BIB-R: a Benchmark for the Interpretation of Bibliographic Records

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

BIB-R: a Benchmark for the Interpretation of Bibliographic Records

Joffrey Decourselle[1], Fabien Duchateau[1], Trond Aalberg[2], Naimdjon Takhirov[3], and Nicolas Lumineau[1]

[1] LIRIS, UMR5205, Université Claude Bernard Lyon 1, Lyon, France
firstname.lastname@liris.cnrs.fr
[2] NTNU, Trondheim, Norway trondaal@idi.ntnu.no
[3] Westerdals - Oslo School of Arts, Communication and Technology - Faculty of Technology, Oslo, Norway taknai@westerdals.no

Abstract. In a global context which promotes the use of explicit semantics for sharing information and developing new services, the MAchine Readable Cataloguing (MARC) format that is commonly used by libraries worldwide has demonstrated its limitations. The semantic model for representing cultural items presented in the Functional Requirements for Bibliographic Records (FRBR) is expected to be a successor of MARC, and the complex transformation of MARC catalogs to FRBR catalogs (FRBRization) led to the proposition of various tools and approaches. However, these projects and the results they achieve are difficult to compare on a fair basis due to a lack of common datasets and appropriate metrics. Our contributions fill this gap by proposing the first public benchmark for the FRBRization process.

Keywords: benchmark, migration, record interpretation, FRBRization, LRM, FRBR, MARC, dataset, evaluation metric

1 Introduction

Cultural institutions are responsible for cataloging and offering access to a large number of cultural items. The most popular format for libraries, MAchine Readable Cataloguing (MARC), available in different implementations such as MARC21 or UNIMARC, has shown some limitations in terms of interoperability, reuse, or information disambiguation [10]. The Functional Requirements for Bibliographic Records (FRBR) and its updated version Library Reference Model (LRM) [17] have been designed to provide a sound and more explicit semantics which will enable new enhancements such as improved navigation and enrichment features [3,5,11]. However, more than twenty years after the original specifications of FRBR, the model is still not widely used in libraries [6]. A major obstacle to the adoption of FRBR is the interpretation of records (e.g., FRBRization), which consists of migrating cultural heritage data from legacy formats (e.g., MARC) to models based on the FRBR semantics[4].

[4] For instance, RDA, the LD4L project or BIBFRAME

In the last decades, the proliferation of FRBRization tools has demonstrated the need for specific enhancements (e.g., clustered deduplication, exploitation of added entries) to improve the process [9]. Despite this progress, it is still very complicated to evaluate and compare FRBRization tools for several reasons. First, the experiments described in papers are rarely reproducible, mainly because the tools and the datasets are not 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 [2]. Last but not least, the current metrics are not sufficient to evaluate all possible cases that might occur during FRBRization. In addition, these metrics imply that the user has to wait until the end of the FRBRization process before obtaining any insight about the resulting quality. To summarize, we advocate that the lack of a common FRBRization benchmark is an obstacle to the adoption of FRBR.

In this paper, we propose BIB-R[5], the **first benchmark for evaluating FRBRization**. It is composed of two datasets and a set of evaluation metrics. The goal of the **first dataset T42** 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 three different cultural institutions and can be used for comparing or experimenting with the data quality that is typically found in real world catalogs. The assessment of the process relies on a **set of metrics** to predict the quality of the output and to evaluate the quality of a FRBRized catalog. An **experimental study** with three recent FRBRization solutions shows the benefits of our benchmark.

In the rest of the paper, we present related work in Section 2 and an overview of our benchmark BIB-R in Section 3. Evaluation metrics for pre-FRBRization and post-FRBRization are presented respectively in Sections 4 and 5 and the datasets are described in Section 6. The experimental study is detailed in Section 7. Section 8 concludes and outlines future work.

2 Related work

Issues related to FRBRization and identification of challenging bibliographic patterns are described in recent surveys [2,16,19]. In this section, we present rule-based FRBRization tools, and we focus on the evaluation of this process.

