Supporting FRBRization of Web Product Descriptions

Naimdjon Takhirov, Fabien Duchateau* and Trond Aalberg

Norwegian University of Science and Technology NO-7491 Trondheim, Norway
{takhirov,fabiend,trondaal}@idi.ntnu.no

Abstract. The FRBR model has the potential for new services and discovery techniques for cultural items such as books, movies and music. In this paper, we present an approach to interpret descriptions found in Web resources and identify the FRBR entities these pertain to. To verify the resulting set of FRBR entities, we have used the Linked Open Data and the verifications have been validated by a group of experts. The results of this work demonstrates applicability of FRBR in a new context and establishes a firm basis for further exploitation.

1 Introduction

The entity relationship model proposed in the IFLA Functional Requirements for Bibliographic Records (FRBR) [6] provides a framework for the entities and relationships that are of interest to end users of metadata. The model builds upon current practice and understanding of what commonly is described in metadata, and defines a formal framework for explicit statements about the entities and relationships that such descriptions pertain to. The interpretation or conversion of bibliographic records is a topic explored in different projects [4, 5, 2, 7]. Projects so far have mainly focused on library catalogs, but the FRBR model is generally accepted as sufficiently generic to serve as a conceptual framework for a broad range of metadata related to cultural items. The major benefit of the FRBR is that it can be used to describe intellectual and artistic contributions at different levels of abstraction. The model enables collocation of entities based on their intellectual equivalence and the relationships between the entities provides a network-based structuring of the entities described in metadata.

For many users the Web is the primary source of information and the total amount of data available online is far larger than the one stored in library catalogs. However, this Web data is often neither well structured nor machine-interpretable, although the emergence of the Semantic Web aims at tackling this issue. For a large portion of Web data there is, unfortunately, no easy transition to the Semantic Web. Simply transforming from one format to another, which is a syntactic approach, does not automatically enable semantic interoperability and input data often needs to be reinterpreted into entities and properties that make sense as well as transformed.

* The author carried out this work during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

In this paper, we propose to fill the gap between these two worlds, using the FRBR model for product information found on the Web. The Web contains a substantial amount of resources that describe products of creative or artistic endeavor such as online stores and community sites. We present an approach to interpret such information and identify the entities of the FRBR model. We advocate that such a representation would enable websites (e.g. e-commerce) to better organize and exploit those products.

2 Functional Requirements for Bibliographic Records

FRBR is a conceptual model of the bibliographic universe published around a decade ago [6]. Intellectual and artistic contributions are modeled in multiple levels of abstraction using the entities: **work**, **expression**, **manifestation**, and **item**. Figure 1 depicts this hierarchy. The three-part epic by J.R.R. Tolkien "The Lord of the Rings" is an abstract work encompassing "The Fellowship of the Ring", "The Two Towers", and "The Return of the King". Work represents a distinct, intellectual or artistic creation. Each of these three works has been translated and published in number of languages and each of those translations/editions is an expression in the FRBR model. This is illustrated by the realization of "The Two towers" in two different languages: the original English version "The Two Towers" and a Norwegian translation "To Tårn". The paperback format in original language (English) published by Mariner Books in 2005 is regarded as a manifestation in FRBR terms. In our example, the paperback edition published in September 2003 and June 2005 are thus regarded as the two separate manifestations of the same (English) expression. Finally, the physical book that one can hold in his hand is an item.

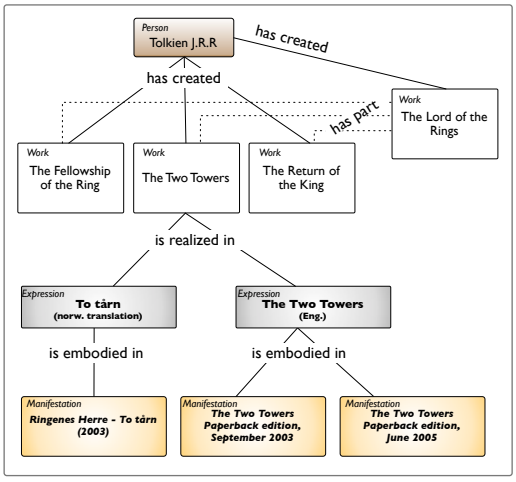

Fig. 1. A Fragment of Lord of the Rings FRBR Work by J.R.R. Tolkien.

