert

Bilder Angezeigt: alle Größen Beliebiger Content

Beim Kochen verströmt Jaroma **Kohl ...**
512 x 312 - 25k - jpg
www.marions-kochbuch.de

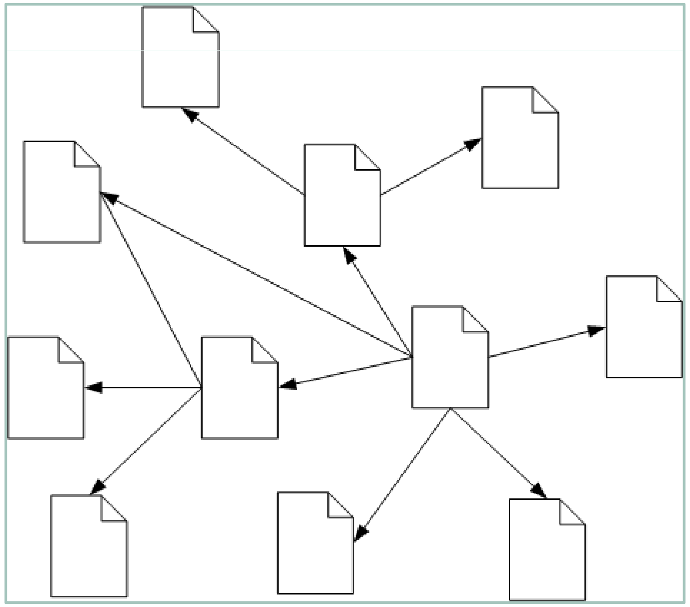
Kohl im Museum gefunden!
375 x 500 - 39k - jpg
photokej.wordpress.com

Preisträger Helmut **Kohl**
1296 x 1951 - 1425k - jpg
www.hss.de

Search on the Semantic Web … is about Textual & Structured Data!

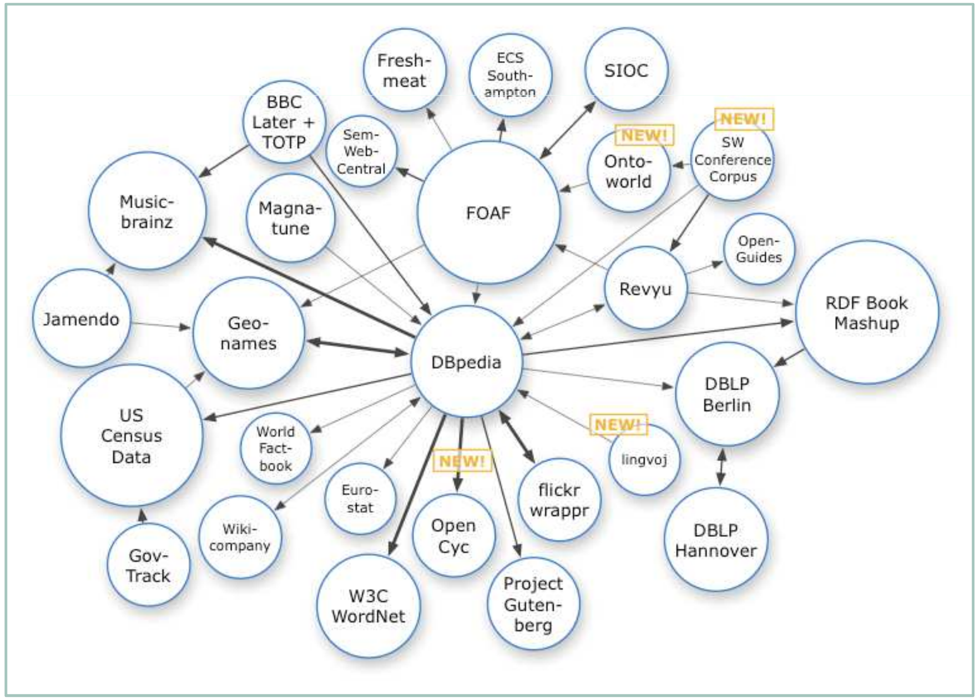
A Web of documents

- Billions of web pages
- Number ever increasing



A Web of data

- Deep Web
- Publicly available ontologies
- Open Linked Data
- More than 5 billion triples



Search from Information Retrieval Perspective

Information Retrieval Model is a quadruple $\langle D, Q, F, R(q_i, d_j) \rangle$

- *D* is a set composed of views (representations) for the **resources** (documents) in the collection.
- *Q* is a set composed of views (representations) for the **user information needs** called queries.
- *F* is a framework for modeling **resource representations, queries and their relationships**.
- ***R(Q*$_i$*,D*$_j$*)*** is a **ranking** function which associates a real number with a query ***Q***$_i$ and document representation ***D***$_j$
- Such ranking defines an ordering among the documents with regard to the query.

- Search based on classical Information Retrieval
  - Resources are text documents
  - User needs (queries) expressed as keyword
  - Simple syntactic matching of keywords against documents

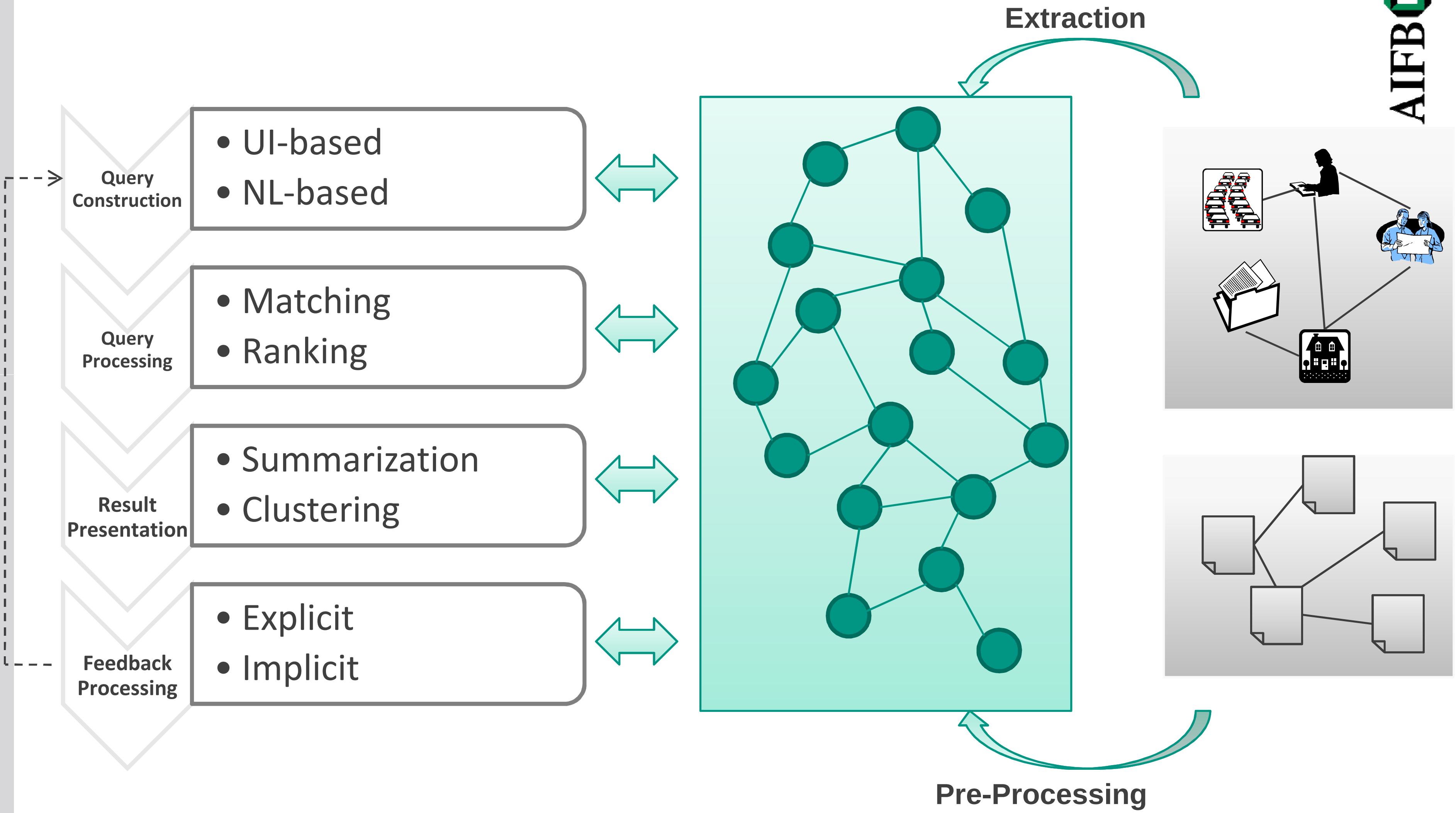
What do we mean by Semantic Search?

- Working definition:

"Semantic Search is a process of information access, where one or several activities can be supported by a set of functionalities enabled by semantic technologies"

- Terminology:
  - Information: documents vs. facts
  - Semantic technologies: knowledge extraction, knowledge representation, reasoning
  - Ontology-enhanced IR vs. ontology-based IR

Semantic Search as a Process



Query Construction
- UI-based
- NL-based

Query Processing
- Matching
- Ranking

Result Presentation
- Summarization
- Clustering

Feedback Processing
- Explicit
- Implicit

Extraction

Pre-Processing

Topics

- **Semantic Search**
  - Overview
  - **Examples of Semantic Search Systems**
  - Details of the Semantic Search Process
  - Ontology-based Interpretation of Keyword Queries
  - Natural Language Interfaces

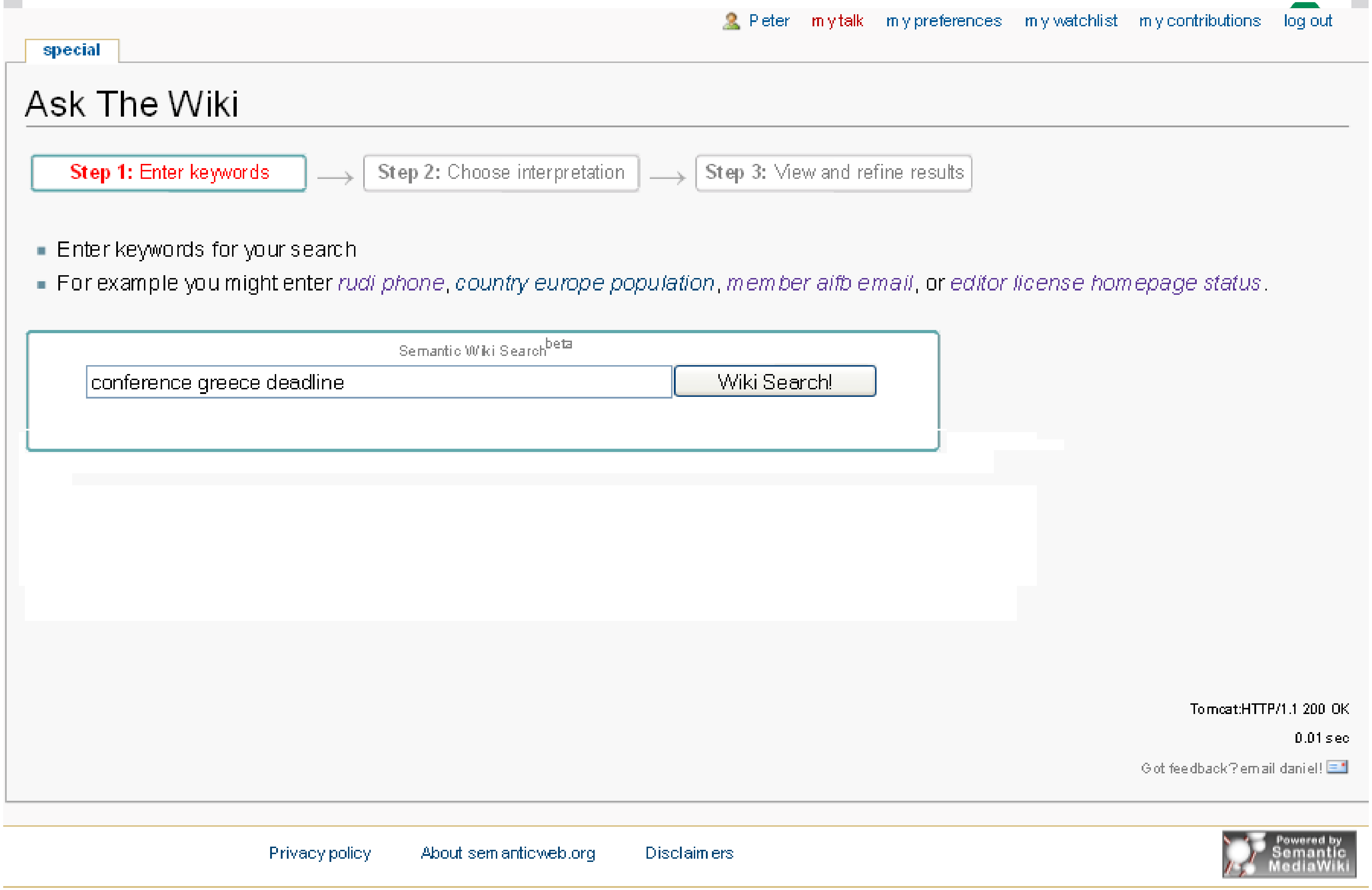
- Information Integration
  - Ontology Mapping
  - Automated Mapping Discovery

Example: Semantic Wiki Search

- Operates on
  - Semi-structured wiki-data (OWL/RDF export)

1. Query construction
   - User expresses information needs in keyword queries
   - Automated translation into structured, conjunctive queries (SPARQL)
2. Query processing
   - User selects intended interpretation
   - Query processing using a triple store
3. Result presentation
   - Structured answers as results (tuple sets)
4. Feedback Processing
   - Faceted browsing to refine queries and results

1. Entering Keywords



Peter my talk my preferences my watchlist my contributions log out

special

Ask The Wiki

Step 1: Enter keywords → **Step 2:** Choose interpretation → **Step 3:** View and refine results

- Enter keywords for your search
- For example you might enter *rudi phone*, *country europe population*, *member aifb email*, or *editor license homepage status*.

Semantic Wiki Search^{beta}

conference greece deadline

Wiki Search!

