Karlsruhe Institute of Technology

AIFB

Exploring the knowledge in Semi Structured Data Sets with Rich Queries

Jürgen Umbrich
Sebastian Blohm

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Overview

- How to use a annotation engine to extract implicit knowledge encoded in semi structured data sets.
- How to discover, in a automatical way, relation patterns between concepts/categories.
- A framework with support for free text search combined with annotation search
- A user-interface, that hides the complexity of a structured query syntax from the end-user

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Motivation: (Semi-) Structured KBs

- Effort to manage unstructured information in (semi-) structured knowledge bases
  - Encyclopedias, like Wikipedia
  - ODP
- Information management is often maintained either manually and/or in a supervised manner.
- KB's reflect the “wisdom of the crowd” and cover a lot of different domains

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Challenges

- Current approaches to **access** information in semi structured KB
  - Keyword search interface
  - Exploring the data set by article and category links or facets

  -> Works well only for small and/or specific data sets
- Inability to incooporate **background knowledge** of users
  - Background knowledge about the domain
  - Knowledge about relations in the result pages

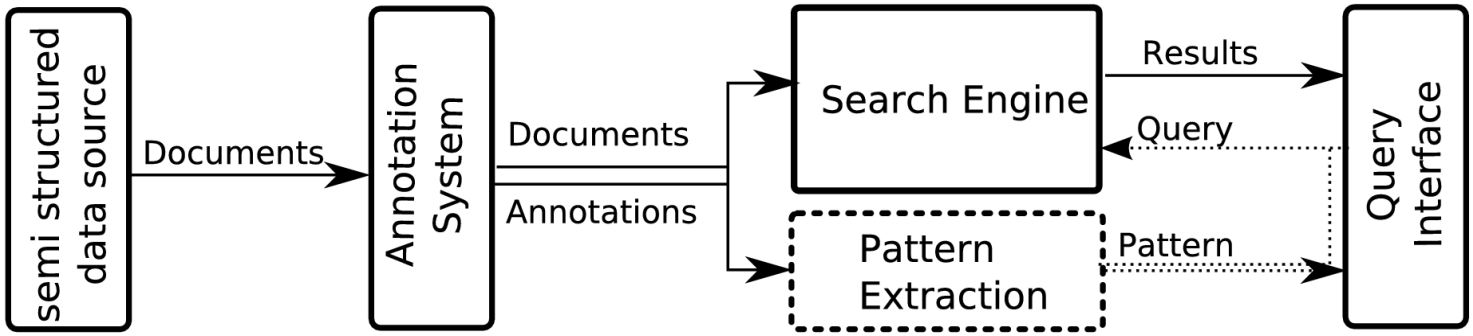
KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Solution Approach:

- Accessing the KB via freetext search combinded with annotation search
  - By extracting implicit knowledge encoded in the KB like **categories and links** between articles,
  - By semantical grounding of extracted information with an ontology
- Support the background knowledge of end-users
  - By automatically extracted relation patterns from the knowledge base

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Solution Architecture: Overview

