Linking FRBR Entities to LOD through Semantic Matching

Naimdjon Takhirov, Fabien Duchateau, Trond Aalberg

Department of Computer and Information Science
Norwegian University of Science and Technology

Theory and Practice of Digital Libraries (TPDL' 2011)
Berlin, Germany

- Vast amount of valuable (and thoroughly documented) metadata in library catalogs
- Need for a transition to semantic formats
- Transition requires a great deal of quality assurance
- Many applications utilizing Linked Open Data

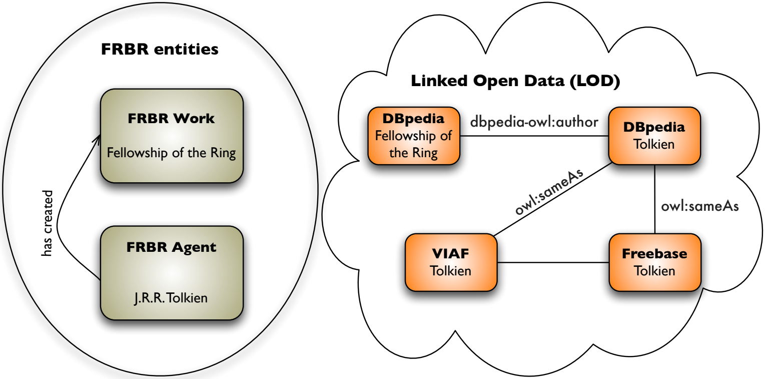
Outline

1. Introduction
2. Matching FRBR Works to LOD
   - Overview
   - Blocking
   - Matching Process
   - Filtering
   - Discussion
3. Experimental Evaluation
   - Protocol
   - Results
4. Conclusion and Future Work

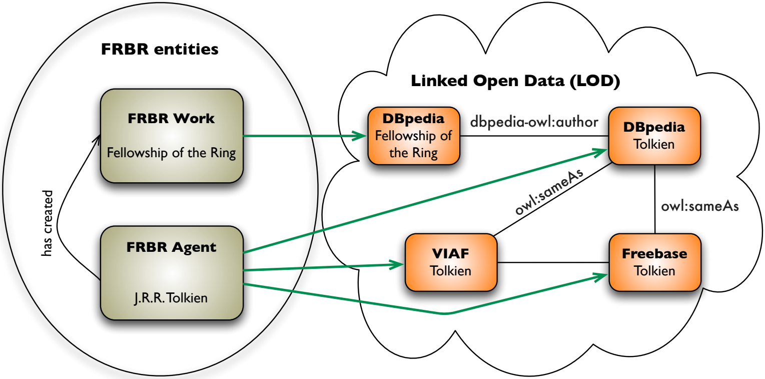
Motivation 1/2

- Enrich and verify FRBR entities with LOD information
- Reuse information and realize potential value of existing metadata
- Connecting FRBR entities to LOD to facilitate information discovery

Motivation 2/2



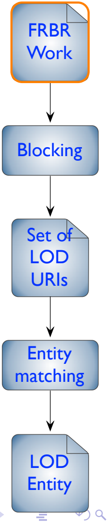
Motivation 2/2



Overview

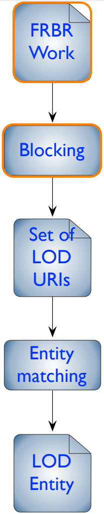
- The problem:
  a set of works $\mathcal{W}$ and a set of LOD entities $\mathcal{L}$;
  for each work $w \in \mathcal{W}$ and $l \in \mathcal{L}$, we note $\mathcal{F}$ the set of attributes shared by w and l;
  For an attribute $f \in \mathcal{F}$ shared by w and l, a similarity function is defined:

$$sim_f(w, l) \rightarrow [0, 1]$$



Blocking

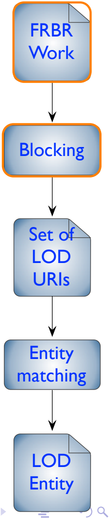
- LOD: millions of entities
  (e.g. Freebase contains ∼12M entities)
- We need a heuristic to select a subset of LOD entities, i.e., reduce the search space



Blocking (2)

- Obtain a subset of LOD entities by querying LOD
- Queries based on the FRBR attributes (e.g. title, date, etc.)
- A set of query tokens for each attribute:

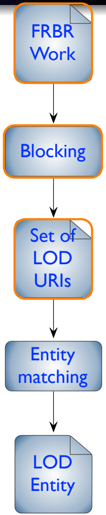
$$\begin{aligned} titles &\rightarrow \{title, normalized\_title\} \\ creators &\rightarrow \{creator_1, ..., creator_k\} \\ types &\rightarrow \{type, ext\_type_1, ..., ext\_type_m\} \end{aligned}$$



Sample queries

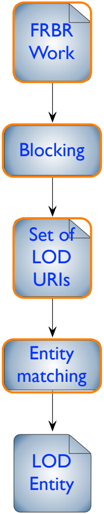
Type of Query	Query	# Entities
title	The fellowship of the ring (LOTR)	0
norm_title	fellowship ring	5
norm_title + ext_type	fellowship ring book	1
norm_title + ext_type	fellowship ring print	0
creator + ext_type	JRR Tolkien book	1
...		

The blocking process has reduced the number of candidate LOD entities against which we can now apply fine-grained matching techniques.