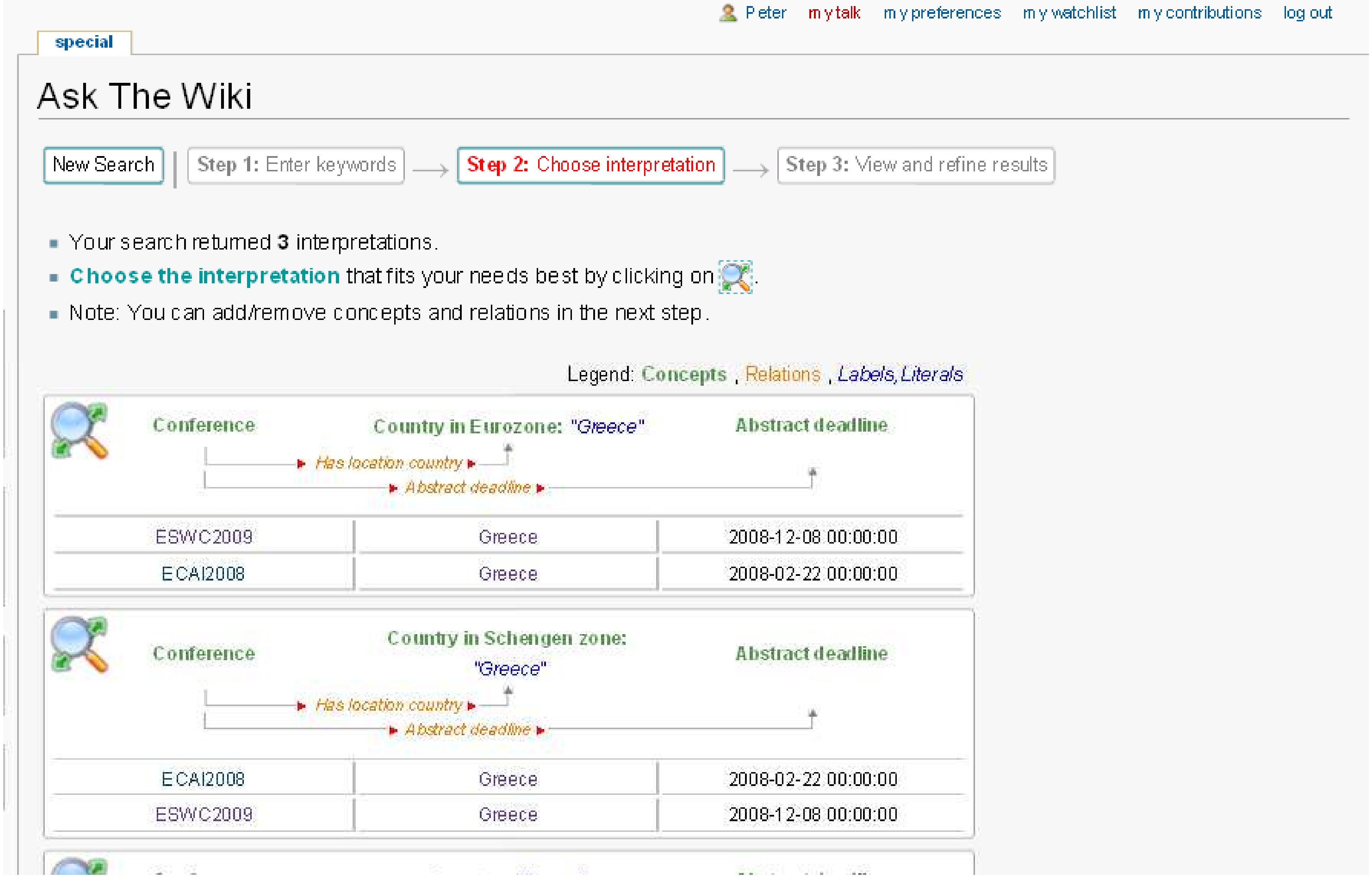
Tomcat:HTTP/1.1 200 OK

0.01 sec

Got feedback? email daniel!

Privacy policy About semanticweb.org Disclaimers

2. Choosing the Interpretation



Peter my talk my preferences my watchlist my contributions log out

special

Ask The Wiki

New Search | Step 1: Enter keywords → **Step 2: Choose interpretation** → Step 3: View and refine results

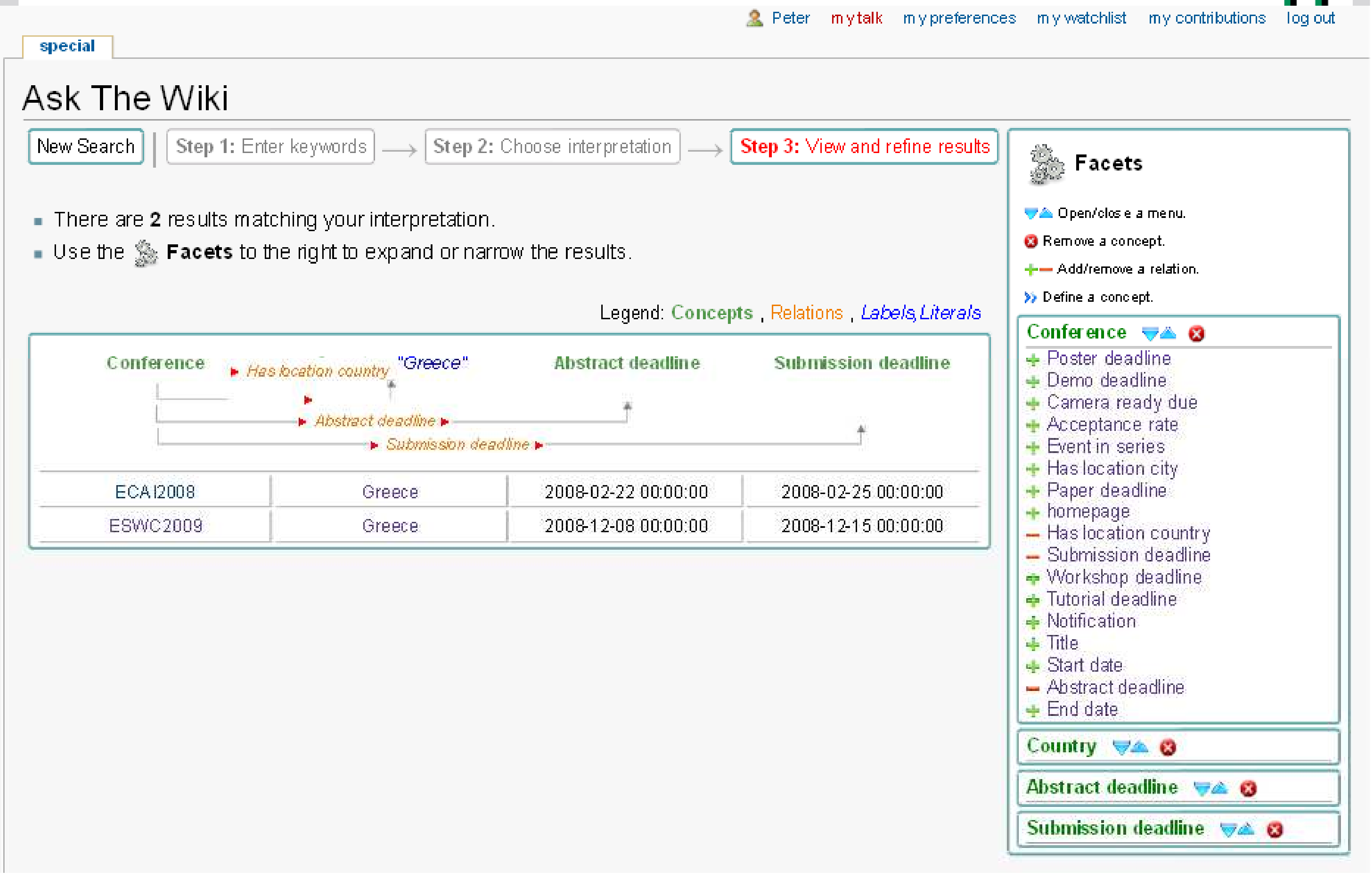
- Your search returned **3** interpretations.
- **Choose the interpretation** that fits your needs best by clicking on 🔍.
- Note: You can add/remove concepts and relations in the next step.

Legend: Concepts , Relations , *Labels, Literals*

Conference	Country in Eurozone: *"Greece"*	Abstract deadline
ESWC2009	Greece	2008-12-08 00:00:00
ECAI2008	Greece	2008-02-22 00:00:00

Conference	Country in Schengen zone: *"Greece"*	Abstract deadline
ECAI2008	Greece	2008-02-22 00:00:00
ESWC2009	Greece	2008-12-08 00:00:00

3. Result Presentation and Refinement



Peter my talk my preferences my watchlist my contributions log out

special

Ask The Wiki

New Search | Step 1: Enter keywords → Step 2: Choose interpretation → **Step 3: View and refine results**

- There are **2** results matching your interpretation.
- Use the **Facets** to the right to expand or narrow the results.

Legend: Concepts , Relations , *Labels, Literals*

Conference	Has location country	Abstract deadline	Submission deadline
ECAI2008	Greece	2008-02-22 00:00:00	2008-02-25 00:00:00
ESWC2009	Greece	2008-12-08 00:00:00	2008-12-15 00:00:00

Facets

▼▲ Open/close a menu.

☒ Remove a concept.

+– Add/remove a relation.

» Define a concept.

Conference ▼▲ ☒

- + Poster deadline
- + Demo deadline
- + Camera ready due
- + Acceptance rate
- + Event in series
- + Has location city
- + Paper deadline
- + homepage
- – Has location country
- – Submission deadline
- + Workshop deadline
- + Tutorial deadline
- + Notification
- + Title
- + Start date
- – Abstract deadline
- + End date

Country ▼▲ ☒

Abstract deadline ▼▲ ☒

Submission deadline ▼▲ ☒

Sample Types of Queries

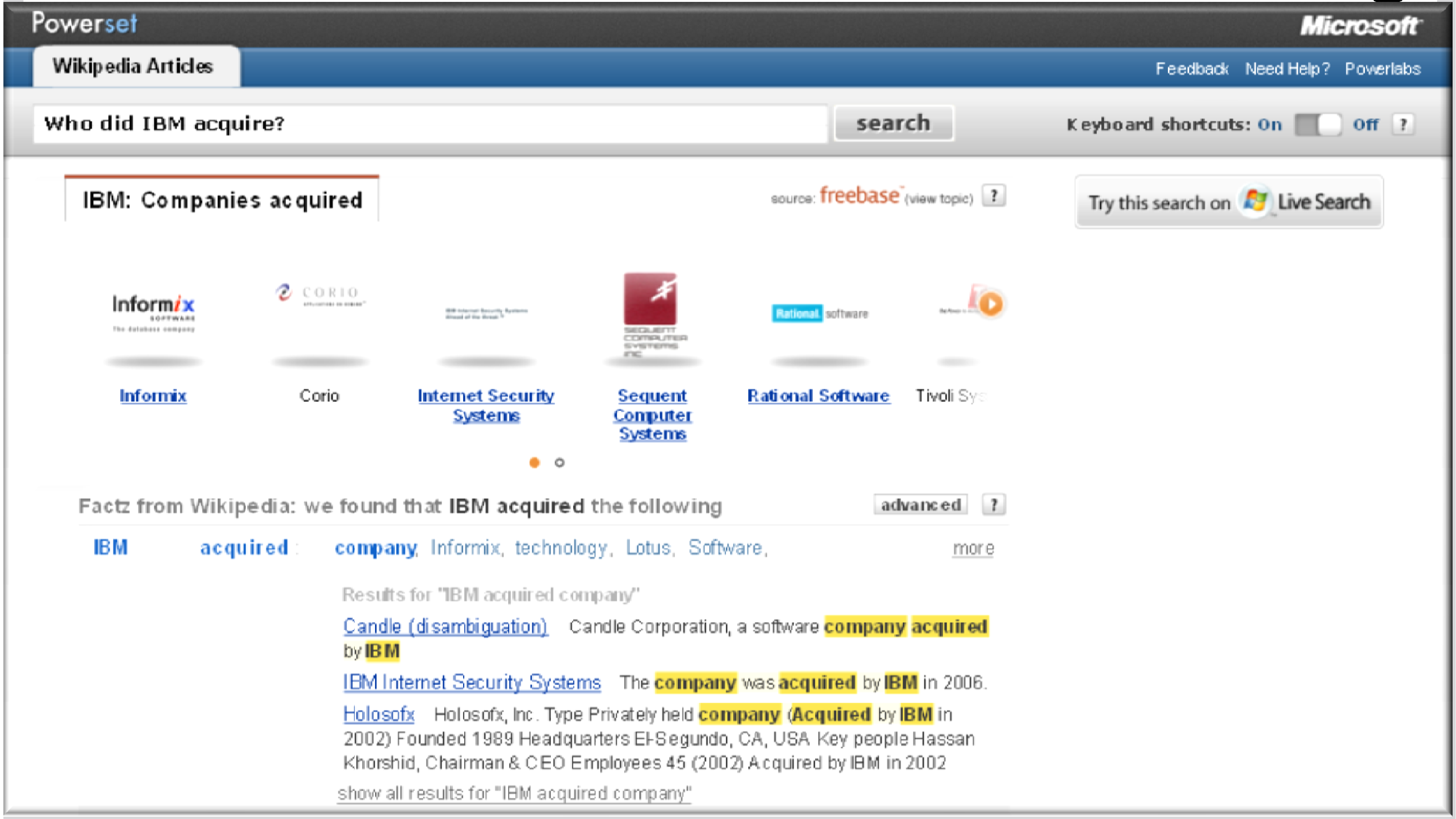
- Search for entities (pages)
  - Find the page describing Rudi Studer
  - ➢ „Rudi Studer“
- Search for single facts
  - Find the capital of Germany
  - ➢ „capital Germany“
- Search for sets of tuples
  - Conference planning: Find conferences taking place in Greece and their deadlines
  - ➢ „conference Greece deadline“

Example: Powerset

- Operates on
  - Un-structured web documents

1. Query construction
   - User expresses information needs in natural language queries
   - Deep linguistic analysis of content and queries
2. Query processing
   - Matching of linguistic structures
3. Result presentation and refinement
   - Structured answers as results (lists of entities)
4. Feedback Processing
   - Faceted browsing to refine queries and results

Powerset



Powerset
Microsoft
Wikipedia Articles
Feedback Need Help? Powerlabs
search
Keyboard shortcuts: On Off ?
source: freebase (view topic) ?
Try this search on Live Search
advanced ?

Who did IBM acquire?

IBM: Companies acquired

Informix | Corio | Internet Security Systems | Sequent Computer Systems | Rational Software | Tivoli Sys

Factz from Wikipedia: we found that IBM acquired the following

IBM acquired: company, Informix, technology, Lotus, Software, more

Results for "IBM acquired company"

Candle (disambiguation) Candle Corporation, a software company acquired by IBM

IBM Internet Security Systems The company was acquired by IBM in 2006.

