Open Datasets for Evaluating the Interpretation of Bibliographic Records

Joffrey Decourselle, Fabien Duchateau, Trond Aalberg, Naimdjon Takhirov, Nicolas Lumineau

Open Datasets for Evaluating the Interpretation of Bibliographic Records

Joffrey Decourselle
LIRIS, UMR5205
Université Lyon 1
Lyon, France
jdecours@liris.cnrs.fr

Fabien Duchateau
LIRIS, UMR5205
Université Lyon 1
Lyon, France
fduchate@liris.cnrs.fr

Trond Aalberg
NTNU
Trondheim, Norway
trondaal@idi.ntnu.no

Naimdjon Takhirov
Westerdals - Oslo School of Arts, Communication and Technology - Faculty of Technology
Oslo, Norway
taknai@westerdals.no

Nicolas Lumineau
LIRIS, UMR5205
Université Lyon 1
Lyon, France
nluminea@liris.cnrs.fr

ABSTRACT
The transformation of legacy MARC catalogs to FRBR catalogs (FRBRization) is a complex and important challenge for libraries. Although many FRBRization tools have provided experimental validation, it is difficult to evaluate and compare these systems on a fair basis due to a lack of common datasets. This poster presents two public datasets (T42 and BIB-RCAT) intended to support the validation of the FRBRization process.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Collection

General Terms
Measurement

Keywords
Migration, record interpretation, FRBRization, FRBR, dataset

1. INTRODUCTION
Libraries have traditionally relied on the MAchine Readable Cataloguing (MARC) format, available in different implementations such as MARC21 or UNIMARC, for the recording and exchange of bibliographic data. The semantics of MARC formats reflect the old-fashioned card catalogue which has obvious limitations and new models, such as the Functional Requirements for Bibliographic Records (FRBR) and its updated version Library Reference Model[1] (LRM), have been developed to provide library systems with a more sound and explicit information model for the next generation of library systems [2].
A major obstacle to the adoption of new models is the migration from the legacy MARC formats and the interpretation and transformation of existing data into the new models (e.g., FRBRization). In the last decade, many tools have been proposed to tackle this challenge [3], but it is very complicated to compare tools: the experiments which are described in the papers are rarely reproducible, mainly because the datasets are not publicly available. A few catalog excerpts are provided, but they do not reflect the reality and the challenges of library catalogs because they are mainly used for illustrating specific cases [1].
In this poster, we present two datasets (T42 and BIB-RCAT) for evaluating the FRBRization process. The goal of the first dataset is to identify the weak and strong points of a tool by testing all possible issues that libraries may face during FRBRization. The second dataset BIB-RCAT is extracted from catalogs of different cultural institutions and can be used for comparing or experimenting with the data quality that is typically found in real world catalogs. The datasets, released under a CC BY-NC licence[2], are available online at http://bib-r.github.io/.

2. DESCRIPTION OF THE DATASETS
In our context, a dataset is a set of collections. Each collection, which contains records, is available in two input formats (MARC21 and UNIMARC) and it is associated with an expert FRBR collection (gold standard). This expert collection has been manually created and verified by a librarian and three digital library researchers. The records have been extracted from real-world catalogs, and modified when needed to reflect bibliographic patterns and cataloging issues found in libraries.

2.1 Bibliographic patterns and issues
In bibliographic data there is a large diversity in the structure of entities and relationships needed to describe each item, but we can identify a set of patterns. Unfortunately, these patterns are often difficult to detect and FRBRize cor-

[1] http://library.ifla.org/1084/

[2] https://creativecommons.org/licenses/

rectly [1, 5]. The most frequent and thus **core pattern** includes a Work, an Expression, a Manifestation and (mostly) the Agent creator of the Work. Its FRBRization is relatively easy, unless the pattern is associated with cataloging issues. The **augmentation pattern** is defined as an additional content to an existing Work, with the assumption that the new content does not alter the main Work (e.g., illustrations, forewords). Several scenarios occur to FRBRize this pattern, for instance the creation of a new Work or a note for the original Work. The **derivation pattern** means that one Work is the modification of another Work (e.g., translations, imitations), and it usually implies the creation of Expression(s) under the same Work or relationships between Works. The **aggregation pattern** is commonly described as a whole-parts relationship (e.g., ensemble, aggregative work). The FRBRization of aggregations mainly results in the creation of relationships between Works (and "super-Works") and optionally new Agents. The **complementary works pattern** aims at modelling a relationship with Works which have the same importance (e.g., sequels, accompanying works). Its FRBRization mainly results in the creation of relationships between Works.

