Karlsruhe Institute of Technology

DARIAH-DE
Digital Research Infrastructure for the Arts and Humanities

Germaine Götzelmann
germaine.goetzelmann@kit.edu
Institute for Data Processing and Electronics

SPOTLIGHT – Automatic annotation of text with Linked Open Data resources for Humanities

Motivation

- **Linked Open Data cloud** as knowledge source
- Machine-readable **semantic** data
- Controlled vocabularies

➡ Usage for **automatic annotation**

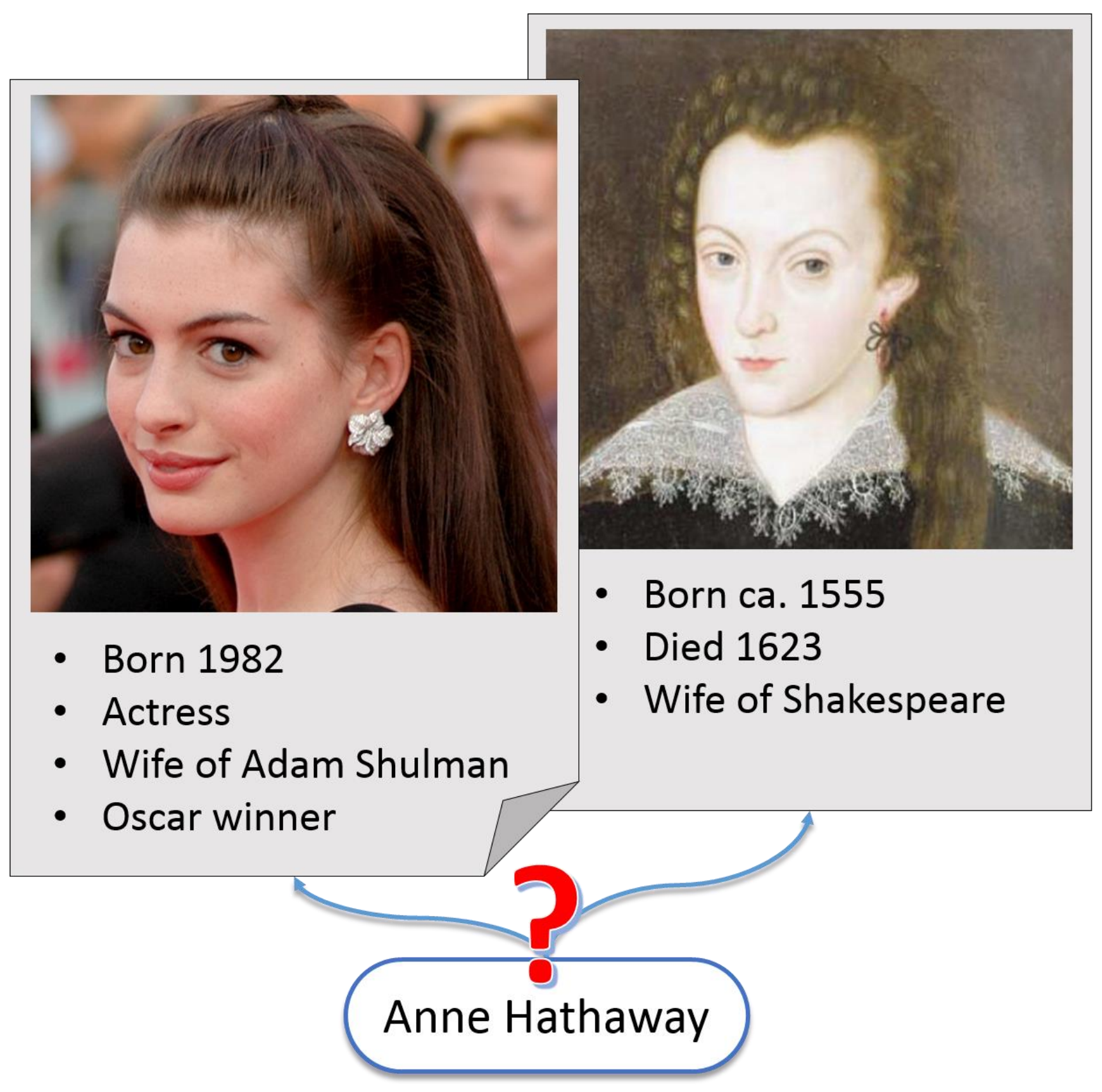
Specific task

Annotation of natural language texts with person data from GND (Gemeinsame Normdatei)