Holosofx Holosofx, Inc. Type Privately held company (Acquired by IBM in 2002) Founded 1989 Headquarters El-Segundo, CA, USA Key people Hassan Khorshid, Chairman & CEO Employees 45 (2002) Acquired by IBM in 2002

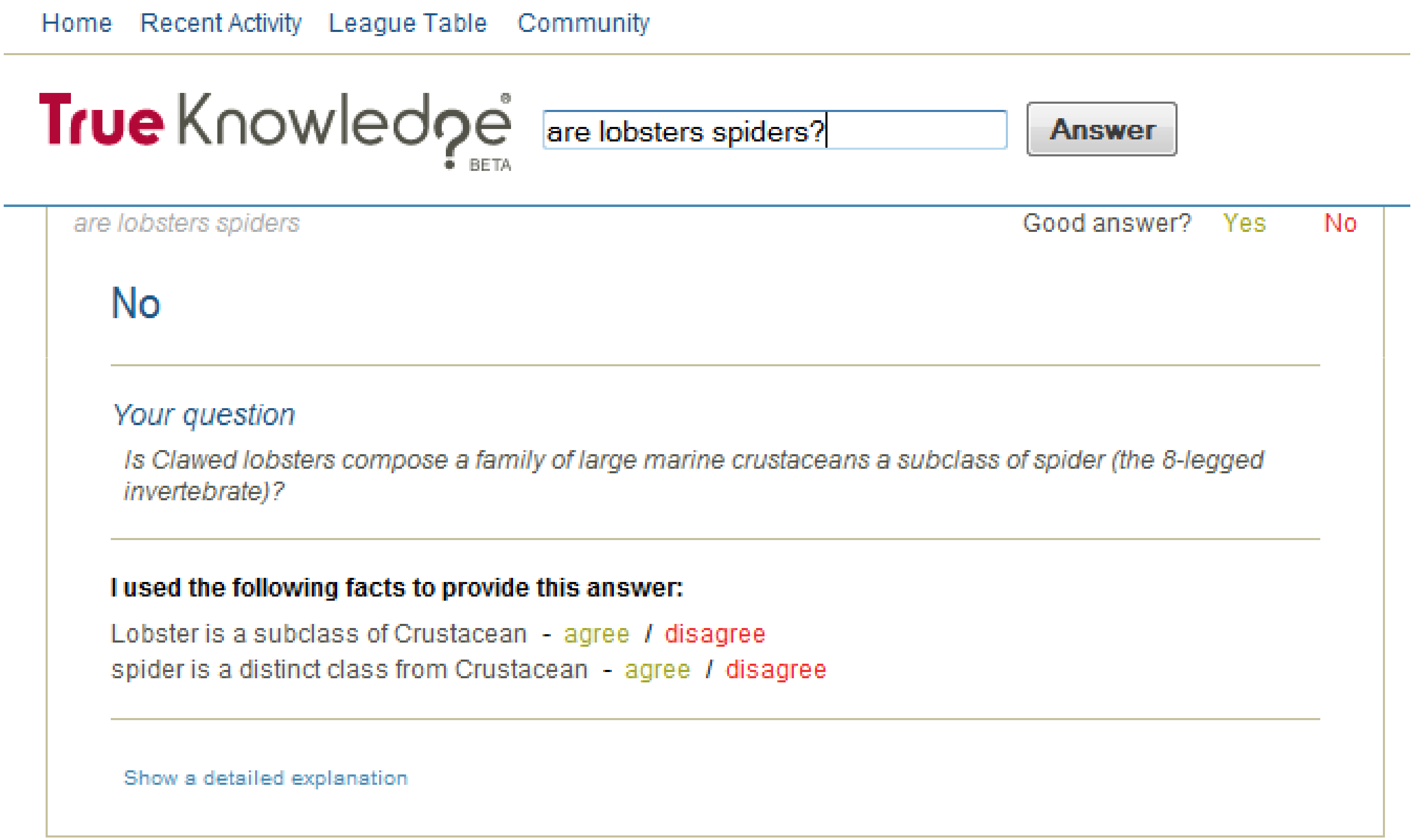
show all results for "IBM acquired company"

Example: TrueKnowledge

- Operates on
  - A structured knowledge base

1. Query construction
   - User expresses information needs in natural language queries
   - Deep linguistic analysis queries, translation to logical formulae
2. Query processing
   - Logical entailment: KB |= Q
   - Explanations: minimal subset KB' |= Q
3. Result presentation and refinement
   - Binary answers, facts
4. Feedback Processing
   - Confirmation of facts in the knowledge base

True Knowledge



are lobsters spiders?

are lobsters spiders

Good answer? Yes No

No

Your question

Is Clawed lobsters compose a family of large marine crustaceans a subclass of spider (the 8-legged invertebrate)?

I used the following facts to provide this answer:

Lobster is a subclass of Crustacean - agree / disagree

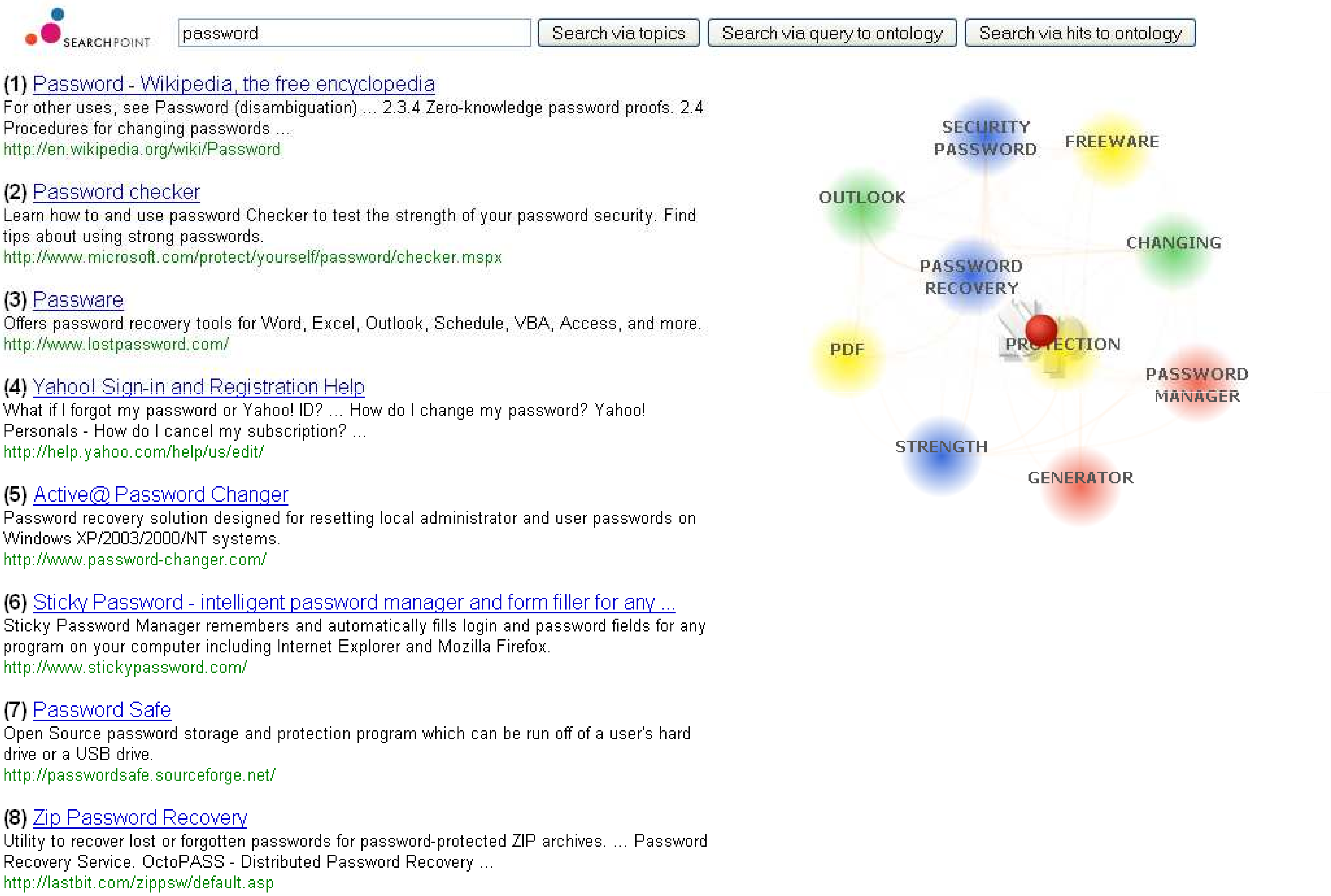
spider is a distinct class from Crustacean - agree / disagree

Show a detailed explanation

Beispiel: Searchpoint

- Inhalte
    - Google (unstrukturierte Web-Dokumente)
- Erzeugen der Anfrage
    - Google (keyword-basierte Anfragen)
- Bearbeiten der Anfrage
    - Google
- Präsentation der Ergebnisse
    - Google Ergebnislisten
    - Clustering/Klassifikation der Ergebnisse
- Auswertung von Feedback
    - Auswahl relevanter Klassen/Cluster durch den Nutzer

SearchPoint



(1) Password - Wikipedia, the free encyclopedia
For other uses, see Password (disambiguation) ... 2.3.4 Zero-knowledge password proofs. 2.4 Procedures for changing passwords ...
http://en.wikipedia.org/wiki/Password

(2) Password checker
Learn how to and use password Checker to test the strength of your password security. Find tips about using strong passwords.
http://www.microsoft.com/protect/yourself/password/checker.mspx

(3) Passware
Offers password recovery tools for Word, Excel, Outlook, Schedule, VBA, Access, and more.
http://www.lostpassword.com/

(4) Yahoo! Sign-in and Registration Help
What if I forgot my password or Yahoo! ID? ... How do I change my password? Yahoo! Personals - How do I cancel my subscription? ...
http://help.yahoo.com/help/us/edit/

(5) Active@ Password Changer
Password recovery solution designed for resetting local administrator and user passwords on Windows XP/2003/2000/NT systems.
http://www.password-changer.com/

(6) Sticky Password - intelligent password manager and form filler for any ...
Sticky Password Manager remembers and automatically fills login and password fields for any program on your computer including Internet Explorer and Mozilla Firefox.
http://www.stickypassword.com/

(7) Password Safe
Open Source password storage and protection program which can be run off of a user's hard drive or a USB drive.
http://passwordsafe.sourceforge.net/

(8) Zip Password Recovery
Utility to recover lost or forgotten passwords for password-protected ZIP archives. ... Password Recovery Service. OctoPASS - Distributed Password Recovery ...
http://lastbit.com/zippsw/default.asp

Topics

- **Semantic Search**
  - Overview
  - Examples of Semantic Search Systems
  - **Details of the Semantic Search Process**
  - Ontology-based Interpretation of Keyword Queries
  - Natural Language Interfaces
- Information Integration
  - Ontology Mapping
  - Automated Mapping Discovery

Query Construction – Task

- **Task:** “Query construction is the first step in the information access process that concerns with assisting the user in the **specification of the information need.**”
- **Result:** “A representation of the information need in terms of primitives of a language supported by the system.”
- Language supported by the system
  - keyword-based queries
  - database queries (SQL)
  - XML-based query language (XQuery, XPath, XML Fragments)
  - Semantic Web query languages (SPARQL, DL conjunctive queries, F-logic)
- Enhancement
  - user query is incomplete representation of the user information need
  - modification through expansion, disambiguation and refinement to achieve a more complete and precise representation of the need
- Translation
  - transformation of keywords- or NL-queries to a formal query

Query Construction – Advanced Approaches

- Interpretation of keywords
  - Keywords map to index containing ontology elements
  - Use **graph exploration** to compute possible connections
  - Map connections to elements of formal query
- Interpretation of Natural Language Queries using domain knowledge
  - Deep parsing query to obtain Part of Speeches (PoS)
  - **PoS maps** against element types of the knowledge base
  - Query elements are mapped against knowledge base instances
  - Typically requires rich lexical models

Query Processing – Task

- **Task:** “Query processing is a step in the information access process where the **need** as specified in the user query **is matched against the system resource model** so as to retrieve the (documents containing the) relevant information.”
- **Result**
  - “(A list of ranked documents containing) the information that satisfy the user need.”
  - Topical information need → documents about some topics (Document Retrieval)
  - Focused information need → document parts, e.g. section, passage (Focussed IR)
  - Exact information need → an answer to a question (Question Answering)
- Matching procedure is dictated by query and document representation
- “Terms-only” matching
  - Keywords-based queries / “Bag of words” resource models
  - Statistical IR approaches such as vector space model, probabilistic model, DFR etc.
- Incorporating syntactic information
  - Syntactic properties of the text language → Language Model
  - Syntactic properties of the query and resource representation → XML retrieval
- Incorporating semantic information
  - Representation of query and resources enhanced with ontology elements → ontology-enhanced IR
  - Representation of query and resources based on ontology elements → ontology-based IR

Result Presentation – Task

Ranking

- **Task:** “Finding appropriate **measures of relatedness** and use them to **rank results.**”
- **Result:** “A score for each of the results is calculated and results are sorted accordingly.”
- Kinds of relatedness
  - Lexical nearness of terms →synonyms, hyponyms, etc.
  - Topical nearness → buzzwords (politics, weather, etc)
  - Structural nearness → classical distance measure based on bag of words model
  - Ontological nearness → Concepts, Relations, Attributes, Instances
  - Heuristics → hyperlinks (Google’ page rank); anchor text, meta tag, page title,
- Ontology-Driven Semantic Ranking for natural language disambiguation
  - Using conceptual distance
    - Minimal path between concepts
    - Distance to common super concept

Result Presentation – Task

Clustering

- **Task:** “Group and label a given collection of patterns into meaningful clusters.”
- **Result:** “Thematically related documents are grouped together in the same clusters.”
- Clustering with background knowledge
  - Background ontology where terms matched to ontology elements
  - Integration of ontological knowledge in vector model, i.e. extending document term vector with additional dimensions representing ontological knowledge

Feedback Processing – Task

- **Task:** “The processing of user feedbacks aims to **exploit information from the interactions** to further satisfy the user information need.”
- Relevance feedback as standard paradigm
  - use information about which results of a query are perceived relevant for
    - – tuning system parameters
    - – suggesting new query (query refinement)
  - Feedback can be explicit or implicit (user behaviour)
  - Example: augment query with terms of relevant documents
  - Example: use feedback for disambiguation of query terms

