Linking FRBR Entities to Linked Open Data through Semantic Matching

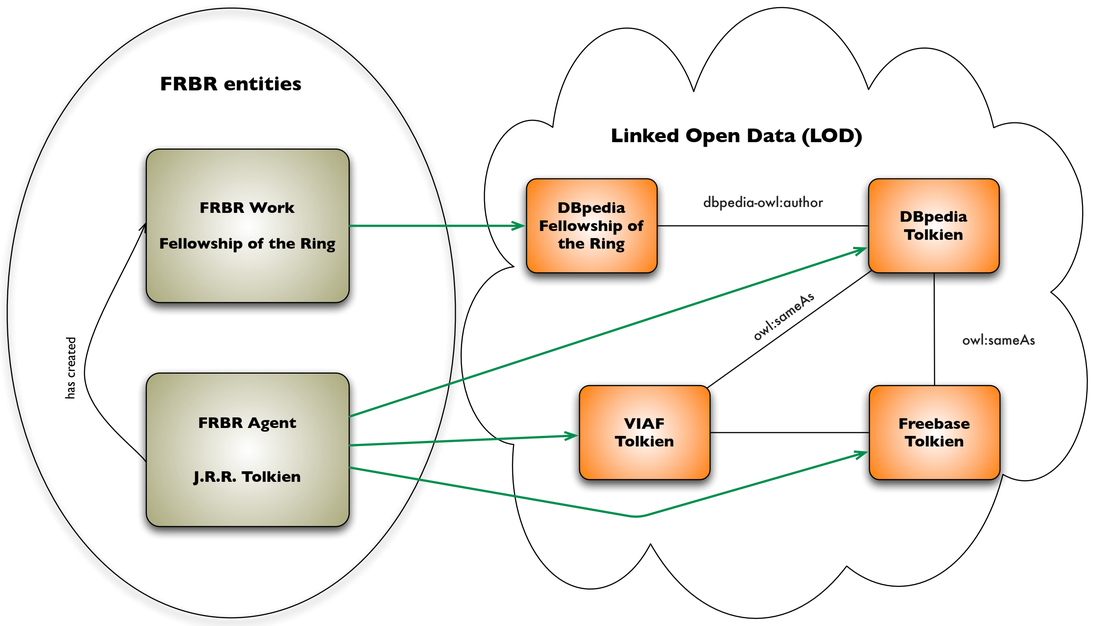
Motivation

Vast amount of valuable (and thoroughly documented) metadata in library catalogs

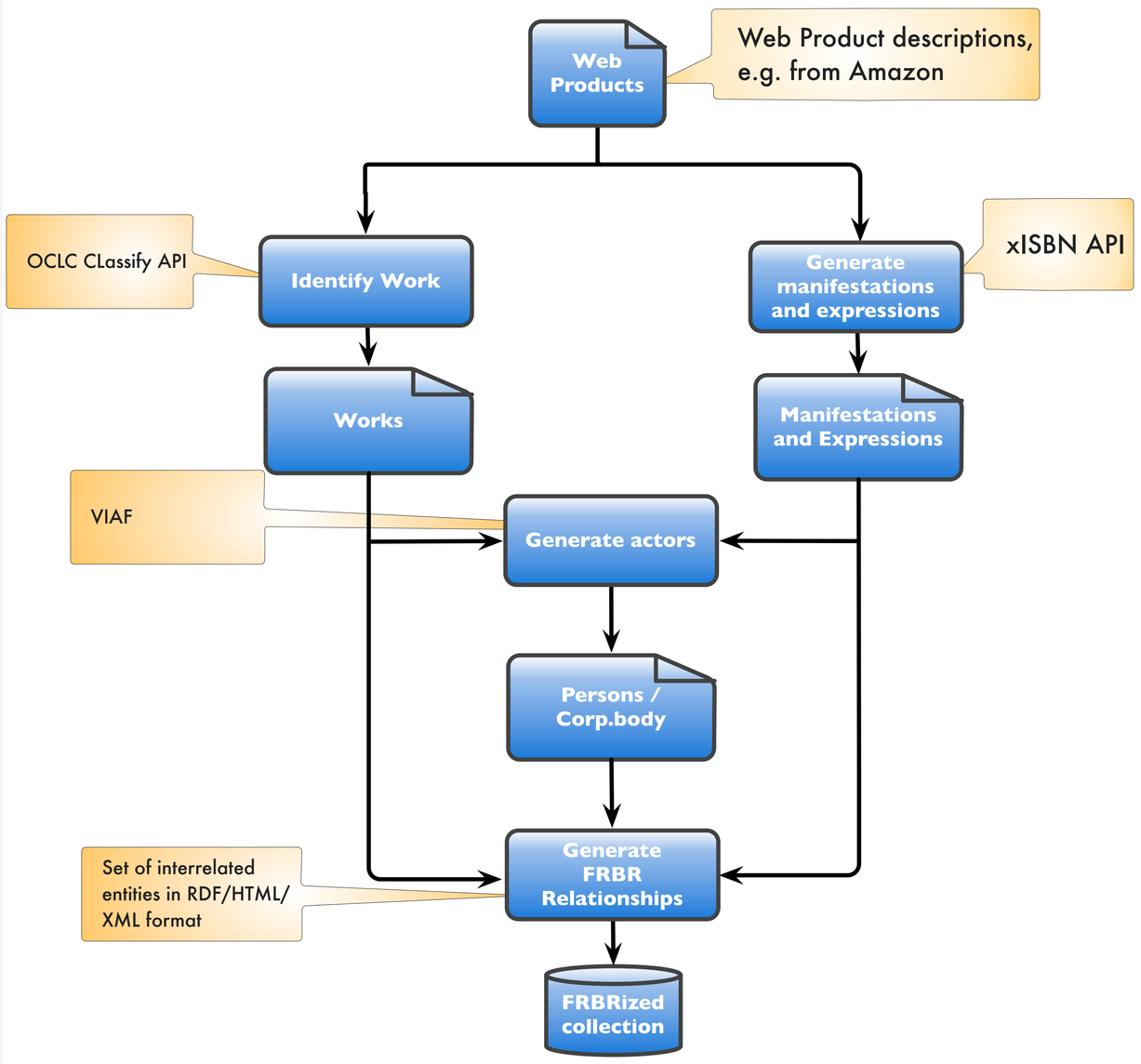
Need for a transition to semantic formats