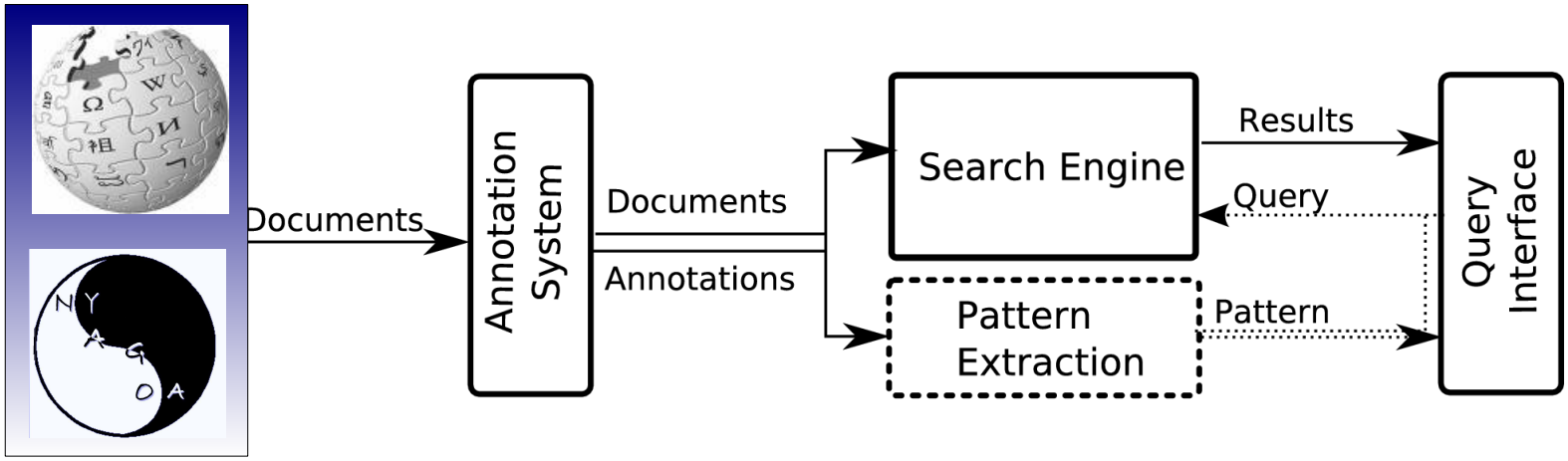


KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

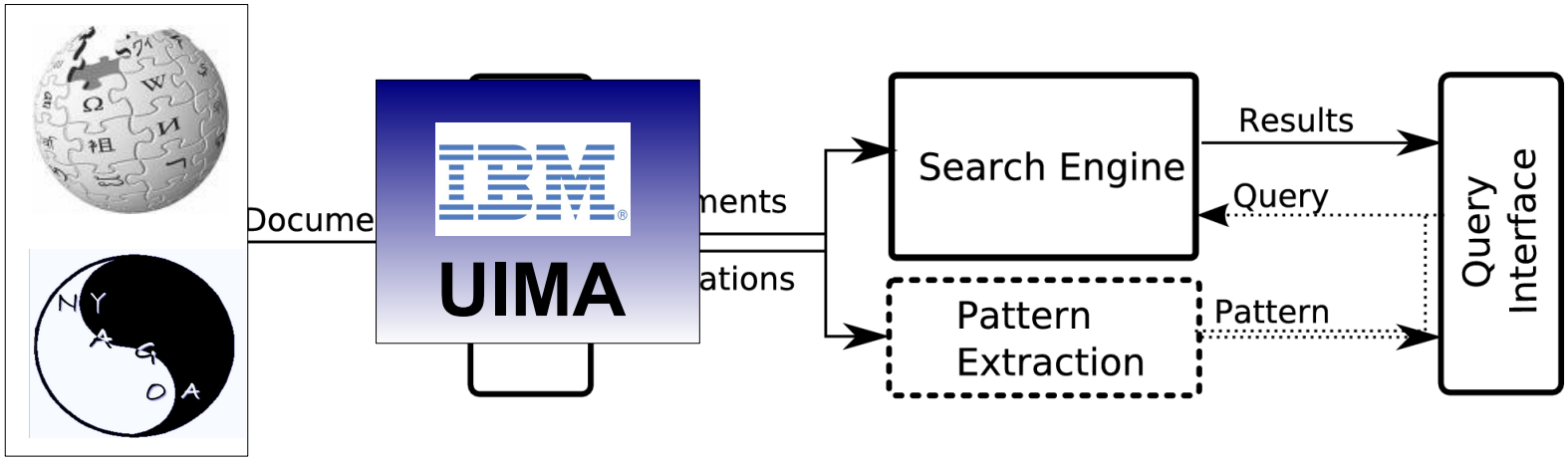
(Semi-)Structured Data Source



- A local **Wikipedia** dump (current from December 2006)
- Wikipedia categories semantically grounded with the **Yago ontology**
  - Mapping between Wikipedia categories and the Yago hierachy