Pre-Processing and Extraction – Task

- **Task:** “Pre-processing and extraction are **offline** tasks that are required to **develop a representation of the resources** available in the system.”
- **Result:** “A model supported by the system that captures the information content contained in the resources.”
- Model supported by the system
  - Bag-of-words
  - Structured model representation (XML documents)
  - Ontology-based model representation
- The more sophisticated the system resource model, the harder is pre-processing and extraction
  - Bag-of-words mostly developed using
    - tokenization
    - Lemmatization only.
  - Identification of syntactic and semantic information from the resources
    - PoS parsing
    - Extraction of instances, relations and expressive axioms

Pre-Processing and Extraction – Approaches

- Domain-Specific
  - more sophisticated domains (e.g. Chemistry)
- Semi-Supervised Information Extraction
  - Deriving Taxonomies
  - Relation Extraction from the Web
  - Mining Wikipedia
- „Open" Information Extraction
  - extract relations that are not user-defined
- Linguistically heavy approaches
  - Deep parsing

Topics

- **Semantic Search**
  - Overview
  - Examples of Semantic Search Systems
  - Details of the Semantic Search Process
  - **Ontology-based Interpretation of Keyword Queries**
  - Natural Language Interfaces
- Information Integration
  - Ontology Mapping
  - Automated Mapping Discovery

Motivation – Complex Information Needs

- Complex Information Need – a scenario

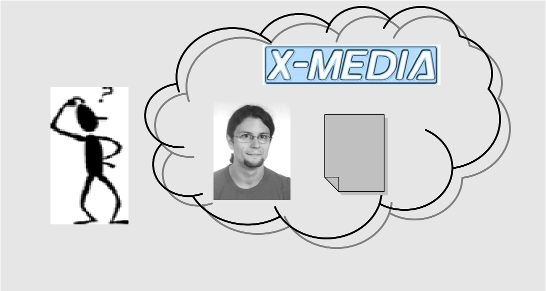
  “A user is searching the publications of the research institute AIFB using the information portal http://www.aifb.uni-karlsruhe.de *. He might look for a* ***publication*** *that*

  1. *was written by an* ***author*** *of the knowledge management research group,*
  2. *deals with the* ***topic*** *of information retrieval and*
  3. ***describes*** *a question answering system deployed in a corporate setting.”*

- Answering such an information need requires more expressiveness
  - More completely interpret the information need
  - **More precisely capture information about resources**

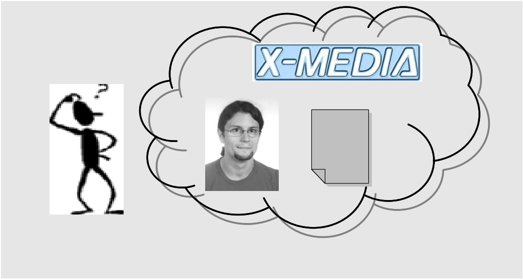
Ontology-based Query Interpretation

User Mental Model

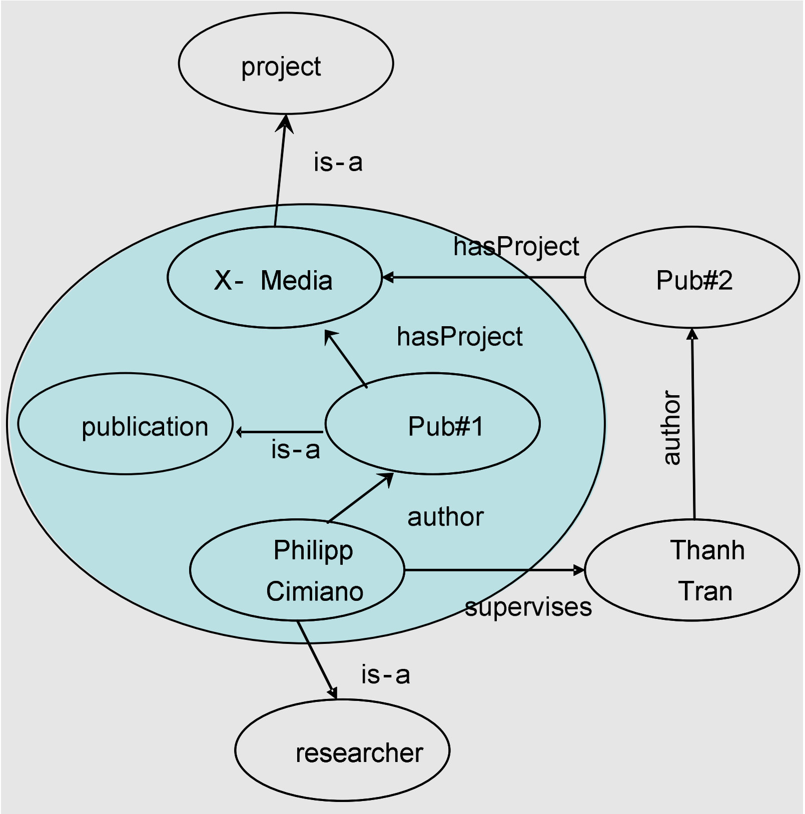


Ontology-based Query Interpretation

USER MENTAL MODEL

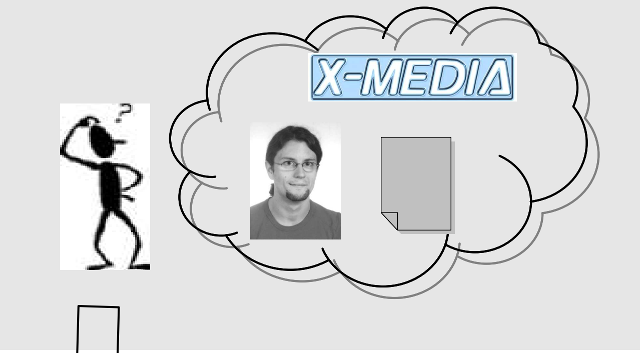


RESOURCE MODEL



Ontology-based Query Interpretation

USER MENTAL MODEL



Query Specification

USER QUERY

„Philipp Cimiano X-Media publications“

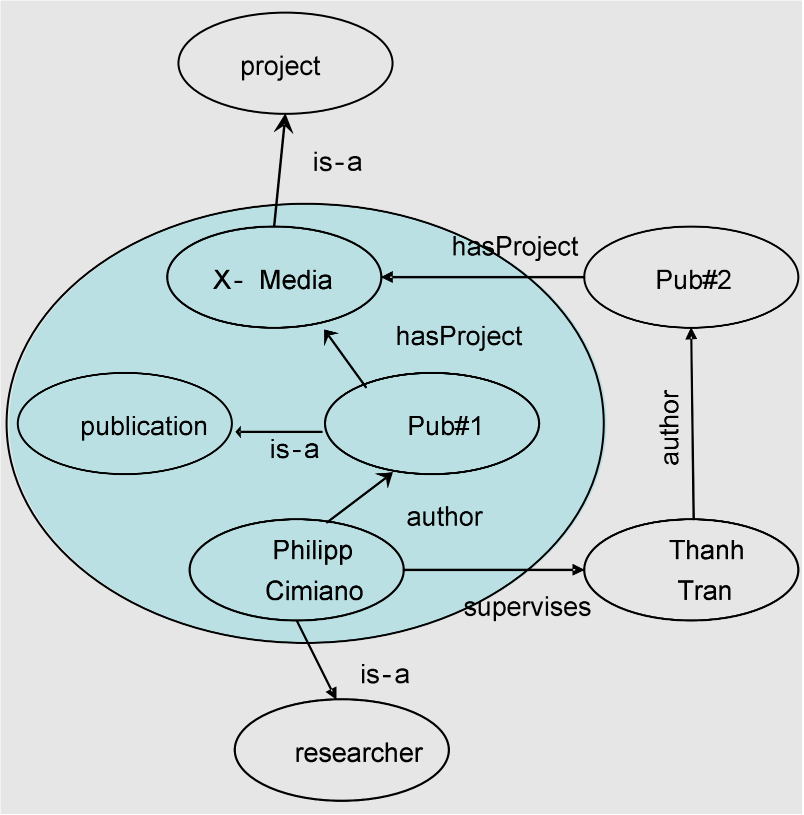
Query Interpretation

SYSTEM QUERY

Query Processing

?z <z , Philipp Cimiano>: author
<z , X-Media>: hasProject
<z , publication>: type

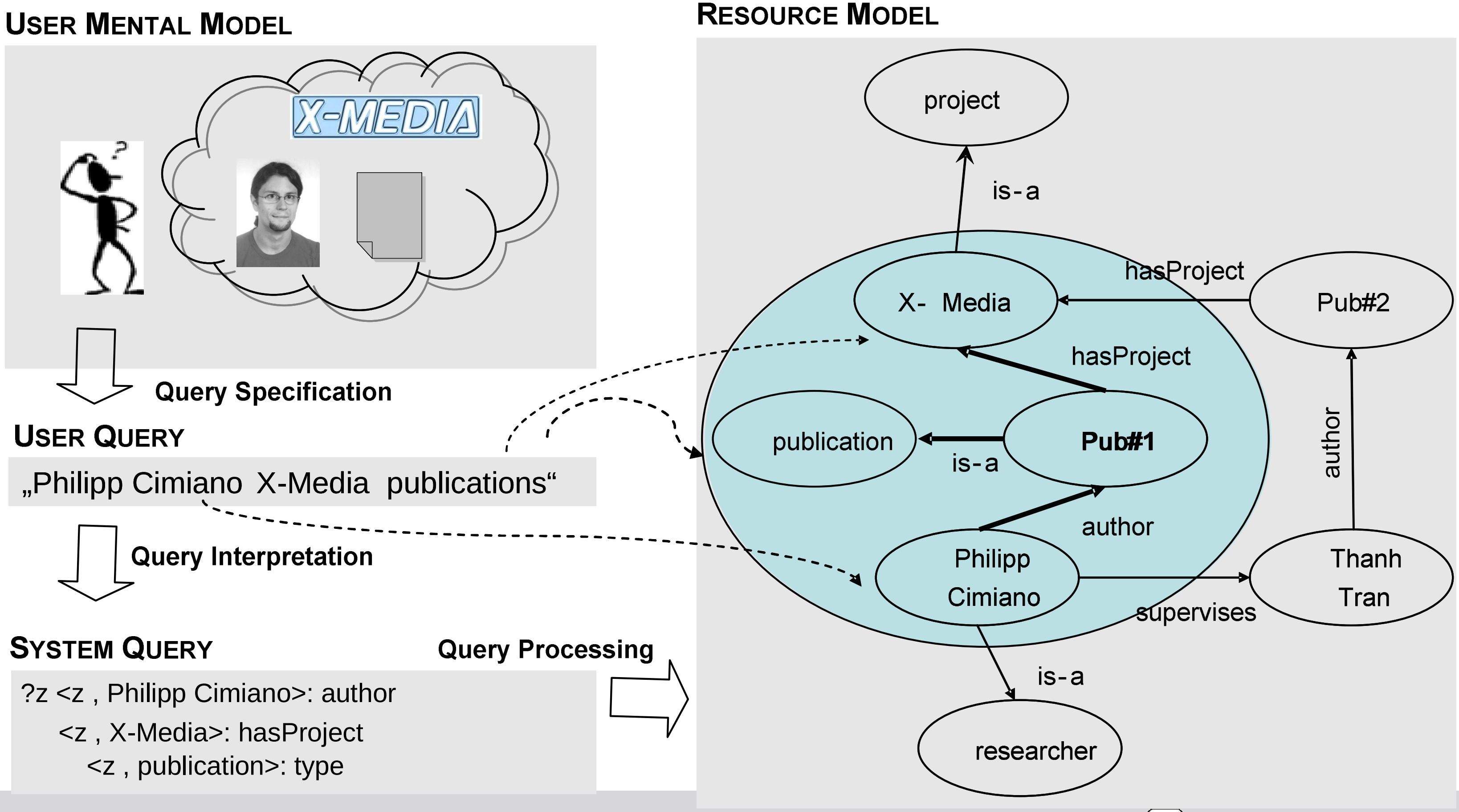
RESOURCE MODEL



Ontology-based Query Interpretation

Procedure

- Map user query elements to ontology elements
- Explore ontology elements to find connections
- Derive system query from connections



USER MENTAL MODEL

Query Specification

USER QUERY

„Philipp Cimiano X-Media publications“

Query Interpretation

SYSTEM QUERY

Query Processing

```
?z <z , Philipp Cimiano>: author
    <z , X-Media>: hasProject
      <z , publication>: type
```

RESOURCE MODEL

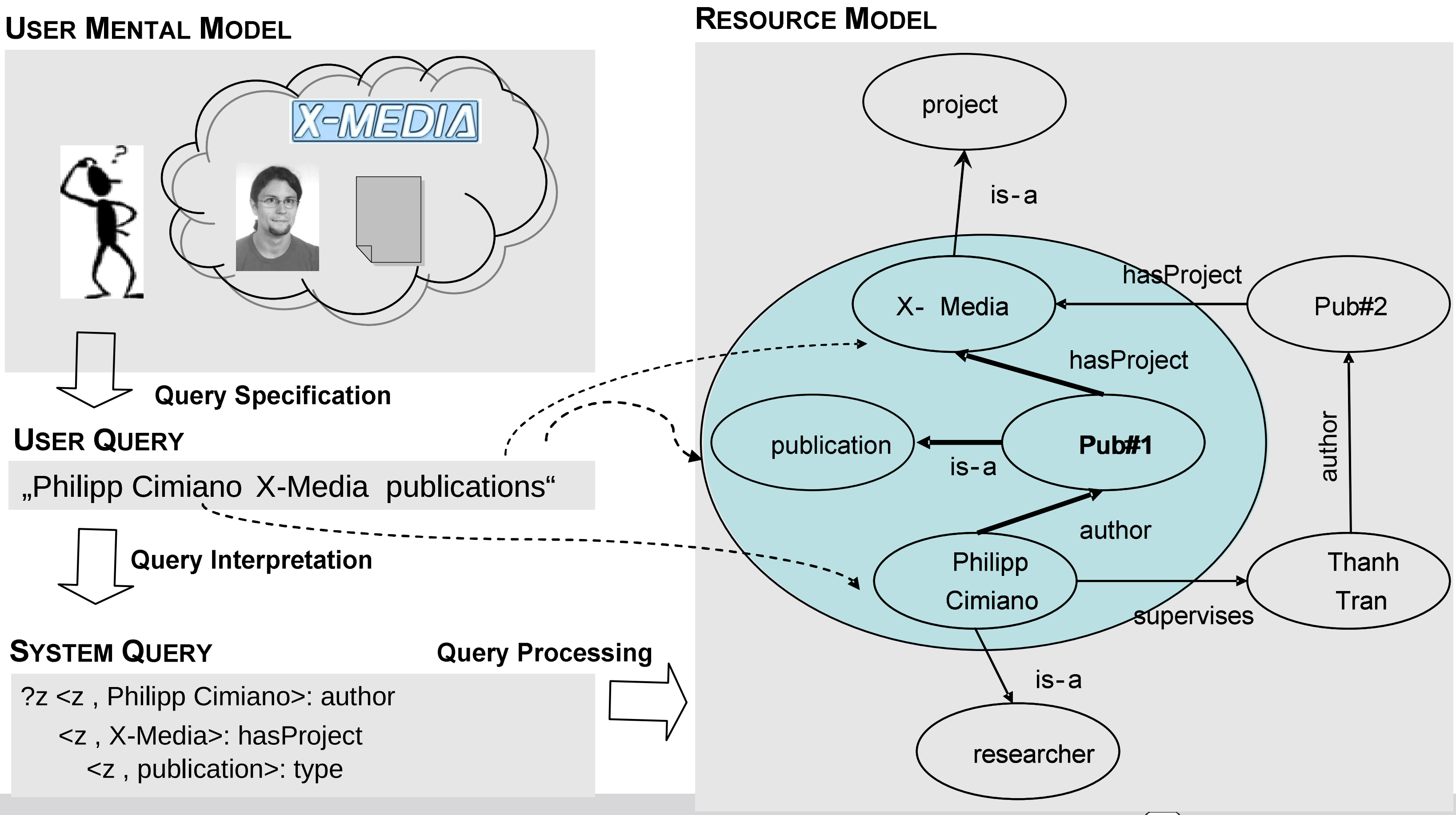
Ontology-based Query Interpretation

Procedure

- Map user query elements to ontology elements
- Explore ontology elements to find connections
- Derive system query from connections