Tools. Due to page limitation, we refer to a recent survey for an exhaustive list of FRBRization tools [9]. The last decade has seen the emergence of rule-based FRBRization tools, since grouping-records tools are not able to process complex structures [2,12]. We focus on three rule-based tools that are publicly available for experiments. The tool VFRBR, developed in the context of the Variations project, aims at FRBRizing catalogs with a focus on the music domain [14,15]. Since music items are often described using added entries, VFRBR's strategy is to interpret these added entries as separate entities. The online catalog Sherzo is the proof of concept that lets users explore musical works,

[5] http://bib-r.github.io/

composers and related entities issued from 185,000 MARC records. Extensible Catalog (XC) is an open-source project for a complete Integrated Library System, which includes the FRBRization tool Metadata Service Toolkit [4]. XC exploits added entries and is therefore able to detect complementary works for instance. The third tool FRBR-ML [18] is based on Aalberg's approach [1]. The authors discuss the possible structures of the FRBR output catalog, and the tool provides enhancements to disambiguate some complex cases by exploiting other catalogs or Linked Open Data knowledge bases.

Evaluation. Only the output of the FRBRization process is evaluated, under various forms: the most frequent option requires a ground truth or gold standard, i.e., an expert FRBRized catalog [13]. The comparison between the expert FRBRized catalog and the FRBRized catalog produced by a tool indicates whether the tool is able to perform an acceptable transformation. One of the main issues is the manual construction of the expert FRBRized catalog. The FRBR-ML approach avoids the tiresome construction of a gold standard by converting the FRBRized catalog back to a MARC catalog [18]. The evaluation is performed between the initial MARC catalog and the converted one. With this type of evaluation, the drawback is that the last transformation (into MARC) may have a negative impact on the quality of the catalog. Besides, the rules that enable this last transformation have to be written too. To evaluate a process, metrics are required. In TelPlus, an aggregation metric is proposed to measure the percentage of aggregated content (e.g., Works, Persons, Places). FRBR-ML is evaluated with three metrics: redundancy, completeness and extension respectively measures duplicate data, loss of data and amount of enrichment.

Discussion. The digital library community has successfully identified the bibliographic patterns, as well as a few FRBRization issues. But there is no collection publicly available for testing each of these challenges. Thus, most FRBRization tools have been tested against private datasets, whose characteristics are not clearly defined. Available metrics either assess the deduplication (aggregation) or compare two MARC collections (redundancy, completeness and extension). Thus, there is no metric which compares the quality of a generated FRBR collection, especially in terms of bibliographic patterns. And the whole FRBRization process is currently not evaluated (e.g., the tuning task). Contrary to other research domains, there is no benchmark for one of the most crucial task in the digital library community. Yet, we advocate that understanding the weak and strong points of the FRBRization process tends to promote novelty and enhancements in the future implementations. In addition, common datasets and evaluation metrics enable a fair comparison between the tools. In the next section, we describe our benchmark for FRBRization.

3 Overview of the benchmark

The FRBrization process has been described and enhanced in the last decade [1,9]. Figure 1 depicts an overview of the FRBRization process. It is composed of three main steps (pre-FRBRization, FRBRization and post-FRBRization).

During pre-FRBRization, librarians are in charge of preparing the input catalog (traditionally in MARC) and tuning the tool. The optional preparation allows to clean fields, to delete empty records, etc. For the tuning task, librarians can configure some parameters (e.g., setting a decision threshold for deduplication), but the main challenge is to add, modify or delete rules. Next, the FRBRization starts using a clean catalog and a customized set of rules. The transformation of each record produces a set of entities and relationships according to the rules applied. A deduplication task is necessary to detect and merge entities that represent the same concept. Finally, the last step is post-FRBRization, during which optional tasks are performed on the raw FRBR collection [9]. We only mention validation and enrichment. The former enables expert to verify and correct the generated FRBR catalog while the latter refers to the task of adding information from external sources. Most FRBRization approaches only evaluate the last step (post-FRBRization, using the FRBRized catalog), but the initial step has a strong impact on the final result in terms of quality and performance.
Our benchmark BIB-R provides metrics and datasets to evaluate pre-FRBRization and post-FRBRization. It focuses on the foundations of FRBR, but it does not take into account the specificities and the complexity of the different implementations (e.g., FRBR-OO).