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Annotation System (AS)



- Unstructured Information Managemen Architecture (UIMA)
- Open source Java framework (Sun Java version)
- Various text analysis engines (TAE) build from scratch
  - Sentence and paragraph splitters, tokenizer ,etc ...
- Easy developement of own TAE (in Java)
  - e.g. a simple data annotator or Wikipedia to Yago category mapping

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

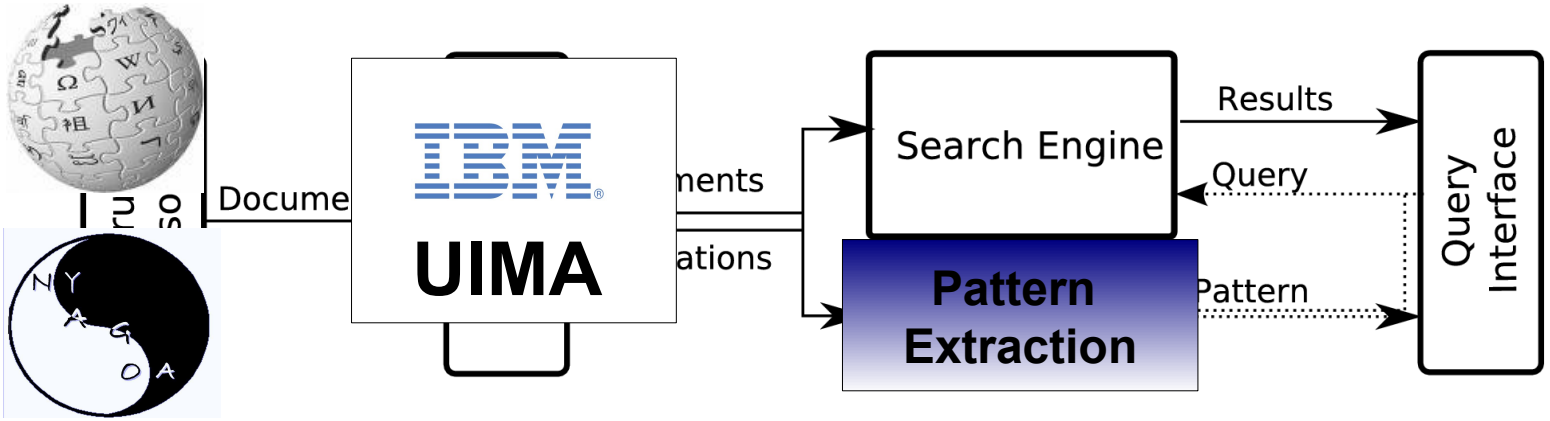
(AS) Article Annotations

Wikipedia article of "Karl Steinbuch"