Assumptions

- Ontology-Mental Correspondence
- Locality of Information Need



USER MENTAL MODEL

Query Specification

USER QUERY

„Philipp Cimiano X-Media publications“

Query Interpretation

SYSTEM QUERY

Query Processing

?z <z , Philipp Cimiano>: author
<z , X-Media>: hasProject
<z , publication>: type

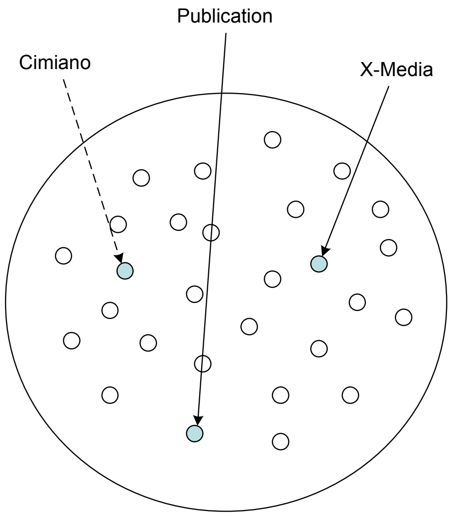
RESOURCE MODEL

Interpretation of Keywords as Conjunctive Queries

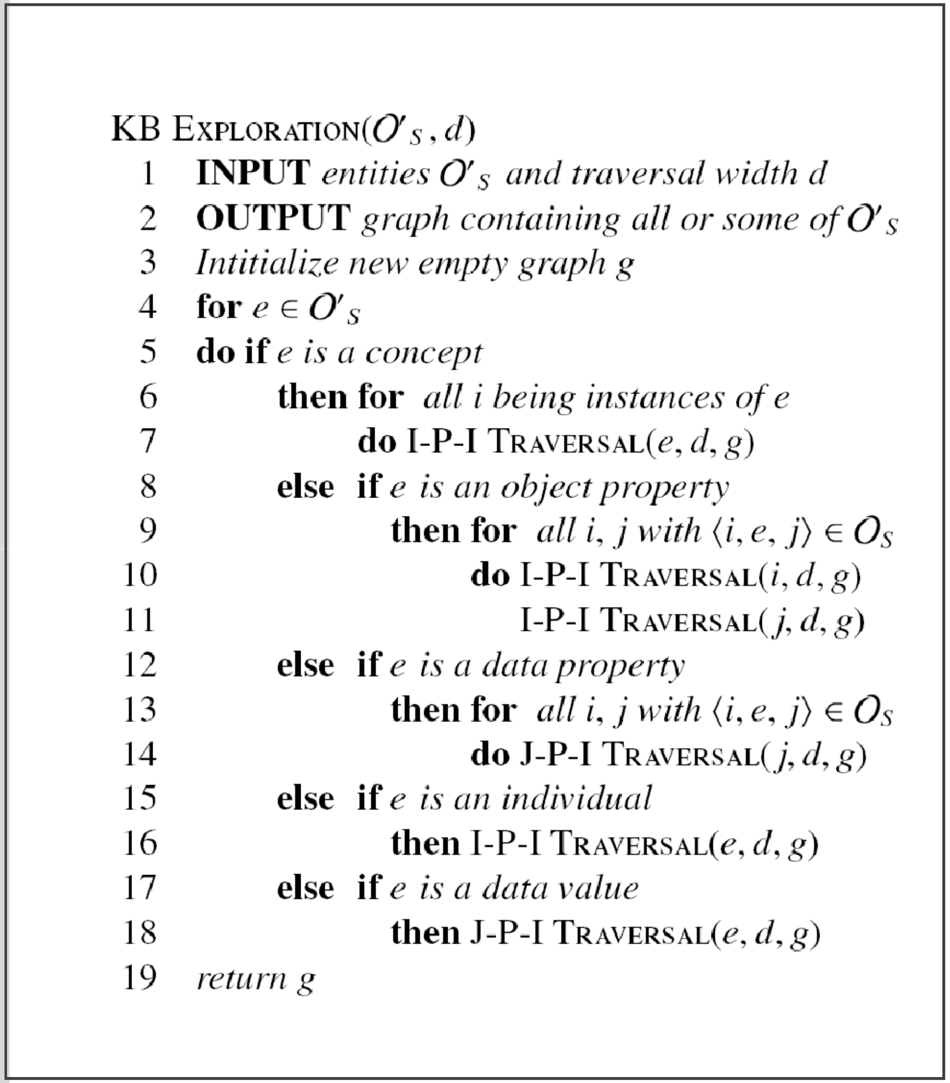
- An instantiation of the generic IR model
- User question **a set of keywords** $Q_U = (k_1, k_2, ..., k_n)$
- System query
  - **Conjunction of terms** of the form $x : C$ and $\langle x, y \rangle : R$
  - Where C is a concept, R is a role, and x, y are variables or individuals taken from a set of variable names, or a set of individual names
- System resource model
  - OWL DL knowledge base
  - **Ontology entities:** sets of individuals, data values, concepts, data ranges, object properties and data properties
  - **Connections between entities** are captured by terminological axioms and assertions (concept and property membership)

Step 1 – Mapping Keywords to Ontology Entities

- Match
  - keywords against ontology entities
- “Robust” matching functions
  - **Syntactic variants**
  - **Spelling variants**
- Matching function
  - Index of ontology elements
  - **Fuzzy search** on index with each keyword
  - Return ontology entities ranked according to syntactic similarity

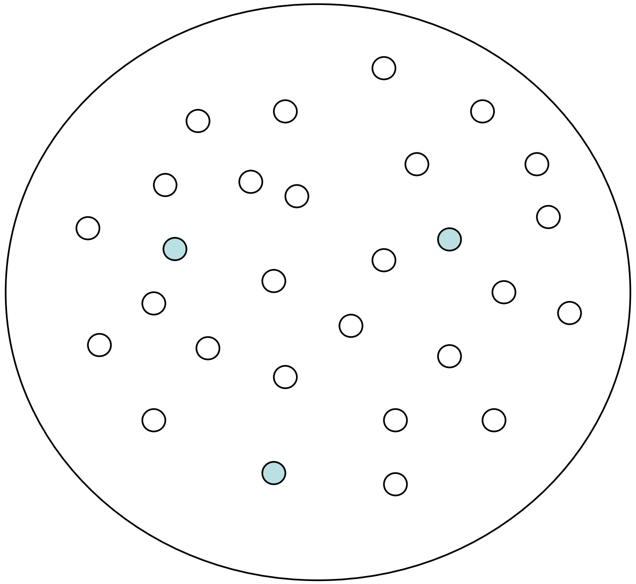


Step 2 – Exploring Connections among Ontology Entities

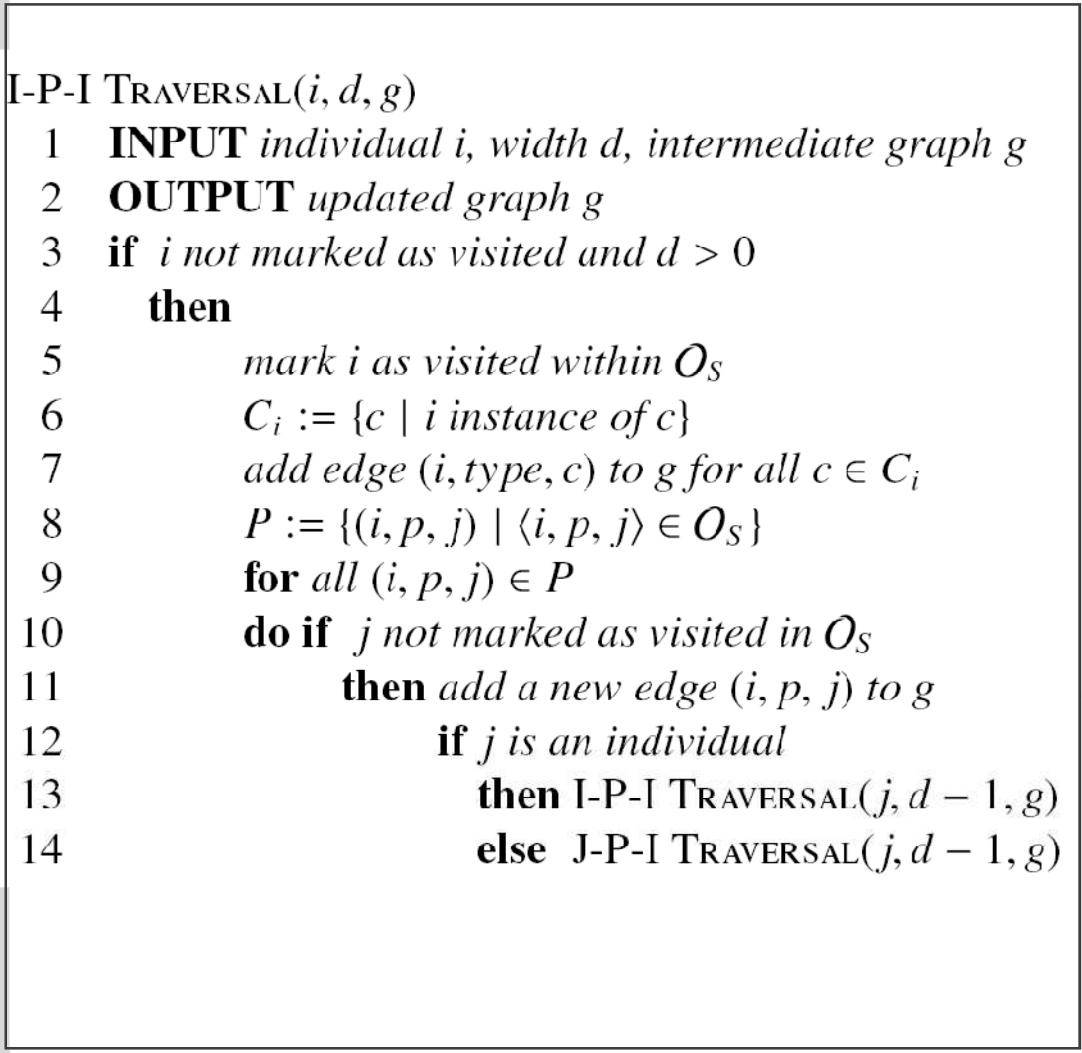


```
KB EXPLORATION($O'_S$, $d$)
INPUT entities $O'_S$ and traversal width $d$
OUTPUT graph containing all or some of $O'_S$
Intitialize new empty graph $g$
for $e \in O'_S$
do if $e$ is a concept
    then for all $i$ being instances of $e$
        do I-P-I TRAVERSAL($e, d, g$)
    else if $e$ is an object property
        then for all $i, j$ with $\langle i, e, j \rangle \in O_S$
            do I-P-I TRAVERSAL($i, d, g$)
               I-P-I TRAVERSAL($j, d, g$)
    else if $e$ is a data property
        then for all $i, j$ with $\langle i, e, j \rangle \in O_S$
            do J-P-I TRAVERSAL($j, d, g$)
    else if $e$ is an individual
        then I-P-I TRAVERSAL($e, d, g$)
    else if $e$ is a data value
        then J-P-I TRAVERSAL($e, d, g$)
return $g$
```

- Algorithms for
  - **Knowledge Base Exploration**
  - Recursive traversal of elements
  - Adopted Depth First Search (DFS) for calculation of paths

