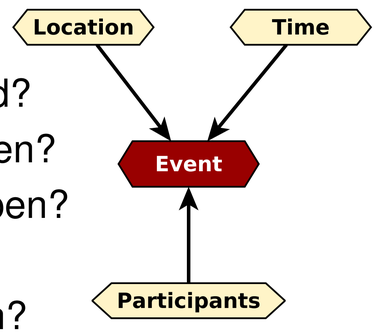
State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities

Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz and Johanna Geiß

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

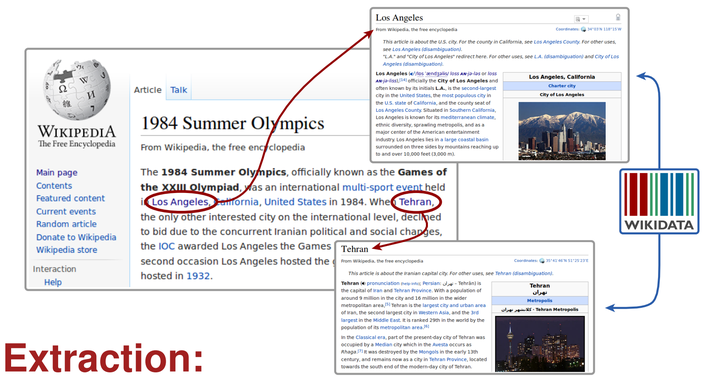
Entities: Components of Events

A pivotal part of **Information Retrieval** from text is the detection of event descriptions. Involved **entities provide the context** of such events [1]. As a result, the identification of events is influenced by the ability to detect and classify entities and link them to a knowledge base. In classic journalism, this is reflected in the well known set of standard questions, the **Five Ws**:



- **Who** was involved?
- **When** did it happen?
- **Where** did it happen?
- What happened?
- Why did it happen?

Entity Extraction and Resolution



Extraction:

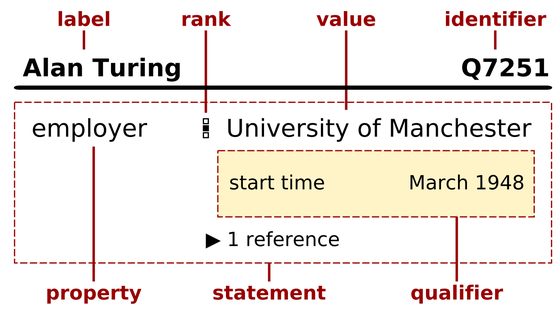
- Identify the entities in the text
- Use gazetteers as support (lists of known entities)

Resolution:

- Classify the entities into groups
- Link them to entries in a knowledge base with additional information

Wikipedia and Wikidata support

Wikipedia texts support the extraction of entities due to the annotation with Wikipedia links. For resolution, Wikipedia also includes **a direct integration of a knowledge base** since Wikidata properties and statements connect entities that correspond to Wikipedia pages [2].



However, there are some **complications in the union of both Wikis...**

Personnel Problems

Fictionality

Not all entities in Wikidata correspond to a real-world analogon. The available classes for persons have large intersections, are partially linked through properties such as *fictional analog (P1074)*, and include among others:

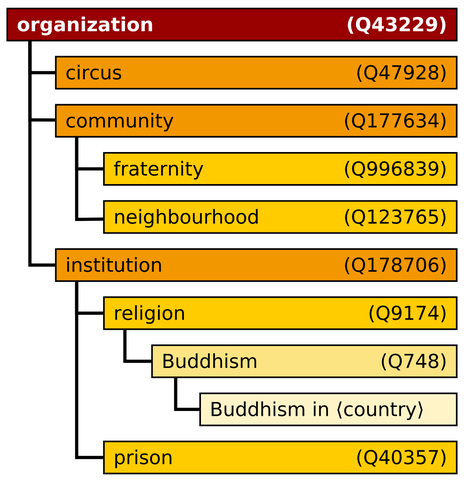
- *human (Q5)*
- *fictional human (Q15632617)*
- *person (Q215627)*
- *fictional character (Q95074)*
- *fictional animal character (Q3542731)*

As a result, identifying persons and building a comprehensive gazetteer of existing persons (real or otherwise) is a difficult and involved process.

Organizational Issues

Levels of organization

Groups of persons in Wikidata are subclasses of *organization (Q43229)*. The corresponding subtree below organization is very large with over 7500 entries. The importance of these subclasses fluctuates severely at all levels of the tree.



Events: Timing is Everything

Discretization of time

Classes like *former entity (Q15893266)* are problematic in a knowledge base since they fix the reference time and are difficult to keep updated.

Locality of events

Many events are directly annotated with geo-coordinates, which makes it difficult to distinguish them from locations. While locations are one aspect of an event, they consist of more than just coordinates.

item label	ID	instance of
Ich bin ein Berliner	Q443	speech
I Have a Dream	Q192341	speech
September 11 attacks	Q10806	terrorist attack
'05 Bali bombings	Q86584	suicide attack
'10 Haiti earthquake	Q43777	earthquake

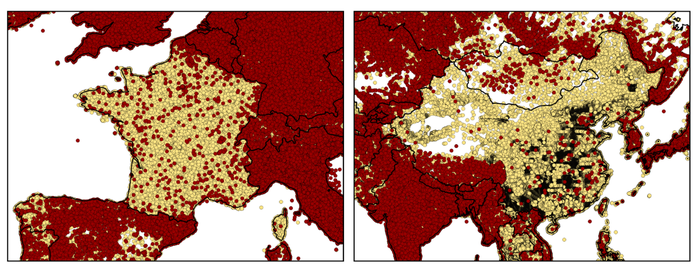
Location, Location, Location

Countries

The distinction between *state (Q7275)* and *country (Q6256)* requires local knowledge, especially since only the former is a subclass of *organization*.

Cities

National municipal hierarchies dilute the list of *human settlements (Q486972)*, such as *commune of France (Q484170)* and *town in China (Q735428)*.



Coverage comparison of populated places in GeoNames (yellow) and human settlements in Wikidata (red).

Approaches and Solutions

Skeleton class hierarchies

The hierarchy of classes in Wikidata is very complex and nuanced. It would be beneficial to have a second, simpler hierarchy (parallel hierarchies are possible).

Legacy properties

Since Wikidata is constantly evolving, its contents change. Users and applications that use Wikidata would benefit from a fixed set of legacy properties.

Fewer discrete properties

Discretized properties such as *large city* or *former entity* hard-code scalar information that should be continuous. They are difficult to maintain, hard to interpret, and ultimately superfluous.

Property constraints

Property constraints in knowledge bases limit possible relations between entities. Wikidata supports constraints on an informal basis already. An inclusion of constraint checking during the data input step would help to ensure adherence of the data to the standards that Wikidata sets for itself.

UIs and tools for data output

Wikidata has many tools for inputting data, yet tools for extracting and using the stored data are more difficult to find. Ideally, retrieving data from the knowledge base should be supported directly and be even easier than adding it.

References

[1] A. Spitz and M. Gertz: **Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events**. *SIGIR'16*, 2016

[2] D. Vrandečić and M. Krötzsch: **Wikidata: A Free Collaborative Knowledgebase** *Communications of the ACM*, 57(10):78–85, 2014

Contact Information:

Andreas Spitz

spitz@informatik.uni-heidelberg.de

http://dbs.ifi.uni-heidelberg.de/

This work was presented at the 2nd Wiki Workshop in conjunction with ICWSM'16, May 17, 2016, Cologne, Germany.