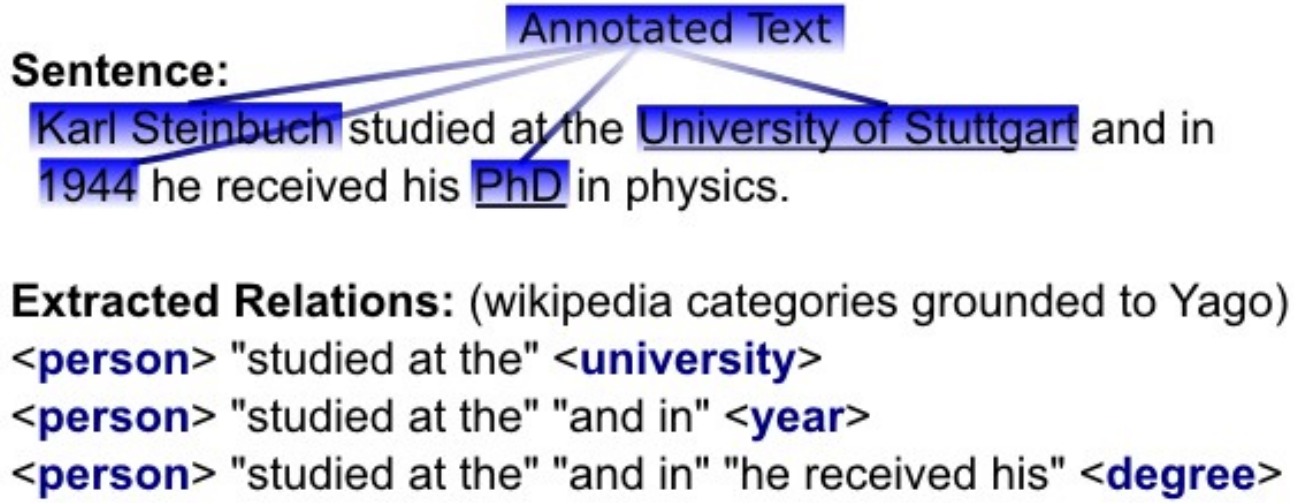
- Semantically enriched annotations
  - Utilising the Yago ontology
  - Each wikipedia category is modelled as a hierachy of yago categories e.g. German_computer_scientist ->... -> scientist ->... -> person

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Annotation System (AS)



Relation Pattern Extraction
Semi - semantic relation patterns



- Extracting relation between wikipedia categories and their respective Yago categories