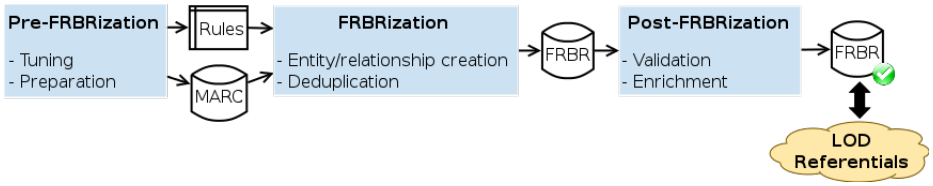


Fig. 1: Overview of the FRBRization process

4 Evaluating pre-FRBRization

The pre-FRBRization aims at preparing the catalog to be FRBRized and the set of rules. The records of a catalog include bibliographic patterns [2], which complicate FRBRization. The different ways of cataloging and the set of rules are subject to potential issues. Since FRBRization may last hours or days, it seems interesting to be able to detect and solve potential problems prior to running the process (e.g., updating the set of rules or cleaning fields and values). To the best of our knowledge, this detection of problems is a novel contribution for digital libraries. Our pre-FRBRization metrics analyze a set of rules according to a catalog for estimating the records that will not be FRBRized correctly. Each metric computes a percentage defined as the number of records concerned with the given pattern/issue divided by the total number of records.

Library practice allows for the description of different entities and relationships according to the nature of the item described. These structures can be

generalized into common patterns, but because the structure mainly is implicit such patterns are difficult to detect and FRBRize correctly [2,16]. The most frequent *core pattern* includes a Work, an Expression, a Manifestation and (mostly) the Agent creator of the Work. 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 optionally new Agents. The *complementary works pattern* aims at modelling a relationship between Works which have the same importance (e.g., sequels, accompanying works). The FRBRization of complementary works mainly results in the creation of relationships between Works. The **metrics COR**, **AUG**, **AGG**, **DER** and **COW** respectively compute the percentage of records with a core pattern, an augmentation, an aggregation, a derivation and a complementary work.

In addition to bibliographic patterns, records may include cataloging errors or quality problems. Authors of the TelPlus project have established six requirements for FRBRization [13], that can be seen as errors in the initial records. We define the **metrics MID**, **MPD**, **MUT**, **MOT**, **MRC** and **MAR** which respectively compute the number of records that include the issues *missing record identifier*, *missing publication date*, *missing uniform title*, *missing original title*, *missing relator code*, and *missing authoritative responsibility*. We propose four new metrics related to cataloging issues. The **metric MTF** deals with *missing type and form of material*, which has an impact for correctly identifying Expressions (and sometimes Works). The **metrics TLE** and **RLE** relate to *title linkage error* and *responsibility linkage error*, which means that the unavailable related record (mainly in UNIMARC) 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). The **metric CPN** deals with these inconsistent *cataloging practices and norms*, which may contain useful information.

The set of rules has not been widely studied and mainly regarded as an artifact that needs to be tuned by librarians. Yet, this tuning has a crucial impact and its analysis can be exploited to improve FRBRization. In case of *missing rule*, a field cannot be processed by a rule, thus causing loss of information. Detecting these fields prior to FRBRization enables librarians to update the set of rules accordingly. A *not used rule* indicates that it is not useful for a given catalog.The *conflicting rules* issue occurs because the set of rules can be built using various techniques (e.g., written by librarians, merged from collected sets).

Actions associated to such rules can either be complementary or conflicting, which may degrade performance or quality during FRBRization. The **metrics MR**, **UR** and **CR** respectively compute the percentage of missing rules, not used rules and conflicting rules. The metric MR can be decomposed into more detailed metrics for a given pattern or issue, for instance **MR-AUG** to calculate the percentage of missing rules for detecting all augmentations. A formal notation of the pre-FRBRization metrics can be found in an online appendix [7].

5 Evaluating post-FRBRization

When the process of FRBRization is finished, librarians typically need to check the FRBRized catalog produced by the tool. This evaluation is the most studied in the literature, because the resulting quality is currently an obstacle to the adoption of FRBR and because most FRBRization tools have demonstrated their capabilities through an experimental validation based on the analysis of the produced FRBR catalog [13,18]. In our context, we have chosen an evaluation based on expert FRBRized catalog. This is the most reliable evaluation, specifically because it directly assesses the quality of the FRBR catalog, including its complex relationships between entities. We have identified seven metrics to compare the FRBR catalog generated by a tool $\mathcal{T}$ and the expert FRBR catalog $\mathcal{E}$. These metrics are useful to understand the weak and strong points of a FRBRization tool, to estimate the manual effort which is needed to complete the FRBRization, or to provide an insight about the rules that should be added to improve the quality.

