Wikidata as universal thesaurus
Theo van Veen, WikidataCon 2017, 28-10-2017
Overview

• The idea and motivation
• Historical newspapers as motivating use case
• Possible approaches, different environments
• Pros and objections
• Conclusions
Why do we use a thesaurus?

• Unique identification of an entity
• Providing context information
• Enabling search using the entity’s properties as captured in the thesaurus (including name variants)

For what reason do we link to Wikidata?

• We use Wikidata as knowledge base and as a hub to other knowledge bases (>2000 external identifiers) to link identifiers and get more properties.
The idea: gradually adopt Wikidata as a universal thesaurus because ...

- Libraries and other institutions from different disciplines increasingly want to get connected
- Connecting to a central hub is more efficient than connecting everything to everything
- Using the same identifier for the same resource is more efficient than linking resources
- Inventing yet another identifier is less efficient than using a (rich) existing one: the Wikidata identifier
- Libraries may want to share responsibility for Wikidata as a common thesaurus
Create trusted links

- For bibliographic data many trusted links are available: from thesaurus to VIAF, from VIAF to ISNI and from ISNI to Wikidata
- In many situations links for persons, events and locations the links have to be created
Wikidata as hub, as central hub and as universal thesaurus

“Everything links to everything”

“Wikidata as central hub”

“Everything links directly to Wikidata”
Bibliographic record

<creator rdf:Resource="http://data.kb.nl/thesaurus/068350767">Albert Einstein</creator>


<creator rdf:Resource="http://data.bnf.fr/11901607/albert_einstein/">Albert Einstein</creator>

<creator rdf:Resource="http://id.loc.gov/authorities/names/n79022889">Albert Einstein</creator>

all become:


With index field: wd_id= Q937
Wikidata as universal thesaurus for libraries

Current situation: many to many links (many identifiers for single resource)

Proposed: everything links to Wikidata (same identifier for single resource)
Lower barriers

• We use a name authority thesaurus for **unique identification** of resources. Sharing a single identification across institutions will make such identification **globally unique** and usable.

• Minimizing the **number of hubs**, minimizing the **number of variations** for identical queries in different databases and minimizing the **required knowledge** will lower the barrier to connect to external sources.

• Sharing a global identifier makes it easier for institutions to connect without dramatically changing their infrastructure.

• Why standardizing “everything” but not resource identifiers?
Motivating use case: KB Historical newspapers
Named Entity Linking

- Named Entity recognition
- Get entities
- Enrichment and training
  - Store article id + resource ids
- Find the best candidate
- List with Einsteins
- Search entity
- Index DBpedia
- DBpedia
- Wikidata
- VIAF
- Etc.

T. v. Veen, WikidataCon, 2017
Enrichment infrastructure using SURFsara HPC cloud

KB research environment

- articles
- Enrichment process
- Enrichment database

SURFsara HPC environment

- Named entity recognition
- search
- DBpedia index
- disambiguation
Continuous improvement of disambiguation algorithm

- All DBpedia titles searched in news articles
- Named Entities searched in DBpedia
- Speedup by using HPC cloud SURFsara
- Using context and machine learning

At the end cycle to first article and overwrite earlier enrichments with newest algorithm
# From conventional entity linking to deep learning and beyond

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Link Recall</th>
<th>Link Precision</th>
<th>Link F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule based</td>
<td>.76</td>
<td>.76</td>
<td>.65</td>
<td>.70</td>
</tr>
<tr>
<td>Machine learning (SVM)</td>
<td>.84</td>
<td>.76</td>
<td>.83</td>
<td>.79</td>
</tr>
<tr>
<td>Neural network</td>
<td>.84</td>
<td>.73</td>
<td>.87</td>
<td>.79</td>
</tr>
<tr>
<td>Extra features e.g. word embedding</td>
<td>.85</td>
<td>.81</td>
<td>.82</td>
<td>.82</td>
</tr>
<tr>
<td>Extra Wikidata data, more training data</td>
<td>.87</td>
<td>.81</td>
<td>.86</td>
<td>.84</td>
</tr>
<tr>
<td>Entity embedding</td>
<td>.88</td>
<td>.86</td>
<td>.85</td>
<td>.85</td>
</tr>
<tr>
<td>New architecture</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Semantic search: index resource identifiers

- Get text for article X
- Get enrichments for article X
- Search articles with Wikidata id’s
- Text + Viaf id + Wikidata id etc.
- Semantic search (SPARQL) providing Wikidata id’s

Indexing

Newspaper index

Wikidata

Enrichment database

Einstein door Epstein, buste van professor Einstein, voorvaardig door den beroemden beeldhouwer Epstein (rechts).
SELECT ?p WHERE {
FILTER NOT EXISTS {
?place wdt:P17 wd:Q55 .
}
}
For the same query in the catalogue the Wikidata identifier is converted to the local thesaurus identifier.
Navigation example

- Semantic query between [ ], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor

Using square brackets the software tries a few Wikidata SPARQL queries and replaces this string by the Wikidata results.
Navigation example

- Semantic query between [ ], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor
Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- **Click on a linked named entity for more information**
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor
Navigation example

- Semantic query between [ ], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor
**Navigation example**

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- **Click on a property for searching more articles about resources with that property**
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor
Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor
Coverage, approaches and environments
Usage with different infrastructural impact

• Federated SPARQL queries to search in a local database and Wikidata

• Mixed use of conventional queries and SPARQL: using the output of SPARQL as conventional query input using Wikidata identifiers

• Generate “just in case links” to construct a query for an external database (lowest implementation barrier)
Mixed use of resource identifiers

• All items are in a local thesaurus and in some cases the Wikidata item contains the local identifier
• Some items in the local thesaurus contain a link to Wikidata
• Mixed use: Items in bibliographic records link to Wikidata or to the local thesaurus
• All items in bibliographic records link to Wikidata
• The local thesaurus is kept for administrative purposes or as backup. New items might be pre-staged in the local administrative thesaurus
**Current coverage**

- Wikidata items: 37,652,877
- KB thesaurus: 3,249,839
- Named Entities in newspapers (number unknown): 1,738,085

Wikidata

KB

Named Entities in historical newspapers
Institutions maintain links to more than one collection with different coverage. In case of the Koninklijke Bibliotheek, for example:

- Library catalogue
- Newspaper collection
- Parliamentary Papers
- DBNL
- Etc...
Use WD identifier when possible (mixed use)

Match and merge

Local thesaurus for administrative purposes only
Pros and cons of Wikidata as universal thesaurus

Pros

• Lower barrier to connect different databases
• It doesn’t require an advanced infrastructure to benefit
• Less maintenance
• Coverage of more domains
• Potentially richer set of properties

Possible objections

• Libraries perceive it as losing control over their thesaurus: unauthorized users may change items
• Different organizations or countries may have a different view on specific items
• There is a risk of duplicate entries for new items
• What if Wikidata disappears?
Summary and conclusions

- Wikidata can serve as universal library thesaurus
- Using Wikidata as single universal thesaurus facilitates identification of entities across organizations and sharing properties
- Replacing thesaurus identifiers with Wikidata identifiers in bibliographic records can be done gradually
- When the transition is complete thesauri can be kept for administrative purposes
- The use of Wikidata identifiers is not restricted to SPARQL: identifiers can also be indexed and used in conventional queries
Any questions?

theo.vanveen@kb.nl

http://www.kbresearch.nl/xportal