BIBSURF - Discover Bibliographic Entities by Searching for Units of Interest, Ranking and Filtering

Trond Aalberg
NTNU
Trondheim, Norway
trondaal@ntnu.no

Tanja Merčun
University of Ljubljana
Ljubljana, Slovenia
tanja.mercun@gmail.com

Maja Žumer
University of Ljubljana
Ljubljana, Slovenia
maja.zumer@ff.uni-lj.si

ABSTRACT

BIBSURF is a system demonstrating search, ranking and filtering of bibliographic RDF data that is organized in form of entities representing intellectual endeavor at different levels of abstraction: item, manifestation, expression, work.

Keywords

Models; RDF; LRM; keyword search; ranking; filtering

1. INTRODUCTION

Memory institutions such as libraries and museums have embraced the Semantic Web as the main enabler for reuse and exploration of cultural heritage data. Significant effort has been invested in the development of reference models and metadata schemas, and large amounts of data are already made available as result of the Linked Open Data movement. Within the library community main focus has been on FRBR and other reference models published by the The International Federation of Library Associations and Institutions (IFLA), which are being merged into a common Library Reference Model (LRM) [6]. The core of this model is the depiction of intellectual endeavor and products at different levels of abstraction: *item*, *manifestation*, *expression*, *work*, as well the *agents* related to these entities. Various vocabularies have been published for implementing and coding such data in RDF, with the RDA vocabulary[1] currently appearing as the most relevant as it builds directly on the IFLA models and the RDA international cataloguing rules.

However, usage of this model in real world applications tailored to the needs of real users has been less systematically explored, although some protoypes and studies can be found [4, 2, 7]. Model and vocabulary development as well as the creation of data sets is often driven by domain expert and needs to be complemented by best practice knowledge from application development and studying users. *BIBSURF* is developed to experiment with search, ranking and "after-search" filtering of library data that is shaped according to the Library Reference Model and coded and stored as RDF. A main difference from traditional library search systems is that we now are interacting with databases containing entities of different types and the system has to make decisions on what type of entities the user has a preference for, what constitutes a meaningful unit in the results listing, what related entities to include in the unit and how to present other relationships. Experience from development and use of the system will also be important to determine how specific bibliographic patterns need be represented. The main contributions of the demonstration are:

- Demonstrate indexing and search on top of LRM data represented and stored in RDF.
- Highlight major design and implementation issues related to indexing and ranking.
- Exemplify solutions for listing and presenting results.
- Contribute with best practice examples for implementing the LRM model.

2. SYSTEM DESCRIPTION

The core of our search system is a database for storing data in RDF, with support for text indexing and keyword search. The current dataset covers a variety of bibliographic patterns and is originally extracted from library catalogues, but enhanced and transformed into rich LRM-data coded in RDF utilizing the RDA vocabulary.

The web-based user interface is inspired by library search interfaces with a single field for entering keywords. Rather than the traditional listing of single publications, BIBSURF organizes the found bibliographic entities in units that correspond to the abstraction level of interest to the end user, as well as improves the search experience by presenting an efficient and comprehensive listing that further can be explored using filtering and hide-show features.

2.1 Indexing and ranking

Keyword search in RDF data is a needed complement to structured queries, particularly when it comes to information needs of users of bibliographic data. A main question when implementing text-indexing and search for RDF is what unit to extract terms from, and what unit to return when querying the index. Different approaches have been explored in research e.g. [3] and most current triplestores have some kind of support for text indexing such as

[1] http://rdaregistry.info/

JCDL '16 June 19-23, 2016, Newark, NJ, USA

ACM ISBN 978-1-4503-4229-2/16/06.

DOI: http://dx.doi.org/10.1145/2910896.2925434

the native repository in the Sesame framework[2] which supports indexing RDF subjects by including the text from associated literal properties [5]. In our context this means that it is possible to index each work, expression etc. as separate units. A more flexible approach is supported in the GraphDB triplestore[3], which extends the subject-based indexing with the possibility to predefine property chains. Essentially this means that it is possible to index larger units of RDF data, sometimes called fragments or RDF molecules [1]. Other alternatives include storing the RDF data in XML databases that support text indexing. Our solution uses property chaining and precomputes RDF fragments which are stored and indexed using the eXist open source database[4]. An important side effect of precomputing fragments is that it removes the severe performance penalties for constructing fragments from the query result runtime.