A person (or corporate body), sometimes also referred to as *actor*, in the FRBR model, is an individual responsible for the creation or realization of a work (e.g., as an author, an illustrator, a translator, etc.). At the same time, this type of entity can be the subject of a work.

Additionally, the FRBR model provides **a set of relationships** between entities beyond the basic relationships shown in Figure 1. This feature helps to cluster a work and its related entities (e.g., an adaptation of a book in a movie), ultimately leading to better user experience when searching and exploring a collection.

3 FRBRizing Web Product Descriptions

One of the main difference between existing FRBRization approaches and our work deals with the input data. Our FRBRization process takes as an input descriptions of products found on the Web, specifically products sold by e-commerce websites (e.g., Amazon). Information about products tend to have different properties from their bibliographic record counterparts found in library catalogs. A first difference is that products do not have the same structural pattern as MARC records, and they are stored in a variety of formats. A second one deals with the identification of products, which are unambiguously referenced by URI, thus providing a basis for reuse and exchange [3]. Additionally, e-commerce websites usually provide faceted navigation where the ranked list of results can be filtered on several dimensions. Yet, Web products can be related to FRBR manifestation level, similarly to library catalogs. For example, the 2005 paperback version of the book "The Two Towers" sold for $10.95 at `hmhbooks.com` is a product which is an original English expression of the work "The Two Towers" by Tolkien.

Subsequently, our approach consists of a set of interrelated operations which result in identification of the different FRBR entities. The FRBRized entities are then connected by establishing appropriate FRBR relationships. The FRBRization workflow is illustrated in Figure 2. From the input product descriptions, we first identify the corresponding works (Section 3.1), then we generate related manifestations, expressions (Section 3.2) and actors (Section 3.3). The process of creating relationships between the FRBR entities, presented in Section 3.4, produces the FRBR collection.

3.1 Identifying a Work

The FRBR work is an abstract distinct intellectual or artistic creation. This entity is the cornerstone of the FRBR model and any FRBRization process needs to include a method for identifying the work entities. Description of a single product on the Web reflect the manifestation level, but attributes of expression and work can often be found in the description of the product. For example, the language of the book is an attribute of the expression while the title and author(s) may refer to the original work.

The following techniques can be used to identify a work:

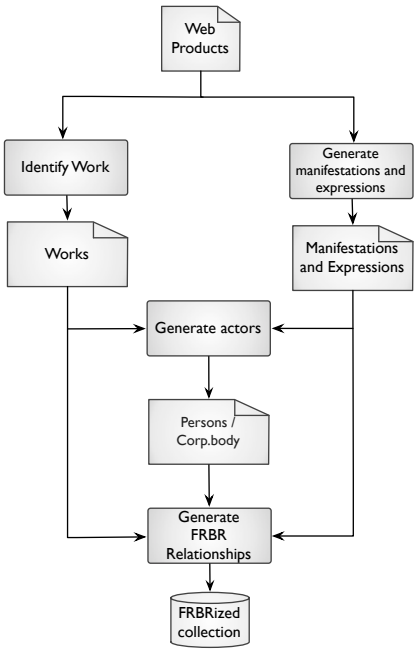

Fig. 2. The FRBRization Workflow

- Creating a work based on title/author and other attributes of the resource if the work has not been created yet; the database of works is then incrementally updated;
- Using an external service to identify a work, e.g. OCLC Classify API[1] for books using ISBN number, ISMN/ISRC music database for music or IMDB API for movies;
- Use z39.50 (or SRW/SRU) protocol to search and fetch the relevant MARC record from publicly available catalogs and then use similar technique to Work-Set algorithm by OCLC [5].

The two latter methods take an identifier as input. On the contrary, the first method requires attributes such as title/author, thus leading to string matching problem. For instance, if the input product description is a translation of a work, we have to make sure that the correct work is discovered.

3.2 Generating Related Manifestations and Expressions

The FRBR expression is often is perceived as the most difficult entity because information that is needed to distinguish between such entities often is ambiguous or missing. On the other hand, because the expression is an intermediary node between work and expression, we are able to generate expressions by clustering the manifestations that are related to a work using any expression level

[1] `http://classify.oclc.org`

attribute. The main challenge is to discover the related manifestations, and we have identified the following methods to achieve this goal:

- Search for author in z39.50 enabled repositories;
- Use external service (e.g. xISBN[2] or ThingISBN[3], Spotify [4]).