The first four metrics deal with data (entities, relationships and properties from the FRBR model). The **metric MD** is related to the missing data issue, i.e., data which appears in the expert catalog $\mathcal{E}$ is missing in the tool's catalog $\mathcal{T}$. This metric computes the ratio between the number of missing data and the total number of data in the expert collection. It can be redefined for each type of data, i.e., MD-E for entities, MD-R for relationships and MD-P for properties. The **metric IAD** deals with incorrectly added data, i.e., duplicate data (e.g., a property which appears twice in an entity, because of a bad deduplication for instance) and incorrect data (e.g., an entity that should not have been created or a property with an unexpected value). It is defined as the number of incorrect data in $\mathcal{T}$ (which is not in $\mathcal{E}$) divided by the total number of data in $\mathcal{T}$. Similarly to MD, the metric IAD can be redefined according to the data type. The **metric DLE** relates to errors in external link (e.g., to a referential or to the Linked Open Data). Either the link does not exist in $\mathcal{E}$ or it has a different value for the same external source. The metric calculates precision, i.e., the number of erroneous links in $\mathcal{T}$ divided by the total number of links in $\mathcal{T}$. The **metric SMD** aims at computing semantic mismatch data, i.e., data which have a different semantics in both catalogs (e.g., a relationship *translated by* which appears as *contributed to*). The metric computes the amount of semantic mismatch data in $\mathcal{T}$ (compared to data in $\mathcal{E}$) with regards to the total number of data in $\mathcal{T}$.

The next metrics deal with patterns. The detection of the pattern in a MARC record is crucial because it provides the FRBR structure. Yet, only part of a pattern may be incorrect and the evaluation should reflect this . Note that it is not possible to verify information about bibliographic patterns without annotations in the expert collection. The **metric MEND** (main entity not detected) relates to the detection of the main entity of a pattern (e.g., an Expression in the case of a translation). It measures the percentage of main entities from $\mathcal{E}$ that have been correctly detected in $\mathcal{T}$ among all main entities from $\mathcal{E}$. The **metric MRND** (main relationship not detected) checks whether the relationship associated to the main entity is correctly identified or not. For instance, an Expression is correctly identified but linked with a *"is a revision"* relationship rather than with a *"is a translation"* relationship. The metric MRND computes the percentage of main relationships from $\mathcal{E}$ that have been correctly detected in $\mathcal{T}$ among all main relationships from $\mathcal{E}$. Finally, the **metric ESE** deals with errors in secondary element(s) of the pattern, which means that the main entity and its relationship have been correctly detected, but other elements (e.g., the translator) are missing or incorrect. The metric ESE computes the percentage of correct secondary elements in $\mathcal{T}$ among all secondary elements from $\mathcal{E}$. A formal notation of the post-FRBRization metrics is given in an online appendix [7]. To use these metrics, it is necessary to have datasets with appropriate features.

6 Datasets

In BIB-R, two datasets allow the assessment of FRBRization tools. 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. This expert collection has been manually created and verified by a librarian and three digital library researchers. All collections included in these datasets are based on the MARCXML and raw MARC formats. The records have been extracted from real-world catalogs, and modified when needed. The datasets are detailed in a report [8] and publicly available at `http://bib-r.github.io/`.

The first dataset **T42**[6] can be used for testing specific cases. In Section 4, we explained that a record has an inherent bibliographic pattern (e.g., core, augmentation) and it may include any number of 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 issue. This dataset currently contains 42 meaningful tests which are crucial for testing specific aspects of FRBRization (a full list of combinations and statistics are available online).

The second dataset **BIB-RCAT**[7] simulates a real-world catalog in which various bibliographic patterns and issues may be found. It currently contains

[6] T42 is a reference to the novel *Hitchhikers Guide to the Galaxy*

[7] BIB-RCAT is a recursive acronym that stands for "*BIB-RCAT Is Basically a Real-world CATalog*"

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, a public Swiss hospital). The size of this catalog (560 records) is smaller than catalogs found in cultural institutions, since the expert FRBR collection requires a tiresome effort to be manually produced and verified.