In addition to bibliographic patterns, records may include cataloging errors. Authors of the TelPlus project have established six requirements for FRBRization [4], that can be seen as errors in the initial records. They deal with **missing information**, namely record identifier, publication date, uniform title, original title, relator code and authoritative responsibility. These errors make it more difficult to FRBRize a record, for instance to discover the correct type of relationships between entities. We propose four new errors that can be found in catalogs. The **missing type and form of material** issue has an impact for correctly identifying Expressions (and sometimes Works). In UNIMARC, we can find **linkage error in title** and **linkage error in responsibility**, which means that the unavailable related record has a negative impact in terms of completeness when FRBRizing. Finally, libraries make use of standards such as the International Standard Bibliographic Description (ISBD), widespread normalization of values (e.g., country codes) or codes specific to individual libraries (e.g., for a book category, value "r" corresponds to a roman). These inconsistent **cataloging practices and norms** usually require human intervention to indicate how to process such fields.

2.2 Dataset T42

All records have an inherent bibliographic pattern (e.g., core, augmentation) and they may include any number of cataloging issues (e.g., missing relator code, title linkage error). The objective of the dataset T42 is to check whether a FRBRization tool is able to handle each possible case. We define a unit test as the combination of a pattern and an optional cataloging issue. Note that we do not include tests with more than one issue, since it would complicate the analysis of the results. We have ensured that the FRBRization is still possible when the issue deals with specific missing information. The dataset contains 42 meaningful unit tests which are crucial for testing specific aspects of FRBRization. For instance the *test 1.0* contains records with the core pattern and without issue, the *test 1.5* combines the core pattern with the missing uniform title issue and the *test 3.8* includes a derivation pattern and a missing relator code issue. The complete list of combinations is available online. Table 1 provides global statistics for the dataset T42 (second column). For example, this dataset includes records in three languages (English, French, German) and eight media types (e.g., books, movies, articles, audio).

Feature	T42	BIB-RCAT
Number of unit tests	42	-
Number of collections	126	3
Number of languages	3	1
Number of media types	8	4
Average (MARC) records	10/test	560
Average fields / record	18	17
Average (FRBR) entities	73/test	1922
Average (FRBR) properties	241/test	9517

Table 1: Statistics for datasets T42 and BIB-RCAT

2.3 Dataset BIB-RCAT

The BIB-RCAT dataset simulates real-world catalogs in which various bibliographic patterns and issues may be found. It contains three collections (MARC21 and UNIMARC formats, and the expert FRBR). It is mainly composed of records from various catalogs (e.g., a public French library). The size of this catalog (560 records) is smaller than the usual catalog in a library, since the expert FRBR collection requires a time-consuming effort to be manually produced and verified. Table 1 provides global statistics for the dataset BIB-RCAT (third column). For instance, the expert FRBR collection contains 1922 entities and 9517 properties.

3. CONCLUSION

In this poster, we present two datasets T42 and BIB-RCAT for evaluating the interpretation of bibliographic records. The first dataset enables to check how a tool performs when facing a specific bibliographic pattern or cataloging issue, while the second dataset reflects the data quality found in libraries. A perspective to this work deals with the definition of new metrics to evaluate the FRBRization process.

4. ACKNOWLEDGMENTS

This work has been partially supported by the French Agency ANRT (www.anrt.asso.fr), the company PROGILONE (www.progilone.com/), a PHC Aurora funding (#34047VH) and a CNRS PICS funding (#PICS06945).

5. REFERENCES

[1] T. Aalberg and M. Žumer. The Value of MARC Data, or, Challenges of FRBRisation. *Journal of Documentation*, 69:851–872, 2013.

[2] K. Coyle. FRBR, Twenty Years On. *Cataloging & Classification Quarterly*, pages 1–21, 2014.

[3] J. Decourselle, F. Duchateau, and N. Lumineau. A Survey of FRBRization Techniques. In *Theory and Practice of Digital Libraries*, pages 185–196, 2015.

[4] H. M. A. Manguinhas, N. M. A. Freire, and J. L. B. Borbinha. FRBRization of MARC Records in Multiple Catalogs. In *JCDL*, pages 225–234. ACM, 2010.

[5] P. Riva. Mapping MARC 21 Linking Entry Fields to FRBR and Tilletts Taxonomy of Bibliographic Relationships. *Library resources & technical services*, 48(2):130–143, 2013.