These methods rely on external sources. Note that using the first method implies to employ FRBRization techniques already proposed for MARC records [7, 5, 1].

From the set of related manifestations, we automatically generate expressions by analyzing attributes pertaining to the expression level. For example, attributes such as language and translator are used to identify expressions. Note that the identifier for an expression is automatically generated at this stage.

3.3 Generating Actors

An actor is a person or corporate body (organization) responsible for the creation or realization of a work. Products available from e-commerce websites usually have information about the responsible for the work, such as author of a book, composer of a music, director of a movie. Generating an actor can be performed using the following methods:

- Create a local authority file or use existing authority files from external sources;
- Search Virtual International Authority File (VIAF) and link actors to VIAF.

Contrary to the first method which is a time-consuming and complex task, the second approach includes a Web-based API and the VIAF collection contains data from many national libraries around the world. At the end of this step, we have generated all the FRBR entities required in our FRBRization process.

3.4 Generating FRBR Relationships

The final phase of the workflow is to generate the actual FRBR entities and establish relationships. Since we have information about each entity from previous steps, this step creates a collection of entities in a specific output format such as a series of SQL statements that can be used to insert into a relational database, HTML, XML, RDF or a simple text file. This step requires that an entity has a unique identifier and can be unambiguously referenced. For manifestations, we already have identifiers. Work and expression entities can be assigned locally generated identifiers since there is no publicly available global unique identifiers for these entities.

[2] `http://labs.oclc.org/xisbn/`
[3] `http://www.librarything.com/api`
[4] `http://developer.spotify.com/en/libspotify/`

4 Experiments

In this section, we demonstrate the use of our approach and the level of quality we have achieved. We have chosen to create our dataset based on search results from Amazon since its database potentially contains a great number of items[5]. Another reason to use Amazon is that some of the sites already implement a feature that is comparable to FRBR model by presenting users with a list of alternative formats. This is, however, solely based on metadata equivalence and is limited to publications that appear under the same title and author.

4.1 Experimental protocol

Using Amazon's Product Advertising API [6], we have searched for works by the 80 best selling fiction authors extracted from Wikipedia[7]. Due to the constraints set forth by Amazon on the number of requests that can be sent in one hour, we have a limited number of items to the first page of the results set (10 items per page). We have performed an automated search using the *ItemSearch* operation on *Books* index. We excluded items representing kindle edition. We also filtered out the products not solely offered by Amazon ("MerchantId"=Amazon). Additionally, we performed search on Amazon's *Video* and *Music* indexes using previously submitted queries on *Books* index. The attributes made available within these products, among others, are, title, author (director for movies), contributor, ISBN, language, release date. As can be seen from Table 1 (column "# of Input Products"), half of the content of the initial set of products were books. Most of these books are published in English language, but the input products include other languages such as Japanese, Chinese or Russian.

In Section 3.1 we have proposed three methods to identify the work corresponding for a product. To avoid implementing z39.50 protocol and reduce latency, we used the Classify API by OCLC. Classify API is a web service from the OCLC Office of Research that can be used to retrieve information, such as work level title, that is common the group of publications that belongs to the same work (identified by the use of OCLC's workset algorithm). The next step in the workflow is to generate related manifestations and expressions. Since we chose OCLC Classify API to identify work, we had a greater chance of match in the same database. Therefore, to obtain a list of related manifestations, we again used an OCLC Service - xISBN. The xISBN Web service returns ISBNs and other information associated with an individual intellectual work that is represented in the WorldCat catalog.

The next step involves the identification of actors. We chose to link actors (persons and corporate bodies) to Virtual International Authority File (VIAF). VIAF is a joint project of national libraries of several countries and it is hosted by OCLC. VIAF's long-term goal is to include authoritative names from many

[5] A blank search on "Books" generates 32,058,092 items (January 2011).
[6] `http://j.mp/amznAPI`
[7] `http://j.mp/fiction_authors`, as of November 2010

libraries into a global service that is available via the Web for free. Using VIAF's public API, we submitted queries for each contributor in the dataset. We used an average of Monge Elkan, Jaro Winkler and Levenshtein to calculate the similarity in the top 30 hits. The final phase of the workflow was to generate the relationship between the FRBR entities. This is achieved using the identifiers created for the FRBR entities in the previous steps. The final output is a set of RDF files for each entity type.