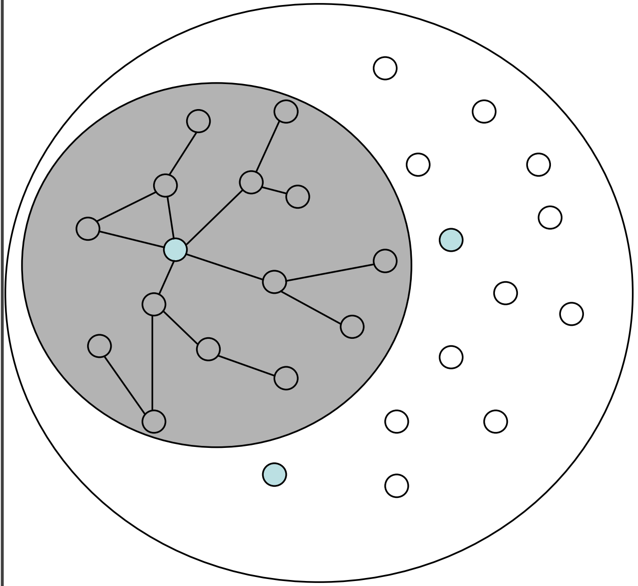


Step 2 – Exploring Connections among Ontology Entities

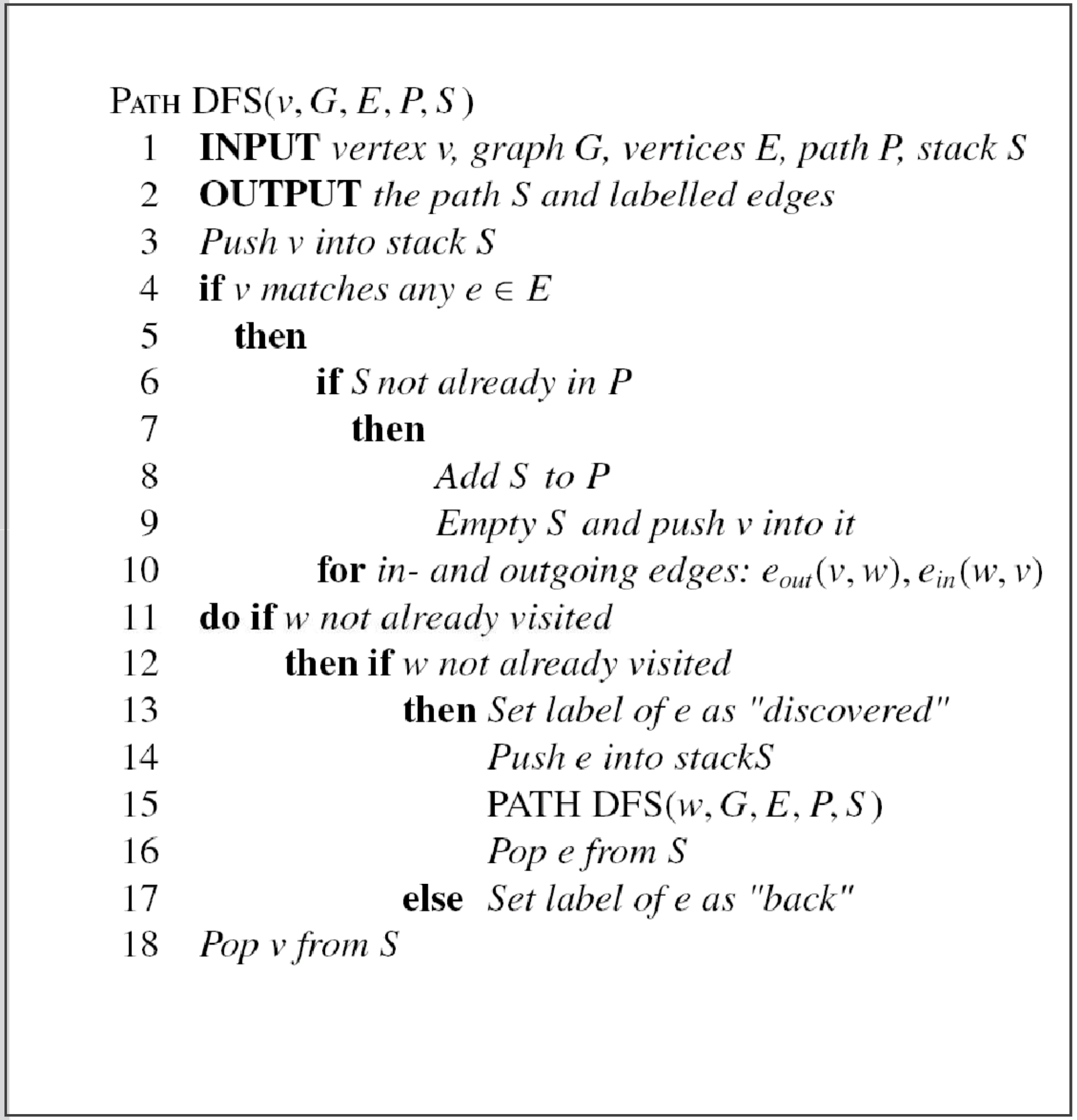


```
I-P-I TRAVERSAL$(i, d, g)$
  INPUT individual $i$, width $d$, intermediate graph $g$
  OUTPUT updated graph $g$
  if $i$ not marked as visited and $d > 0$
    then
        mark $i$ as visited within $O_S$
        $C_i := \{c \mid i \text{ instance of } c\}$
        add edge $(i, type, c)$ to $g$ for all $c \in C_i$
        $P := \{(i, p, j) \mid \langle i, p, j \rangle \in O_S\}$
        for all $(i, p, j) \in P$
        do if $j$ not marked as visited in $O_S$
            then add a new edge $(i, p, j)$ to $g$
                if $j$ is an individual
                  then I-P-I TRAVERSAL$(j, d - 1, g)$
                  else J-P-I TRAVERSAL$(j, d - 1, g)$
```

- Algorithms for
  - KB Exploration
  - **Recursive traversal of elements**
  - Adopted Depth First Search (DFS) for calculation of paths

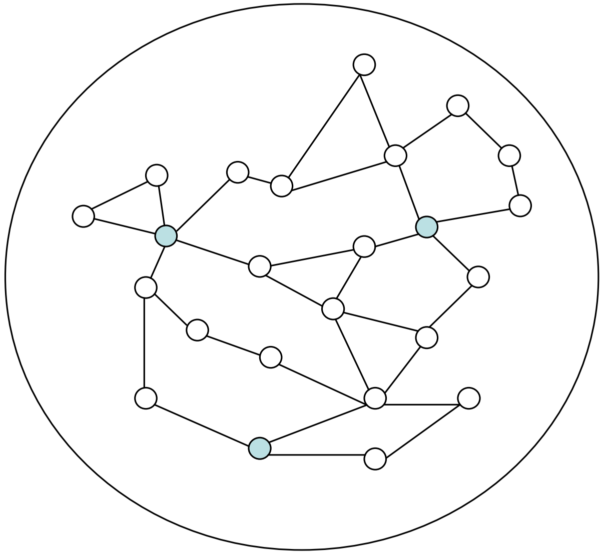


Step 2 – Exploring Connections among Ontology Entities

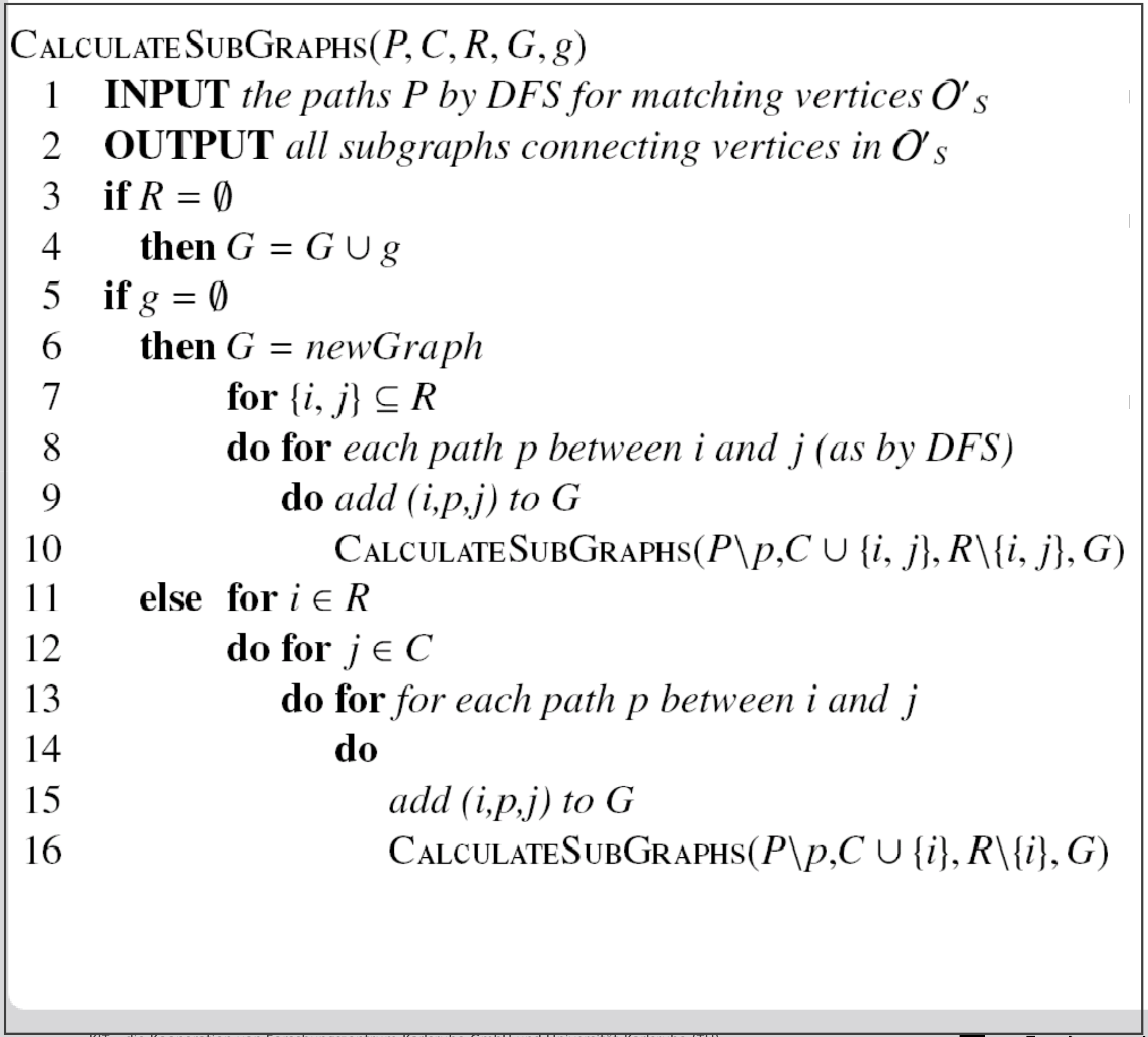


```
PATH DFS($v, G, E, P, S$)
INPUT vertex $v$, graph $G$, vertices $E$, path $P$, stack $S$
OUTPUT the path $S$ and labelled edges
Push $v$ into stack $S$
if $v$ matches any $e \in E$
   then
         if $S$ not already in $P$
            then
                  Add $S$ to $P$
                  Empty $S$ and push $v$ into it
         for in- and outgoing edges: $e_{out}(v, w), e_{in}(w, v)$
do if $w$ not already visited
      then if $w$ not already visited
               then Set label of $e$ as "discovered"
                    Push $e$ into stack$S$
                    PATH DFS($w, G, E, P, S$)
                    Pop $e$ from $S$
               else Set label of $e$ as "back"
Pop $v$ from $S$
```

- Algorithms for
  - KB Exploration
  - Recursive traversal of elements
  - **Adopted Depth First Search for calculation of paths**



Step 3 – Deriving Conjunctive Queries from Connections

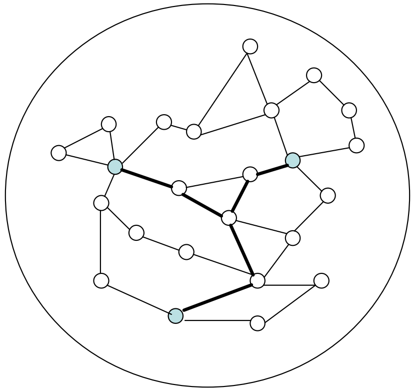


```
CALCULATESUBGRAPHS($P, C, R, G, g$)
INPUT the paths P by DFS for matching vertices $O's$
OUTPUT all subgraphs connecting vertices in $O's$
if $R = \emptyset$
  then $G = G \cup g$
if $g = \emptyset$
  then $G = newGraph$
       for $\{i, j\} \subseteq R$
       do for each path p between i and j (as by DFS)
          do add (i,p,j) to G
             CALCULATESUBGRAPHS($P \setminus p, C \cup \{i, j\}, R \setminus \{i, j\}, G$)
  else for $i \in R$
       do for $j \in C$
          do for for each path p between i and j
             do
                add (i,p,j) to G
                CALCULATESUBGRAPHS($P \setminus p, C \cup \{i\}, R \setminus \{i\}, G$)
```

Compute possible subgraphs

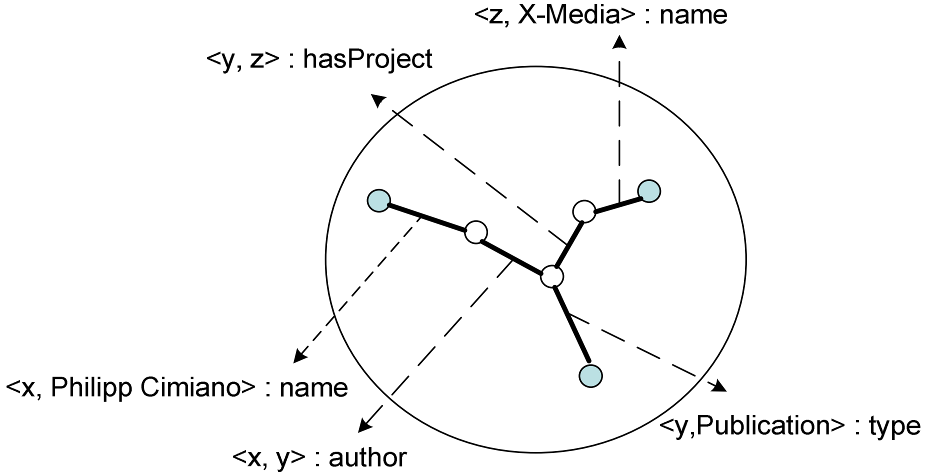
Mapping Connections to Queries

Rank Queries



Step 3 – Deriving Conjunctive Queries from Connections

- Compute possible subgraphs
- **Mapping Connections to Queries**
  - concept member connections and property member connections map to corresponding expressions
  - Vertices matching query elements become constants otherwise variables
- Ranking Queries
  - the smaller the length of the path, the more likely is the corresponding interpretation (locality assumption)
  - length of the longest path of the connection graph