2.2 Units of interest to users

The decisions on what units to index and return for queries need to be aligned with the expectations of the end users of the system: what we have named as *UIU* - Unit of Interest to User. In BIBSURF, we index works, expressions and manifestations based on the assumption that from any given node of these types we can follow property chains to collect the bag of terms that are relevant for each entity. An example for indexing works is shown in Figure 1 and a query explaining the logic behind this chaining could be "Orient Christie Suchet Movie". In this case it is natural to assume that the top ranked item should be the movie adaptation of Agatha Christie's "Murder on the Orient Express" where David Suchet plays the famous detective. To create the bag of terms related to this unit, one needs to traverse the graph to find the related entities that naturally describe the index centroid, as shown using arrows. The indexing of the original novel would follow other paths and will reach Suchet through the expression where he is the narrator, but this unit will not include the term Movie and will thus be ranked lower.

2.3 Display and interaction

Various forms of grouping items when displaying the results in bibliographic search are increasingly being used. Because of the nature of existing library records such groupings are hard to do consistently and systematically and most implementations are only loosely based on LRM. As our data is elaborated with a rich set of relationships, we have the possibility to experiment with groupings and listings beyond what is found in available library search interfaces. In our display we utilize tabbed boxes for showing e.g. works with tabs for each category of expressions. Other techniques included are hide and show features to display subtrees such as content listings for manifestations.

The use of "after search" filtering is a technique that naturally complements the grouped display lists. Checkboxes for the various facets of the search results, such as the person names related to entities in the result, enable users to easily explore subsets of the results.

To support comparative studies of user preferences we have developed three alternative displays: respectively work-centric, expression-centric as well as a manifestation-centric. In addition, we have a feature for looking up each UIU both as RDF-fragment and as a SVG-based graph.

[2]http://rdf4j.org

[3]http://ontotext.com/products/graphdb/

[4]http://exist-db.org/exist/apps/homepage/index.html

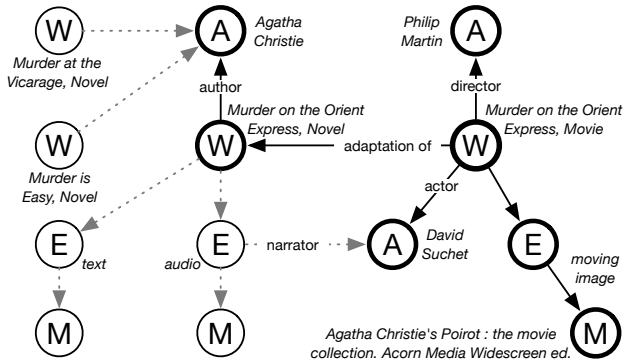

Figure 1: Indexing works.

2.4 Conclusion

BIBSURF is a search application developed to support the exploration and implementation of search systems for RDF-based bibliographic data. The system is currently used in research to extract best practice knowledge for implementing the Library Reference Model and represents an efficient approach for searching such data. Future work includes generalizing the system to support other reference models and adding support for large scale user studies by implementing logging of search sessions.

3. REFERENCES

[1] Li Ding, Tim Finin, Yun Peng, Paulo Pinheiro da Silva, and Deborah L. McGuinness. Tracking RDF Graph Provenance using RDF Molecules. Technical Report TR-05-06, UMBC, 2005.

[2] Zorana Ercegovac. Multiple-Version Resources in Digital Libraries: Towards User-Centered Displays. *JASIST*, 57(8):1023–1032, 2006.

[3] Wangchao Le, Feifei Li, Anastasios Kementsietsidis, and Songyun Duan. Scalable Keyword Search on Large RDF Data. *IEEE Transactions on knowledge and data engineering*, 26(11), November 2014.

[4] Tanja Merčun, Trond Aalberg, and Maja Žumer. FrbrVis: An Information Visualization Approach to Presenting FRBR Work Families. In *TPDL 2012*, volume 7489 of *LNCS*. Springer, September 2012.

[5] Enrico Minack, Leo Sauermann, Gunnar Grimnes, Christiaan Fluit, and Jeen Broekstra. The Sesame LuceneSail: RDF Queries with Full-text Search. NEPOMUK Technical Report, 2008.

[6] Pat Riva and Maja Žumer. Introducing the FRBR Library Reference Model. In *IFLA WLIC 2015*, Cape Town, South Africa, August 2015.

[7] Krzysztof Sielski, Justyna Walkowska, and Marcin Werla. Methodology for Dynamic Extraction of Highly Relevant Information Describing Particular Object from Semantic Web Knowledge Base. In *TPDL 2012*, volume 7489 of *LNCS*. Springer, September 2012.