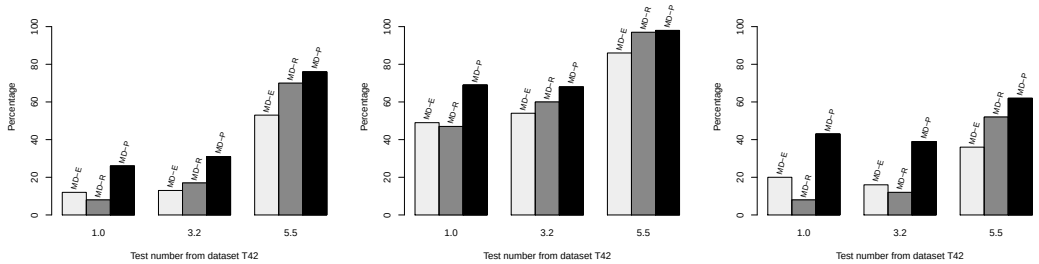
7 Experiments

In this section, we demonstrate the benefits of our benchmark BIB-R for the evaluation of FRBRization. Three tools, which are publicly available[8], have been used in these experiments: FRBR-ML, Extensible Catalog (XC) and Variations VFRBR. These tools are detailed in the Related Work (Section 2). The rest of this section describes three experiments using our benchmark: how to evaluate the strengths and weaknesses of FRBRization tools, how to compare tools in a real-world FRBRization scenario and how to facilitate the tuning of a tool. Due to page limit, only a few interesting results are presented, but all plots are publicly available in an online appendix [7].

7.1 Assessing strengths and weaknesses

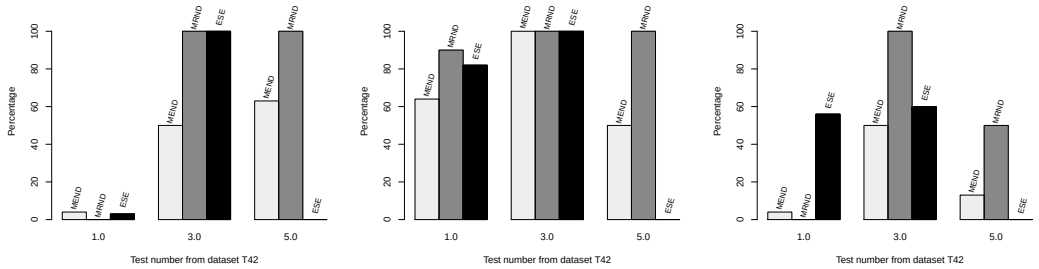
This first experiment aims at demonstrating the benefit of the dataset T42 when it comes to evaluating the strengths and weaknesses of FRBRization tools. For the three tools, we have run each test from the dataset T42 and the evaluation is performed using post-FRBRization metrics. Note that we have not tuned the rules and rely on the set or rules provided with the tools, although they have been developed for different purposes. The first finding is about **missing data**. None of the tools completely FRBRize the data contained in the MARC records. Figure 2 illustrates this trend by showing the missings in terms of entities (MD-E), relationships (MD-R) and properties (MD-P) for various tests. With the core pattern (test 1.0), the tools may miss entities such as Concepts. Tools may also be implemented to merge some properties. For instance, XC merges the subtitle into the title, thus missing the subtitle property. The scores of VFRBR for missing data are strongly impacted by the fact that it does not create Work entities. The more complex the record becomes (tests 3.2 and 5.5), the more losses in the FRBRization. Secondly, only XC generates **incorrectly added data**, mainly in terms of properties and relationships (see online appendix). These additional data are in fact misplaced data, i.e., which should have been put in another entity or which should have linked other entities (e.g., the abstract is placed in the Work entity rather than in the Expression). Another study deals with the **detection of patterns**. Figure 3 depicts the scores obtained by the three tools for correctly detecting the bibliographic patterns without any cataloging issue (i.e., tests x.0). FRBR-ML obtains good results for detecting

[8] FRBR-ML (previously named marc2frbr), Extensible Catalog and Variations VFRBR (adjusted version, only to facilitate compilation)



(a) FRBR-ML (b) VFRBR (c) XC

Fig. 2: Experiment results for evaluating missing data



(a) FRBR-ML (b) VFRBR (c) XC

Fig. 3: Experiment results for evaluating bibliographic patterns

the core pattern (test 1.0). For other patterns (tests 3.0 and 5.0), it discovers half of the main entities (metric MEND) but it fails for the main relationship (metric MRND). VFRBR is not able to detect most patterns with its basic set of rules, even for the core pattern. This is mainly because this tool does not always create Works and Expressions. XC achieves good results with some patterns (core, complementary works) but not for derivations. Next, we note that all tools produce **semantic mismatch**, but only for the relationships (metric SMD-R). This issue occurs in 36 tests for FRBR-ML, 24 tests for VFRBR and 21 tests for XC (out of 42 tests), but the scores of the metric SMD-R are mostly below 10%, thus indicating that less than 10% of the relationships have a different semantics than in the collection annotated by experts. Since the more complex relationships are usually found in patterns, these results are also dependent on the ability of the tool to detect patterns. To summarize, our dataset T42 and post-FRBRization metrics are useful for understanding the failures of a tool.