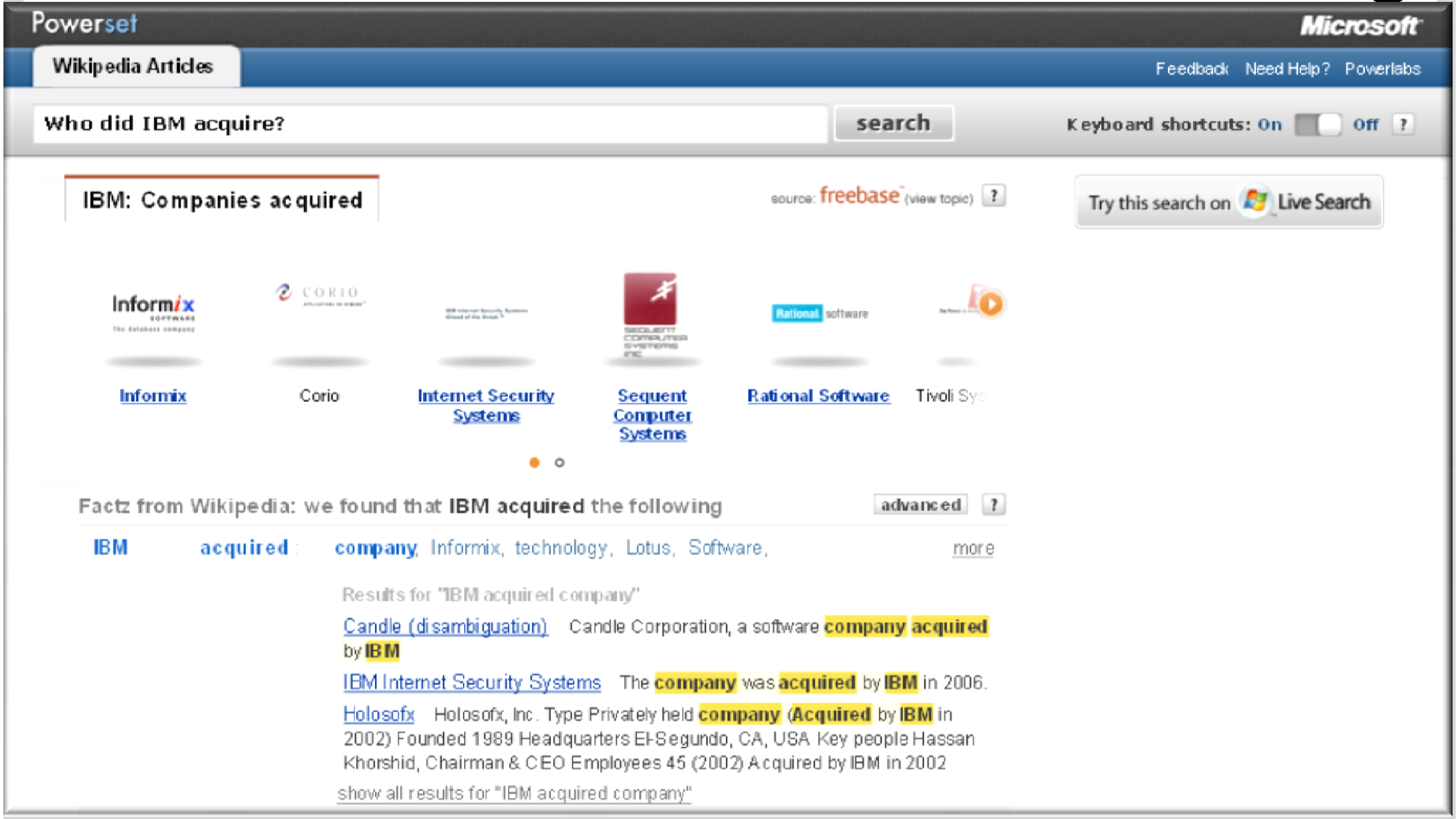


Topics

- **Semantic Search**
  - Overview
  - Examples of Semantic Search Systems
  - Details of the Semantic Search Process
  - Ontology-based Interpretation of Keyword Queries
  - **Natural Language Interfaces**

- Information Integration
  - Ontology Mapping
  - Automated Mapping Discovery

Powerset



Who did IBM acquire?

IBM: Companies acquired

source: freebase (view topic)

Informix · Corio · Internet Security Systems · Sequent Computer Systems · Rational Software · Tivoli Sy

Factz from Wikipedia: we found that **IBM acquired** the following

IBM acquired : company, Informix, technology, Lotus, Software, more

Results for "IBM acquired company"

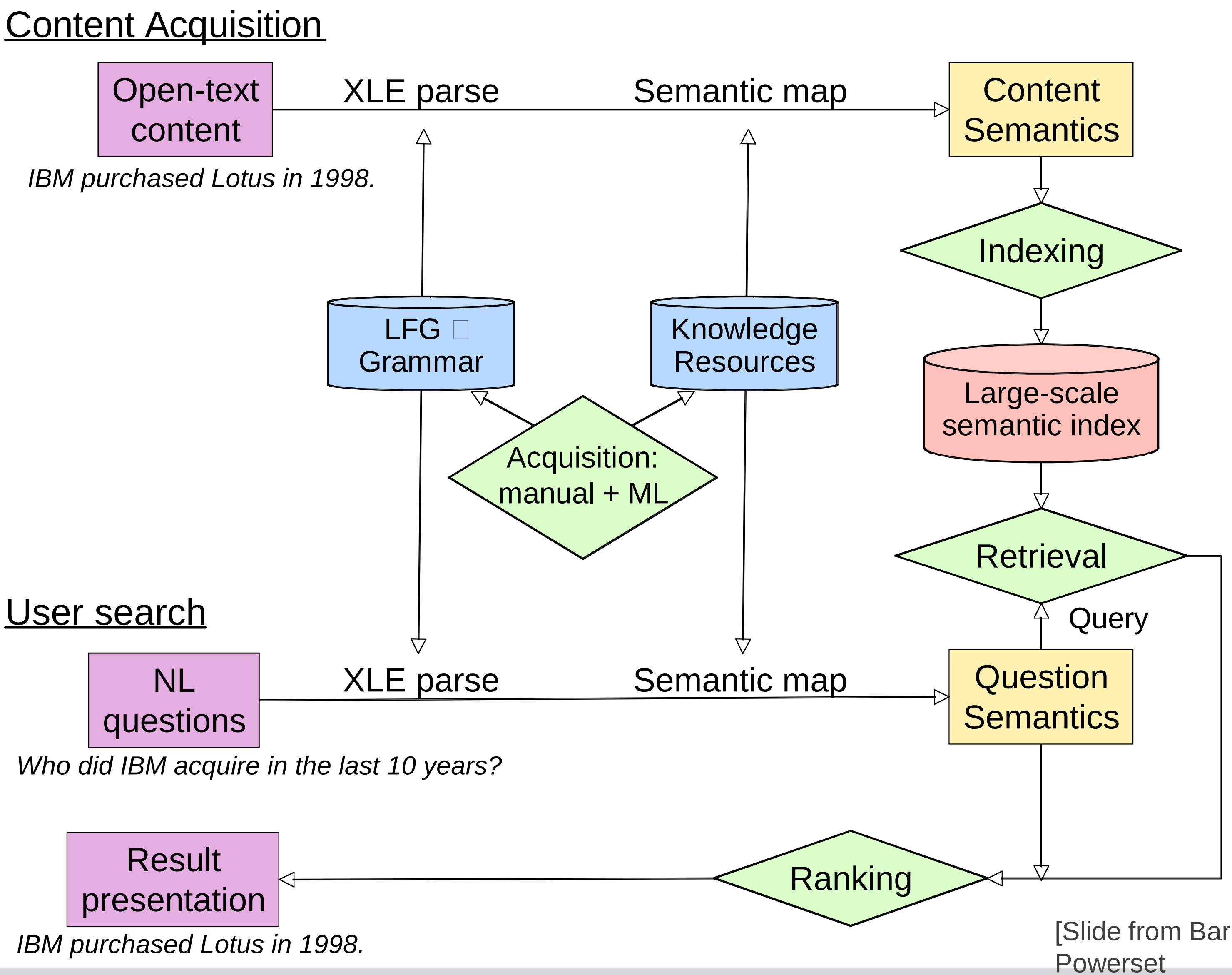
Candle (disambiguation) Candle Corporation, a software **company acquired** by **IBM**

IBM Internet Security Systems The **company** was **acquired** by **IBM** in 2006.

Holosofx Holosofx, Inc. Type Privately held **company** (**Acquired** by **IBM** in 2002) Founded 1989 Headquarters El-Segundo, CA, USA Key people Hassan Khorshid, Chairman & CEO Employees 45 (2002) Acquired by IBM in 2002

show all results for "IBM acquired company"

Powerset: Natural Language Search Architecture



Content Acquisition

Open-text content

XLE parse

Semantic map

Content Semantics

IBM purchased Lotus in 1998.

Indexing

LFG □ Grammar

Knowledge Resources

Large-scale semantic index

Acquisition: manual + ML

Retrieval

User search

Query

NL questions

XLE parse

Semantic map

Question Semantics

Who did IBM acquire in the last 10 years?

Result presentation

Ranking

IBM purchased Lotus in 1998.

[Slide from Barney Pell], Powerset

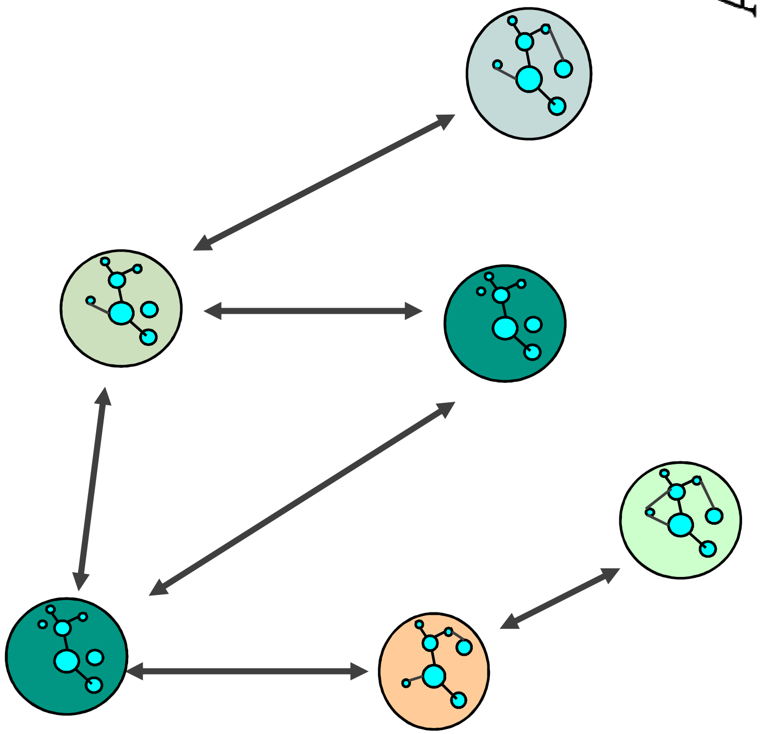
Topics

- Semantic Search
  - Overview
  - Examples of Semantic Search Systems
  - Details of the Semantic Search Process
  - Ontology-based Interpretation of Keyword Queries
  - Natural Language Interfaces
- **Information Integration**
  - Ontology Mapping
  - Automated Mapping Discovery

Ontology Mapping – Problem and Scope

- The Problem
  - ***Heterogeneous ontologies require mappings for interoperability***
  - Numerous independent Ontologies
  - No single Domain Model
  - Modeling same or overlapping Knowledge

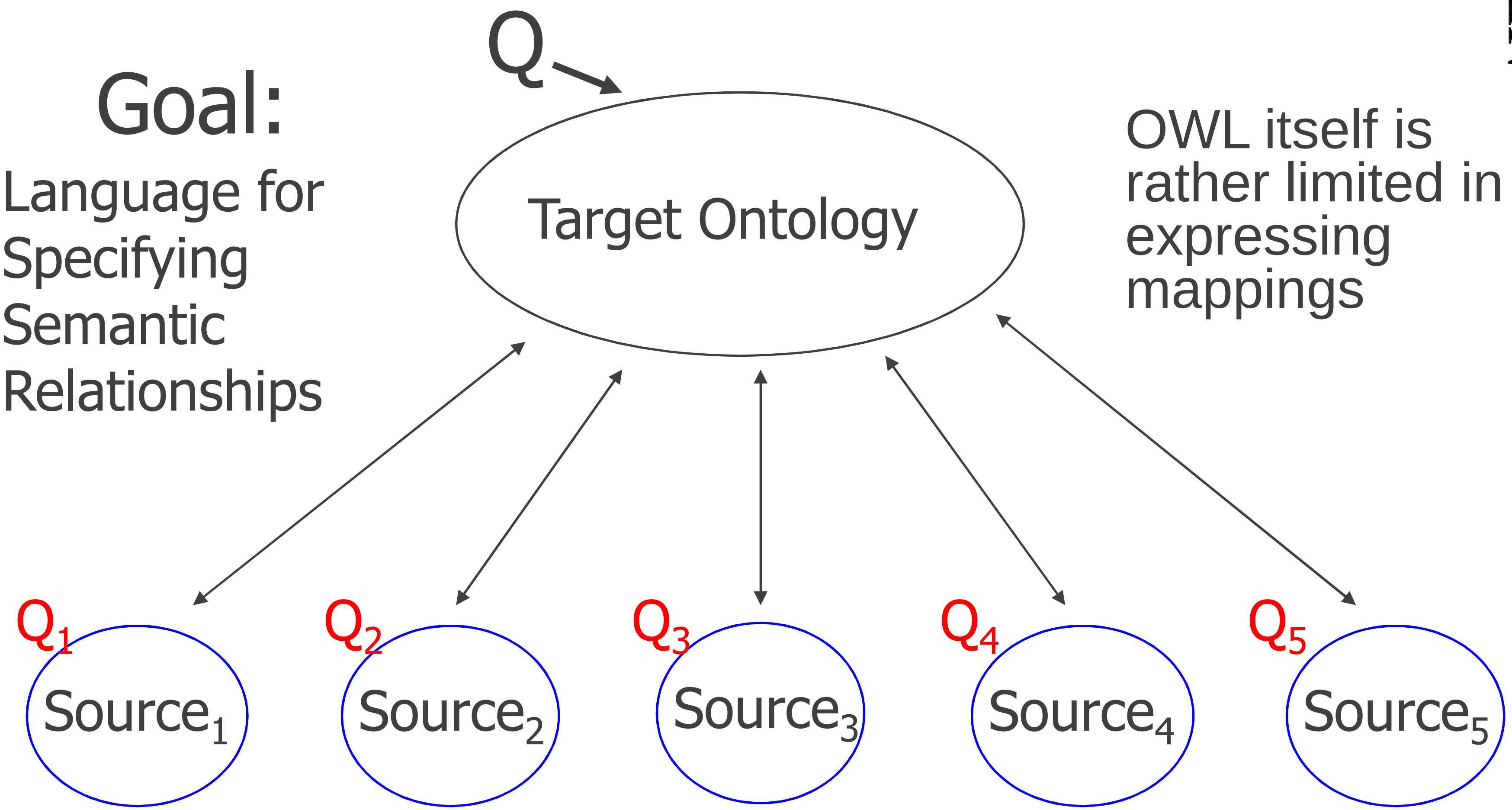
- Main challenges
  - Identifying mappings (correspondences between Entities)
  - Representing these Relations
  - Utilizing Mapping for querying, reasoning, ontology integration, translation and exchange