4.2 Quantitative Results

Table 1 summarizes the results of this experiment. The second column provides the number of Amazon products grouped by product type and the number of actors extracted from these products. In the third column, we show the number of discovered entities during the FRBRization process using Classify API, xISBN and VIAF services. Contrary to what could have been expected, the number of discovered works (739) is less than the number of input products (1216). This occurs because the set of input products contains different products that correspond to the same work (e.g. Norwegian "To tårn" and English "The Two Towers" both corresponding to "The Two Towers" work). We notice that the number of manifestations strongly increased (from 1656 to 28245) because we fetched all manifestations of works provided by the xISBN service. More specifically, we successfully discovered more books and videos while related music and DVDs were more difficult to fetch. The total number of actors we extracted was 2221 while 70% of them were found in VIAF (1569). The last column describes the

FRBR Entity	**# Input Resources**	**# Discov. Entities**	**# FRBRized Entities**
Work		739	684
Expression			5074
Manifestation	1656	28245	28245
-Book	856	27588	27588
-Video	102	542	542
-DVD	190	113	113
-Music	508	2	2
Actor	2221	1569	2221

Table 1. Results of the FRBRization

number of entities in our FRBRized collection, i.e., after removing unidentified works and actors. The initial set of input products we populated from Amazon contained 1656 items while the number of FRBRized manifestations is **28245**. The FRBRized collection includes works, translations, and movie versions of those works. Out of total 739 generated works, we have obtained a match for **684** works in Classify. The unidentified 55 works were mainly not in English language. Dealing with the actors, the final collection contains **2221** actors since the generated actors based on VIAF were automatically assigned locally generated identifiers. Finally, the following issues were encountered during the experiment:

- Search results from Amazon often needed to be cleaned. Some data was deemed as dirty, e.g. if the title referred to a movie rather than a novel such

as "The Lord of the Rings: The Return of the King (Widescreen Edition)". Since we performed search on Amazon database, we could not always limit our list to only works by our initial set of authors. This happens because authors could be mentioned in descriptive text of the resource;
- We could not FRBRize the whole set of product descriptions because the Classify service did not have an entry for all requested products. To solve this issue, we could aggregate the results from similar services (e.g. z39.50);
- VIAF had several identical entries for number of authors (e.g., "Arthur Rankin Jr."). In this case, the system chooses higher ranked item and if the score is identical, the item is chosen in random manner.

5 Conclusion

In this paper, we have demonstrated that the use of the FRBR model as a semantic data model is not only limited to library catalogs, but can be applied to product information found on the Web too. The main benefit of FRBRizing product information on the Web is that FRBR provides support for knowledge-like representation of the data enabling a broad support for exploratory interfaces where users are presented with a list of works for each author and can navigate relationships to learn about and find other versions or preferred editions of a given work. In the future, we plan to study support for more complex tasks such as automatic detection and extraction of aggregate works. Our FRBRization process can be further improved by using text analysis techniques. This means we automate the process of identifying entities and establishing relationships. At the application level, we plan to infer interesting relationships between the works linked to LOD.

References

1. T. Aalberg. A Process and Tool for the Conversion of MARC Records to a Normalized FRBR Implementation. In *Proc. of ICADL*, 2006.
2. N. Freire, J. L. Borbinha, and P. Calado. Identification of FRBR Works Within Bibliographic Databases: An Experiment with UNIMARC and Duplicate Detection Techniques. In *Proc. of ICADL*, 2007.
3. A. Gerber and J. Hunter. A compound object authoring and publishing tool for literary scholars based on the IFLA-FRBR model. *International Journal of Digital Curation*, 4(2), 2009.
4. K. Hegna and E. Murtomaa. Data mining MARC to find: FRBR? In *68th IFLA General Conference and Council*, Glasgow, Scotland, 2002.
5. T. B. Hickey, E. T. O'Neill, and J. Toves. Experiments with the ifla functional requirements for bibliographic records (FRBR). *D-Lib Magazine*, 8(9), 2002.
6. IFLA Study Group on the FRBR. Functional requirements for bibliographic records, final report. *UBCIM publications ; new series*, 19(1), 1998.
7. H. M. Manguinhas, N. M. A. Freire, and J. L. B. Borbinha. FRBRization of MARC records in multiple catalogs. In *Proc. of JCDL*, Gold Coast, Australia, 2010.