7.2 Comparing tools in real-world context

The objective of this second experiment is to compare FRBRization tools in a real-world context using post-FRBRization metrics. The post-FRBRization metric DLE is not presented, since the expert FRBRized collection cannot include a link for each existing authority files or knowledge bases. All tools rely on their basic set of rules (no tuning). Table 1 provides the results for the three tools. We note that they are able to identify only a few patterns (scores above 90% for the metrics MEND and MRND). VFRBR is the only tool to FRBRize half of the secondary elements of the patterns (ESE value equal to 55%). All tools successfully manage not to add incorrect data or produce different semantics (metrics IAD and SMD). However, they do not FRBRize almost half of the data (metric MD), mainly because of the incorrectly detected patterns. These average results for the three tools are understandable for several reasons: contrary to dataset T42, these real-world records from the dataset BIB-RCAT can combine several bibliographic patterns and issues. In addition, almost half of them include cataloging practices, which complicate the interpretation of the records. Finally, some additional entities (e.g., Concept) are not processed and created. The basic set of rules are not sufficient for achieving an acceptable quality. To conclude, this experiment showed that our dataset BIB-RCAT and associated metrics are useful to compare tools in a real-world context.

	FRBR-ML	VFRBR	XC	FRBR-ML tuned
MEND	94%	98%	94%	1%
MRND	100%	100%	100%	29%
ESE	99%	55%	100%	21%
MD	44%	45%	45%	13%
IAD	0%	0%	0%	0%
SMD	0%	0%	0%	0%

Table 1: Results of FRBR-ML, VFRBR and XC for the dataset BIB-RCAT

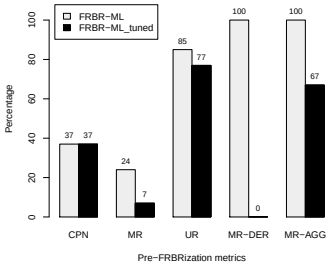


Fig. 4: Applying predictive metrics on BIB-RCAT for FRBR-ML basic rules and tuned rules

7.3 Facilitating the tuning

In this last experiment, we show how our pre-FRBRization metrics can help updating the set of rules. Only the FRBR-ML tool was used in this experiment, but the scenario could be applied to any tool. As shown in Table 1, the results of FRBR-ML for dataset BIB-RCAT could be improved. To provide insight to the expert, we compute predictive scores for the basic set of rules on the dataset BIB-RCAT. A subset of these scores is detailed in Figure 4. The white bar stands for the results with the basic set of rules. For instance, we note that 37% of the records contain cataloging practices (metric CPN). The basic set of rules contains many not used rules for the dataset BIB-RCAT (score of UR equal to 85%) and

it lacks 24% of rules to take into account all fields from the dataset BIB-RCAT. Finally, the metrics for specific patterns indicate that 100% of the rules are missing to tackle derivations (metric MR-DER) and aggregations (metric MR-AGG). Based on these predictive scores, an expert has enhanced the basic set of rules of FRBR-ML. This update took 4 hours mainly for correcting minor changes (e.g., add rules for missing subfields) and implementing new templates to handle relator codes and missing concepts (e.g., augmentations, parent works). The enhanced set of rules has been tested with the prediction metrics (black bars in Figure 4). Now, only 7% of the rules are missing to process all fields, and a few not used rules have been deleted (metric UR). The most significant enhancement deals with the pattern detection: all rules to identify derivations have been added, but the set still misses 67% of rules to process aggregations. Finally, FRBR-ML tuned with this enhanced set of rules was used to FRBRize the BIB-RCAT dataset. The results of this new FRBRization is shown in Table 1 (column *FRBR-ML tuned*). As expected, the quality of this enhanced FRBRization is better than with the basic set of rules, especially for the patterns. Adding relevant new rules enables us to reduce the amount of missing data, but 29% of relationships and 21% of secondary elements in the patterns are still missing. This experiment demonstrates how the predictive metrics help librarians update the set of rules and thus improve the quality of the FRBRization.