  - In sentences and paragraphs
- Relation patterns support query creation in the UI

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Crawling - Indexing - Searching

- Integrates UIMA (Version 1.4 , current published Version is 2.x)
- Query functionality to combine text with annotation search
- XML-Fragment or Xpath Syntax

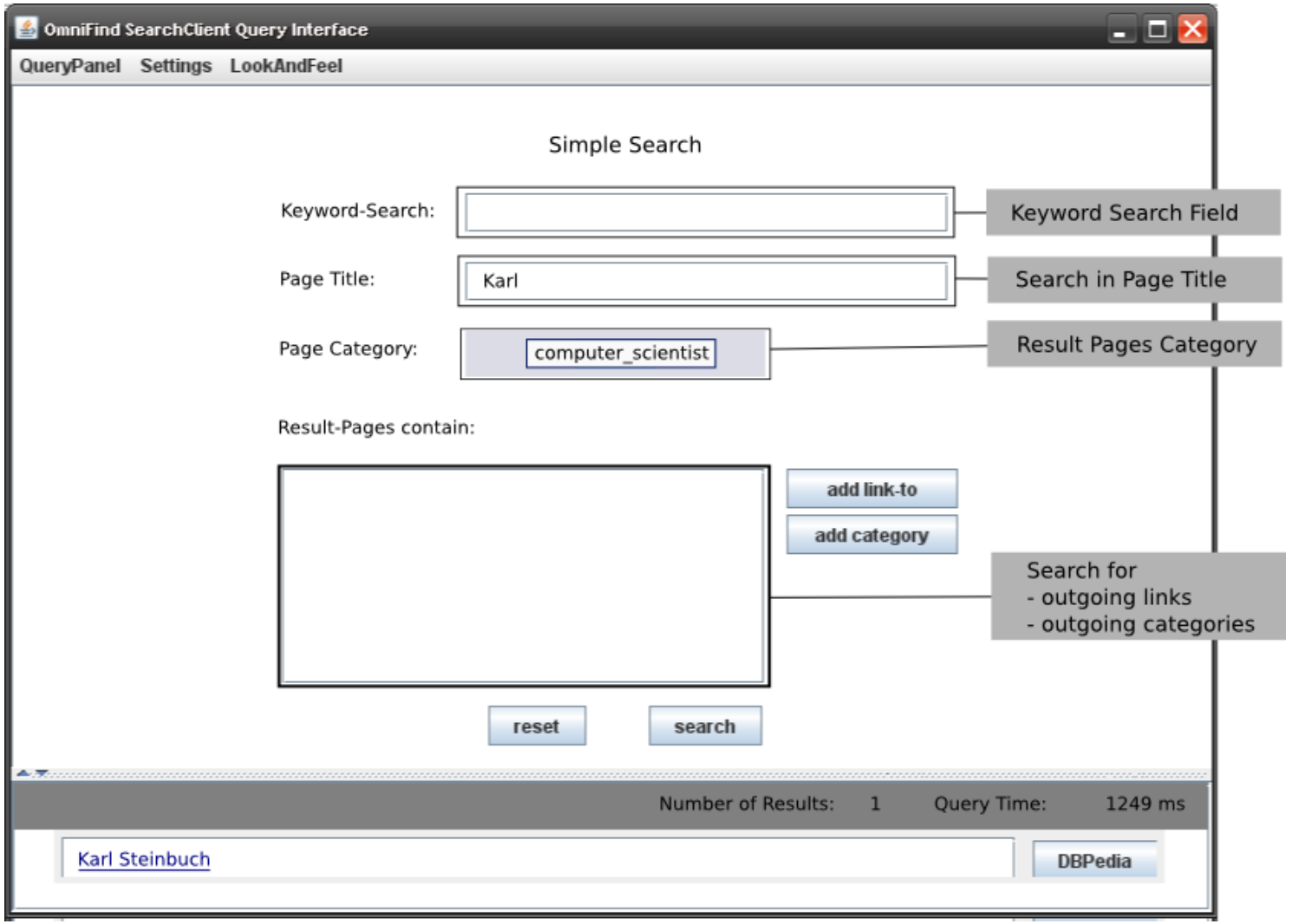
KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