Matching (1)

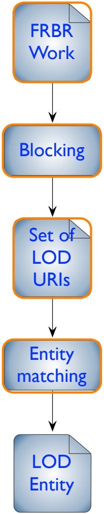
- Given a (reduced) set of LOD entities we need to match against FRBR entities
- Common attributes
  (name, creator, type, category, date of creation)
- Some attributes exist on both FRBR and LOD entities, while the others may be lacking



Matching (2)

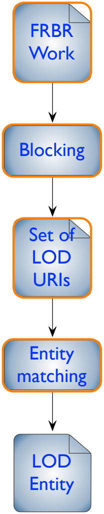
Individual similarity measures between properties of FRBR entity and properties of a LOD entity.

The nature of attributes are different. E.g., the *type* is word from a finite set of values while date can be in variety of formats (3 April 1978, 04.03.1978 or 03.04.1978 etc)

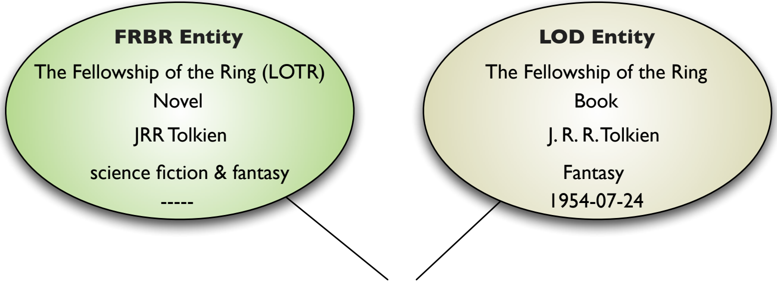


Matching (3)

- Attributes title and creator: three terminological similarity measures (Jaro Winkler, Monge Elkan and Scaled Levenshtein)
- Attribute categories: intersection of the sets of common categories
- Attribute type: using a small Wordnet-based taxonomy, evaluation based on the concepts that appear in the taxonomy
- Attribute date: extract only year as temporal granularity for works, hence the binary individual similarity



Matching (4)



FRBR Entity

The Fellowship of the Ring (LOTR)

Novel

JRR Tolkien

science fiction & fantasy

LOD Entity

The Fellowship of the Ring

Book

J. R. R. Tolkien

Fantasy

1954-07-24

Similarity measures:

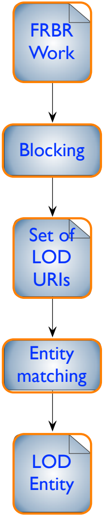
Title:	**0.77**
Type:	**0.29**
Creator:	**0.81**
Categories:	**0.00**
Date:	**0.00**

Matching (5)

- A global similarity value derived from individual ones
- A weighted average function to aggregate the values of all individual similarities
- Flexible with regard to applying weights
- Example: the DBpedia entity *The_Fellowship_of_the_Ring* and the work have a global similarity value equal to 0.37. As a comparison, the DBpedia entity related to the movie *The_Fellowship_of_the_Ring* obtains a similarity value of 0.22.

Filtering Candidate Matches

- Filter the candidate matches using one of the following strategies to filter candidate LOD entities:
  - selecting those with a similarity value above a given threshold
  - type-based constraint (e.g. "book")
  - top-k



Discussion

- Some entities are missing on the LOD
- Due to variations in the spellings and depending on the filter threshold, sometimes no LOD entity is returned by the blocking process, although the corresponding LOD entity might exist
- Not only used for verification purposes, but can also be a ground for adding new entities to the LOD
- A validation step is important

Experiment Protocol

- DBpedia Lookup Engine[1] as blocking process to reduce the set of DBpedia results as a set of candidates
- 684 FRBR works (extracted from product information found on Amazon), 343 with corresponding DBpedia entity
- Eight human judges performed manual validation

[1] http://lookup.dbpedia.org/api/search.asmx/KeywordSearch?QueryString=berlin

Quality Results

	Top-1	**Top-2**	**Top-3**
Number of True-Positives	**189**	197	201

Most of the correct matches (189) are ranked at the top. At top-3, we only discover 12 more entities.

Impact of threshold

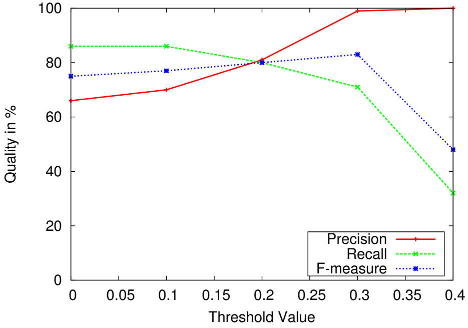


Figure: Quality Results (precision, recall, f-measure) w.r.t. a Threshold Filter

Impact of Weights

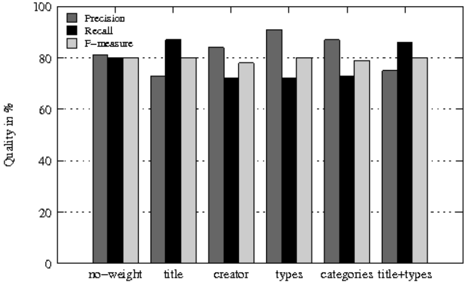


Figure: Quality Results w.r.t. the Weights of Individual Similarities

Conclusion

- A methodology to link a FRBR entity to its corresponding LOD entity
- A query builder as blocking process and refined similarity measures as matching process
- Most of the correct results at the top-1
- Verification and semantic enrichment
  - Integrating with various LOD sources
  - Linking an entity to a specialized database (e.g. *MusicBrainz* for music work)

Questions or Comments?

Linking FRBR Entities to LOD through Semantic Matching

Naimdjon Takhirov

takhirov@idi.ntnu.no