Challenges

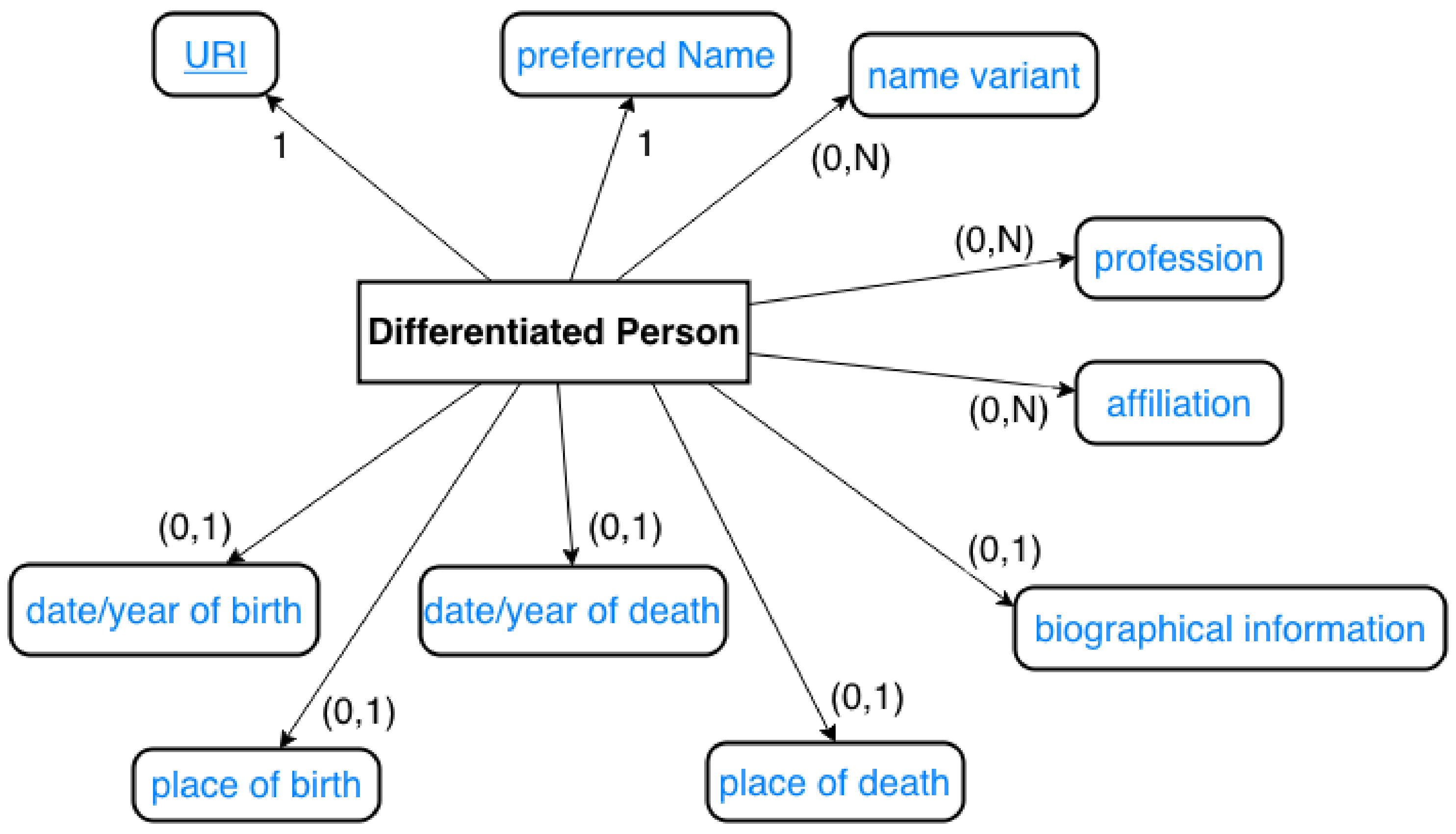
Name variation problem

Name ambiguity problem

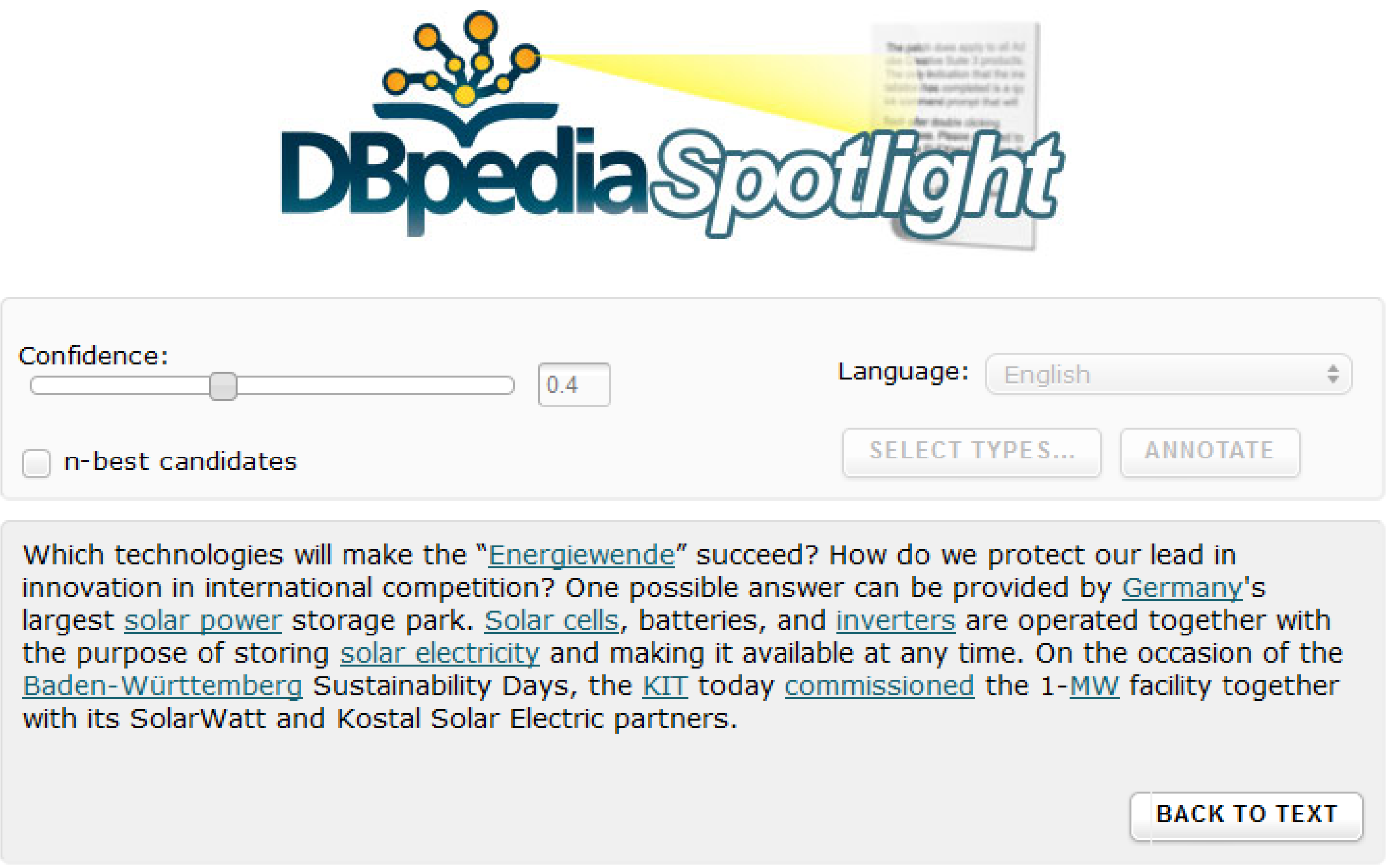


Annotation Technology

- Adaption of *DBpedia Spotlight* with information from Wikipedia/DBpedia
- Extension for other knowledge bases needed
- *Spotlight* requires three basic types of information about every entity:
  - **URI**: Linked Data (LD) identifier for linking
  - **Surface form**: representation of the entity found in text
  - **Context:** words occurring in text surrounding the surface form
- The *statistical backend* additionally considers for every entity: significance, likeliness of its names, likeliness of specific context