End-User Interface

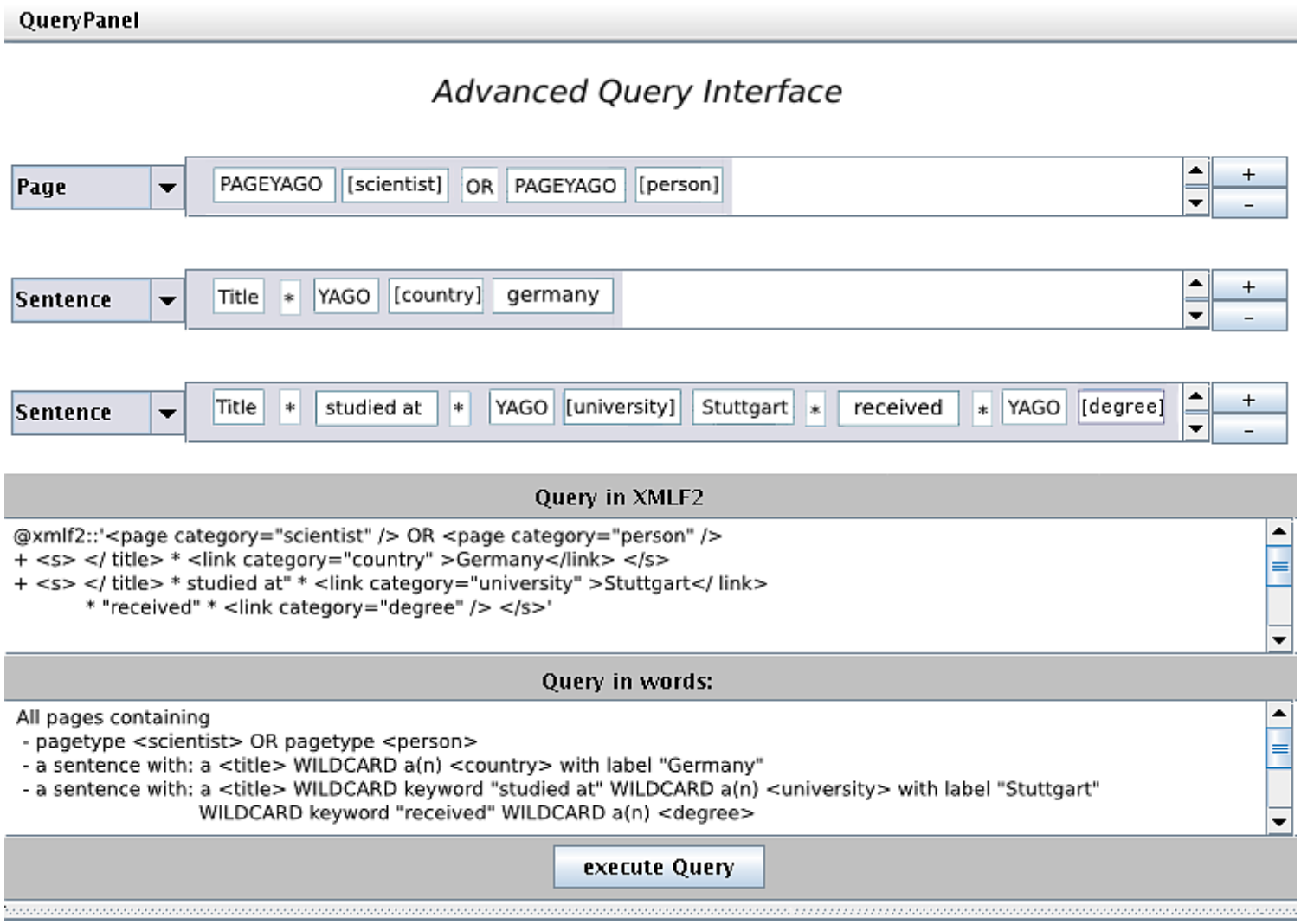
- Hides the compexity of the underlying query syntax
- Two query creation modes (Simple and Advanced)
- Uses know query concepts (e.g. Textfield and DropDown menu)
- Implementation in Java Swing