8 Conclusion

In this paper, we described BIB-R, the first benchmark for evaluating the interpretation of bibliographic records. It includes a set of metrics and two datasets (T42 and BIB-RCAT). Extensive experiments with our dataset T42 have been performed with three recent tools (FRBR-ML, Variations VFRBR and Extensible Catalog) to demonstrate the possibility to identify strengths and weaknesses. Our experimental validation is also the first to compare FRBRization tools with the same datasets and metrics. Finally, we showed how the pre-FRBRization metrics can be useful to help librarians update the set of rules. The release of this benchmark brings different perspectives. We plan to add more records in the real-world dataset BIB-RCAT. The main challenge is to update the FRBR expert collection. Next, we could enhance the benchmark to enable evaluation of ergonomics (quality of graphical user interfaces), performance (execution time) and quality of the semantic enrichment (for instance based on the Knowledge Base Population challenge[9]).

9 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).

[9] http://www.nist.gov/tac/2016/KBP/

References

1. Aalberg, T.: A Process and Tool for the Conversion of MARC Records to a Normalized FRBR Implementation. LNCS: Digital Libraries: Achievements, Challenges and Opportunities 4312, 283–292 (2006)
2. Aalberg, T., Žumer, M.: The Value of MARC Data, or, Challenges of FRBRisation. Journal of Documentation 69, 851–872 (2013)
3. Alemu, G., Stevens, B., Ross, P., Chandler, J.: Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. New Library World 113, 549–570 (2012)
4. Bowen, J.: Moving Library Metadata Toward Linked Data: Opportunities Provided by the eXtensible Catalog. International Conference on Dublin Core and Metadata Applications (2010)
5. Buchanan, G.: FRBR: Enriching and Integrating Digital Libraries. In: Proceedings of Joint Conference on Digital Libraries. pp. 260–269 (2006)
6. Coyle, K.: FRBR, Twenty Years On. Cataloging & Classification Quarterly pp. 1–21 (2014)
7. Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., Lumineau, N.: Appendix: BIB-R: a Benchmark for the Interpretation of Bibliographic Records. Tech. rep., LIRIS, NTNU (2016), http://liris.cnrs.fr/~fduchate/docs/appendix/appendix-tpdl16.pdf
8. Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., Lumineau, N.: Open Datasets for Evaluating the Interpretation of Bibliographic Records. In: Proceedings of Joint Conference on Digital Libraries. ACM (2016)
9. Decourselle, J., Duchateau, F., Lumineau, N.: A Survey of FRBRization Techniques. In: Theory and Practice of Digital Libraries. pp. 185–196 (2015), https://hal.archives-ouvertes.fr/hal-01198487
10. Denton, W.: FRBR and the History of Cataloging. Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools (2007)
11. Dickey, T.J.: FRBRization of a Library Catalog: Better Collocation of Records, Leading to Enhanced Search, Retrieval, and Display. Information Technology & Libraries 27, 23–32 (2008)
12. Hickey, T.B., O'Neill, E.T.: FRBRizing OCLC's WorldCat. Cataloging & Classification Quarterly 39, 239–251 (2005)
13. Manguinhas, H.M.A., Freire, N.M.A., Borbinha, J.L.B.: FRBRization of MARC Records in Multiple Catalogs. In: Hunter, J., Lagoze, C., Giles, C.L., Li, Y.F. (eds.) JCDL. pp. 225–234. ACM (2010)
14. Notess, M., Dunn, J.W., Hardesty, J.L.: Scherzo: A FRBR-Based Music Discovery System. In: International Conference on Dublin Core and Metadata Applications. pp. 182–183 (2011)
15. Riley, J.: Enhancing Interoperability of FRBR-Based Metadata. International Conference on Dublin Core and Metadata Applications (2010)
16. Riva, P.: Mapping MARC 21 Linking Entry Fields to FRBR and Tilletts Taxonomy of Bibliographic Relationships. Library resources & technical services 48(2), 130–143 (2013)
17. Riva, P., Žumer, M.: Introducing the FRBR Library Reference Model. IFLA Conferences (2015), http://library.ifla.org/1084/
18. Takhirov, N., Aalberg, T., Duchateau, F., Žumer, M.: FRBR-ML: A FRBR-based framework for semantic interoperability. Semantic Web Journal 3, 23–43 (2012)
19. Zhang, Y., Salaba, A.: Implementing FRBR in libraries: key issues and future directions. Neal-Schuman Publishers (2009)