Mapping Systems for Ontology Integration

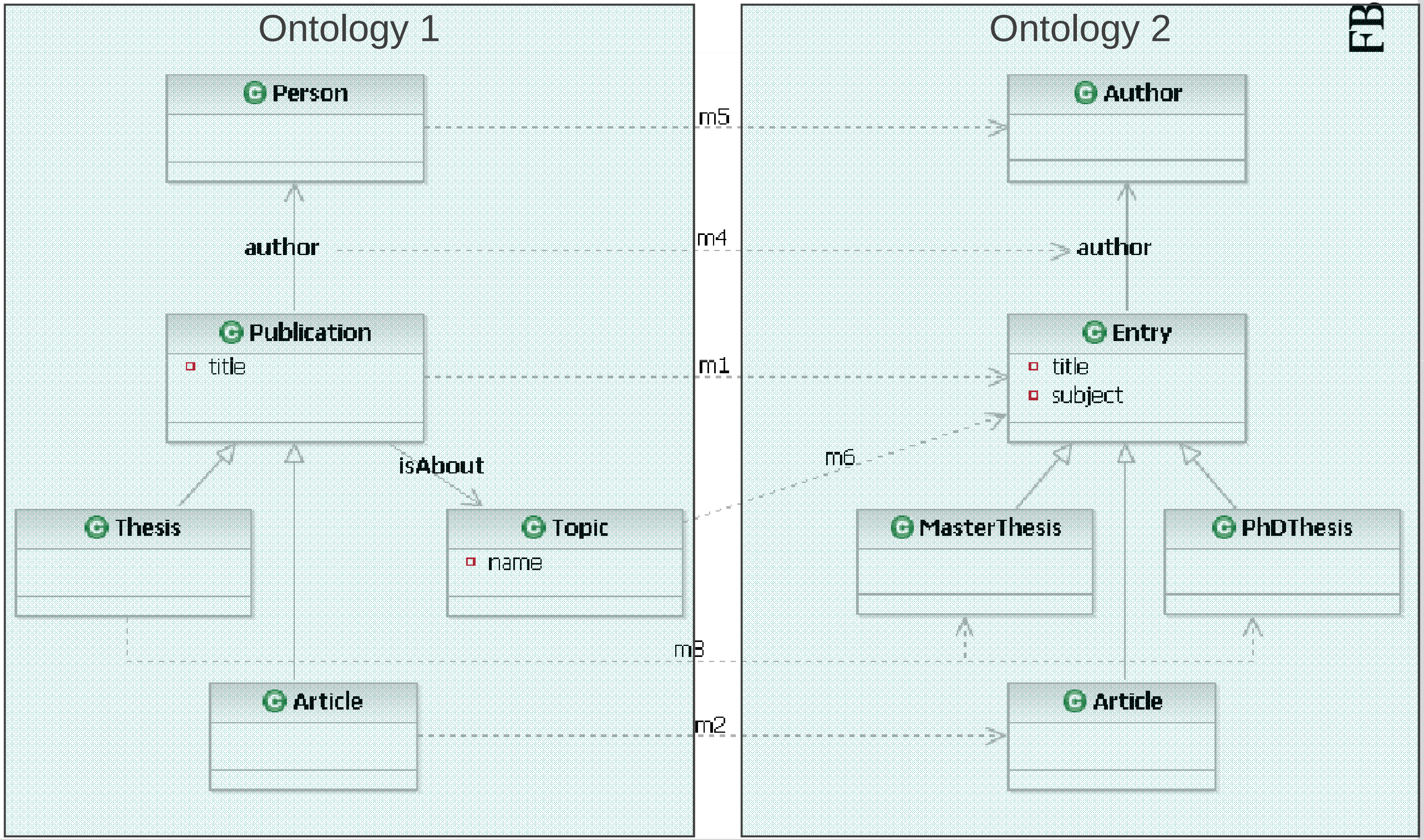
Goal:

Language for Specifying Semantic Relationships



OWL itself is rather limited in expressing mappings

Sample Mapping



OWL DL Mapping System

- An OWL DL mapping system is a triple ($\mathcal{S}$, $\mathcal{T}$, $\mathcal{M}$), where
  - $\mathcal{S}$ is the source OWL DL ontology
  - $\mathcal{T}$ is the target OWL DL ontology
  - $\mathcal{M}$ is the mapping between $\mathcal{S}$ and $\mathcal{T}$

- Mapping: set of assertions
  - $q_S \sqsubseteq q_T$ (sound mapping)
  - $q_S \sqsupseteq q_T$ (complete mapping)
  - $q_S \equiv q_T$ (exact mapping)
  - where q_S and q_T are conjunctive queries over $\mathcal{S}$ and $\mathcal{T}$, respectively, with the same set of distinguished variables

- Semantics defined via translation into FOL, computing answers against $\mathcal{S} \cup \mathcal{T} \cup \mathcal{M}$

[Haase and Motik, IHIS05]

Examples

- Correspondences between atomic elements
  - s: Publication(x) $\sqsubseteq$ t:Entry(x)
  - s:author(x,y) $\sqsubseteq$ t:author(x,y)
- Correspondences between complex class descriptions
  - s:Thesis(x) $\sqsubseteq$ t:PhDThesis$\sqcup$t:MasterThesis(x)
- Even more complex mappings
  - s: Publication(x)$\wedge$ isAbout(x,y) $\wedge$ name(y,z) $\sqsubseteq$ t:Entry(x)$\wedge$ subject(x,z)

Ontology Mapping – Techniques and Tools

- Great number of Techniques
  - Syntactic, Semantic, External
  - Element-Level, Structure-Level
  - Schema or Instance Level mapping

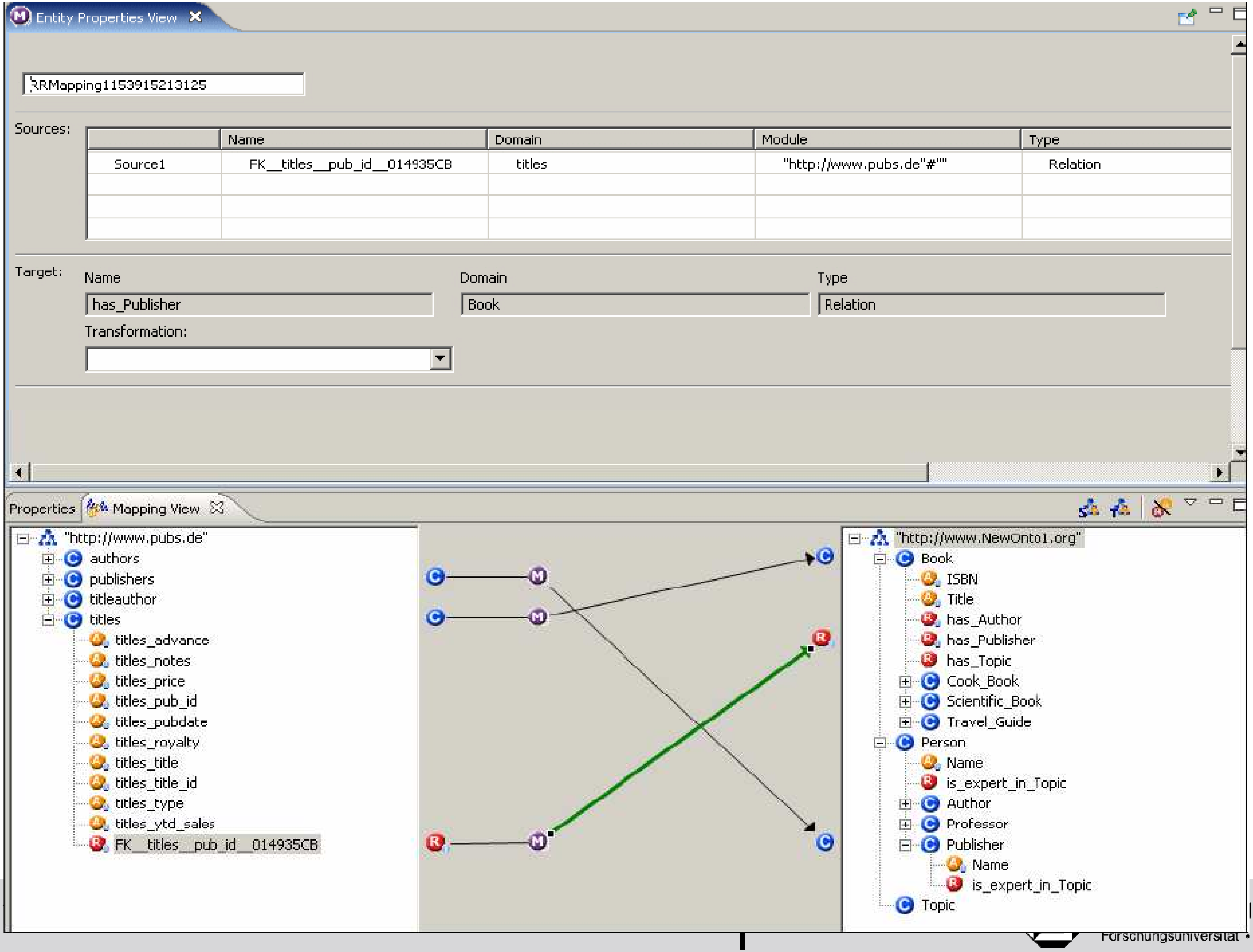
- Mapping Tools
  - Several mapping systems already available

    (GLUE, PROMPT, FOAM, ONION, MAFRA)
  - Manual, visual creation of mappings between ontologies
  - Integration of (relational databases): automated ontology lifting and query answering

    (OntoMap, ODEMapster)

- Best results
  - Find best approximate Matches -> Similarity
  - Semi–automatic
  - Requires human Domain Expert

Ontology Mapping with OntoMap



Entity Properties View

RRMapping1153915213125

Sources:

	Name	Domain	Module	Type
Source1	FK__titles__pub_id__014935CB	titles	"http://www.pubs.de"#""	Relation

Target:

Name: has_Publisher

Domain: Book

Type: Relation

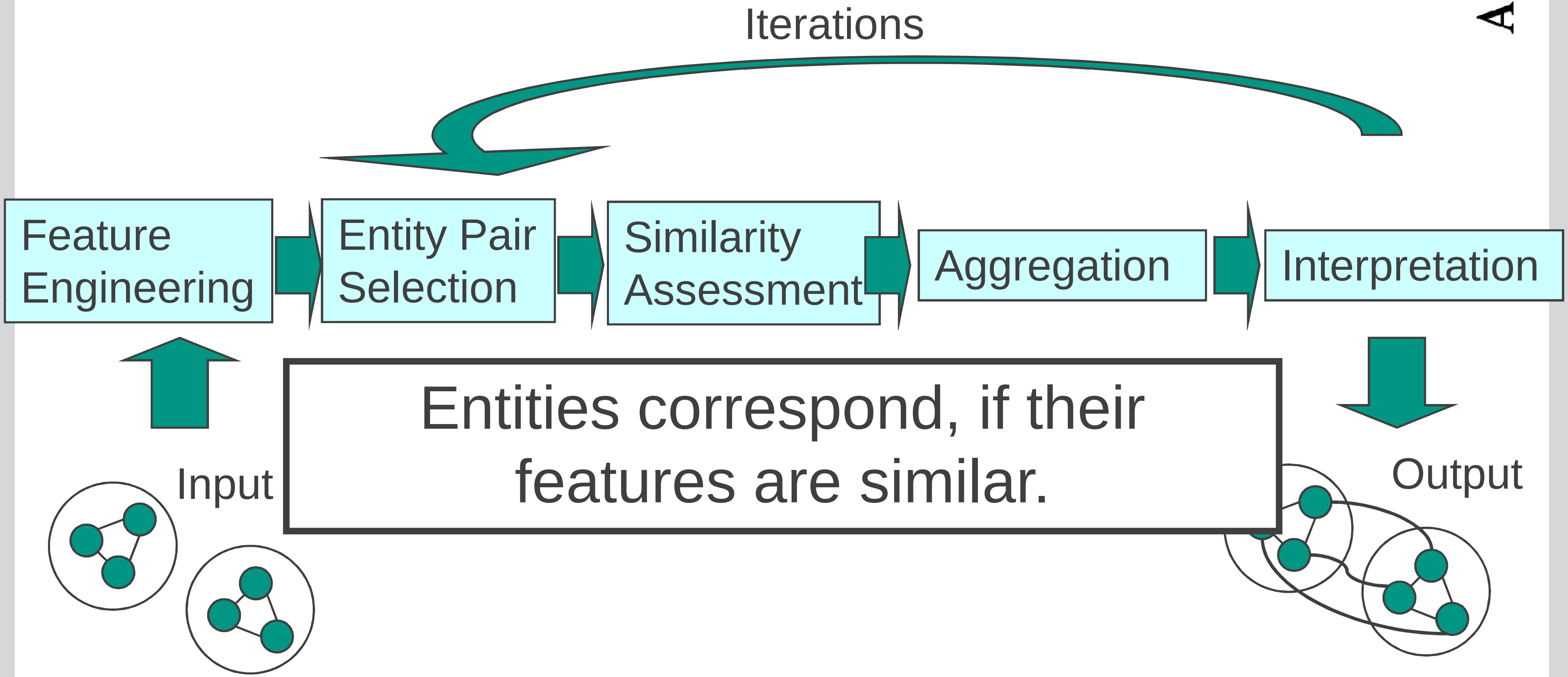
Transformation:

Topics

- Semantic Search
  - Overview
  - Examples of Semantic Search Systems
  - Details of the Semantic Search Process
  - Ontology-based Interpretation of Keyword Queries
  - Natural Language Interfaces

- **Information Integration**
  - Ontology Mapping
  - Automated Mapping Discovery

Automated Mapping Discovery Process



Entities correspond, if their features are similar.

Features and Similarity Measures

	Feature	*Similarity Measure*
Concepts	label	String Similarity
	subclassOf / superclassOf	Set Similarity
	instances	Set Similarity
	…	
Relations	Domain, Range	Set Similarity
	…	
Instances		

Similarity Measures

- String similarity

$$sim_{String}(s_1, s_2) = \max(0, \frac{\min(|s_1|, |s_2|) - ed(s_1, s_2)}{\min(|s_1|, |s_2|)})$$

- Set similarity

$$sim_{Set}(S_1, S_2) = \underset{e_i \in S_i}{avg} \underset{e_j \in S_j}{\max}(sim(e_i, e_j))$$

Aggregation of multiple similarity measures

- Weighted combination method
  - Manually
  - Machine learning

$$sim(e,f)=\sum_k w_k sim_k(e,f)$$

- Non-weighted combination method
  - Average
  - Maximal
  - Minimal

Summary

- Semantic Search: Search on the level of meaning
  - Interpretation of the user‘s information need

- Ontology Mapping: Integration of heterogeneous information sources