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

User Interface: Simpe Mode



User Interface: Advanced Mode



KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Conclusion

- Automatical extraction of (semi) semantic relation patterns from semi structured KB
  - Using text information extraction tools
- Explore and query semi structured KB's with a combination of freetext and annotation search
  - Document or entity centric search
  - Instance and/or concept search

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Future Work

- Evaluation of the architecture
  - Annotation time (current measures: ~3sec/doc)
  - Indexing time/size
  - Query time related to query complexity
- Automatical discovery of semantic relation patterns (SRP)
  - e.g. *[concept:person] [relation:studiedAt] [concept:university]*
  - Automatically build a concept-relation ontology based on the Wikipedia corpus
- Extending UI interoperability capabilities
  - Support the UI with semantic relation patterns

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Exploring the knowledge in Semi Structured Data Sets with Rich Queries

Karlsruhe Institute of Technology

AIFB

Questions?

- **Contact**
  juergen@umbrich.net
  blohm@aifb.uni-karlsruhe.de

Work was supported by

X-MEDIA

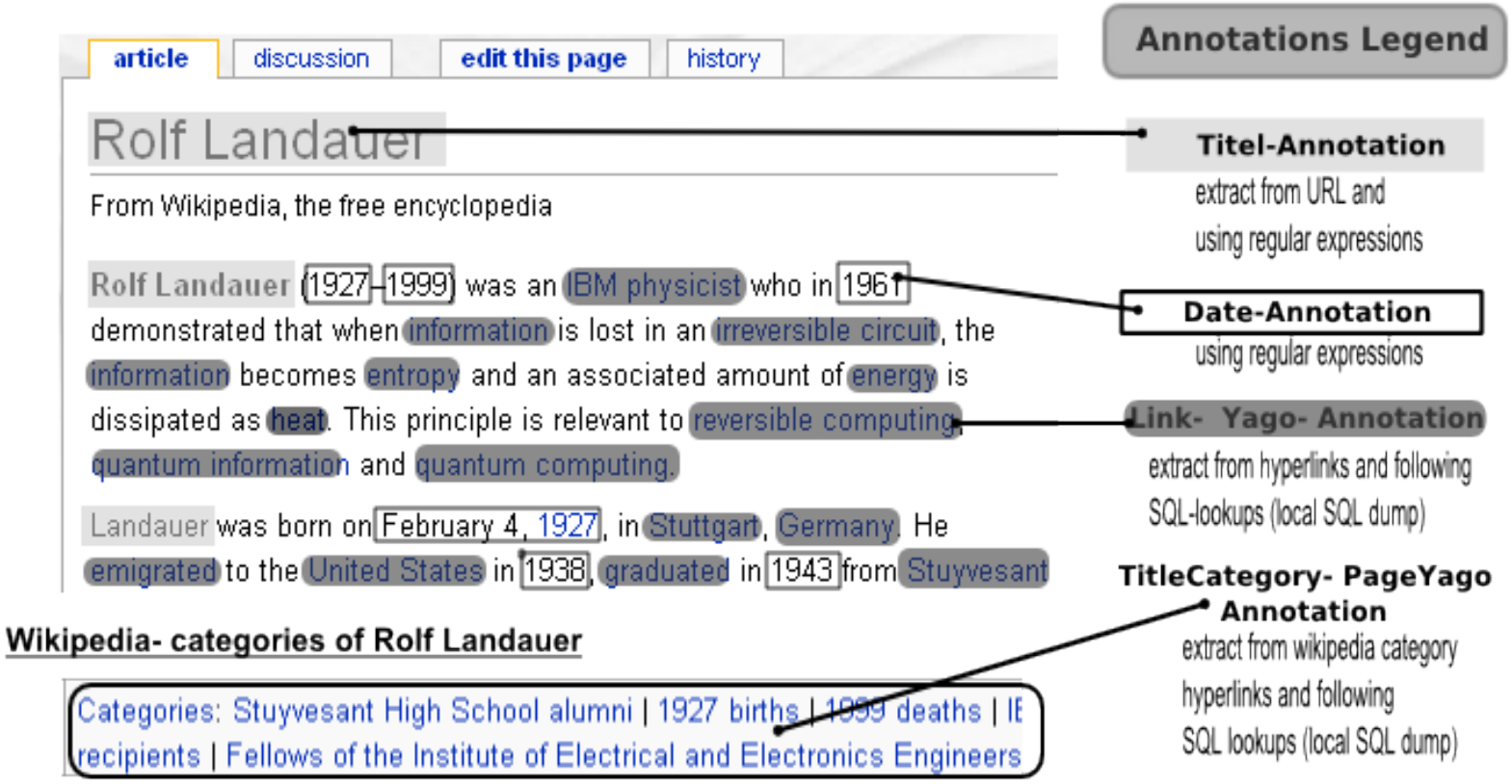
MFG Stiftung Baden-Württemberg

Article Annotations

- Annotate Wikipedia articles by
  - Title and title occurences in the article
  - Article categories
  - Links, link title and link categories
  - Date information from various date formats
- Annotation semantically grounded with the Yago ontology
  - Each wikipedia category is modelled as a hierachy of yago categories
    e.g. american_tennis_player -> player -> person -> causal_agent

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)
Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft
Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Sample: Annotated Wikipedia article



KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

OmniFind: Query Syntax

- XML-Fragment or Xpath Syntax
- Access to the KB: API or User-Interface
- Sample Query

```
@xmlf2::'
  <page category="scientist" /> OR <page category="person" />
  + <sentence>
        </title> * "born in" * <link category="country">Germany</link>
     </sentence>
  + <sentence>
        </title> * "studied at" * <link category="University"/>
        * <link category="city">Karlsruhe</link>
        * "received" * <link category="Degree"/>
  </sentence>
'
```


KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Table of Content

- Motivation
- Architecture
- Annotation Engine
- User Interface
- Future Work

KIT – die Kooperation von Forschungszentrum Karlsruhe GmbH und Universität Karlsruhe (TH)

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825