Transition requires a great deal of quality assurance

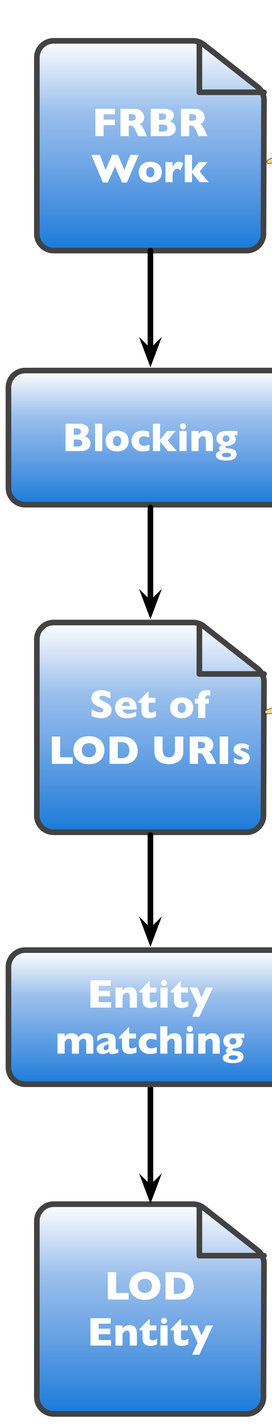
Many applications utilizing Linked Open Data



FRBRization



Linking FRBR entities to Linked Open Data



Input from the FRBRization Process

For each work, find the corresponding LOD entity

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

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: e.g,
titles -> {title , normalized _title }

Obtain a reduced set of candidate entities from the Blocking step

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

Similarity between titles/creators: three terminological similarity measures (Jaro Winkler, Monge Elkan and Scaled Levenshtein)
Similarity between categories: intersection of the sets of common categories
Similarity between types: using a small Wordnet-based taxonomy, evaluation based on the concepts that appear in the taxonomy
Similarity between dates: extract only year as temporal granularity for works, hence the binary individual similarity.

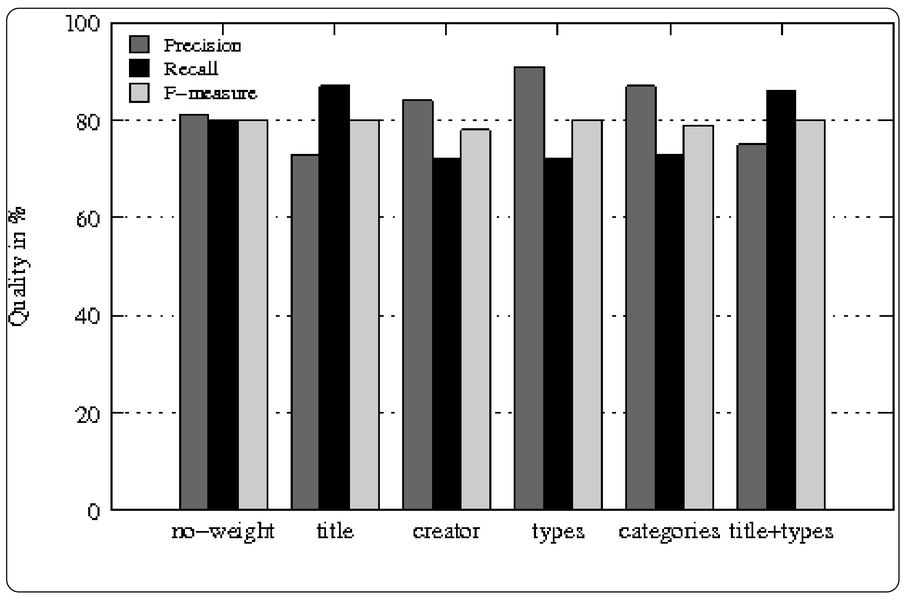
All individual similarities are aggregated to produce a global similarity value between the FRBR entity and a LOD entity.
Selection of the corresponding entity based on this global similarity value.

Experiment results

DBPedia Lookup Engine as blocking process to reduce a 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



A weight on the title which enables the promotion of recall (87%), i.e., allows us to discover more correct matches, but at the expense of precision
Type-based constraint filters out some candidates to promote precision (92%)

	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

Naimdjon Takhirov (takhirov@idi.ntnu.no)
Fabien Duchateau (fabiend@idi.ntnu.no)
Trond Aalberg (trondaal@idi.ntnu.no)

FRBRpedia demo: *http://j.mp/frbrpedia*

NTNU – Trondheim
Norwegian University of Science and Technology