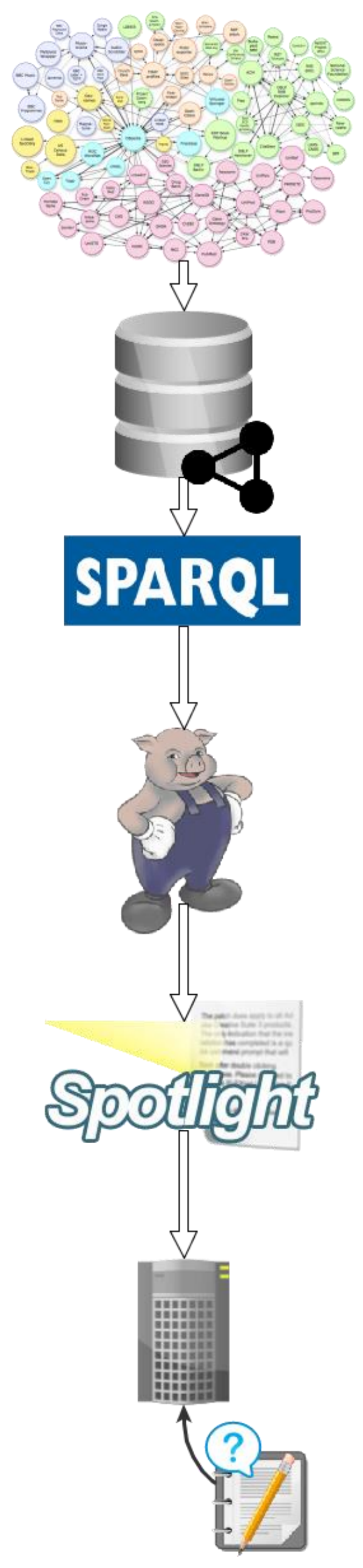
Possible fields with information for entities in GND person data



Which technologies will make the "Energiewende" succeed? How do we protect our lead in innovation in international competition? One possible answer can be provided by Germany's largest solar power storage park. Solar cells, batteries, and inverters are operated together with the purpose of storing solar electricity and making it available at any time. On the occasion of the Baden-Württemberg Sustainability Days, the KIT today commissioned the 1-MW facility together with its SolarWatt and Kostal Solar Electric partners.

Web demo of *DBpedia Spotlight*

Data Acquisition Workflow



Workflow for data creation from Linked Open Data cloud to annotation task

1. Creation of local RDF store
2. Querying of relevant triples with SPARQL
3. Stemming/counting records with *pignlproc*/*Apache Pig*
4. Building of model with modified *DBpedia Spotlight*
5. Deployment on server for text annotation (Rest Interface)

Potential

- Data from the Linked Open Data cloud enhances knowledge in *sparse domains* of research, where information may not pass the relevance criteria of Wikipedia.
- Works hand in hand with *domain experts*, giving *suggestions* for annotations
- Reusable in every scientific field, with knowledge bases published with LD principles