Benchmarking and Evaluating the Interpretation of Bibliographic Records

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

International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Benchmarking and Evaluating the Interpretation of Bibliographic Records

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

Received: date / Accepted: date

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 many limitations. The conceptual reference model for bibliographic information presented in the Functional Requirements for Bibliographic Records (FRBR) is expected to be the foundation for a new generation of catalogs that will replace MARC and the digital card catalog. The need for transformation of legacy MARC records to FRBR representation (FRBRization) has led to the proposal of various tools and approaches. However, these projects and the results they achieve are difficult to compare due to lack of common datasets and well defined and appropriate metrics. Our contributions fill this gap by proposing BIB-R, the first public benchmark for the FRBRization process. It is composed of two datasets that enable the identification of the strengths and weaknesses of a FRBRization tool. It also defines a set of well defined metrics that evaluate the different steps of the FRBRization process. Those resources, as well as the results of a large experiment involving three FRBRization tools tested against our benchmark, are available to the community under an open licence.

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

Trond Aalberg
NTNU
Trondheim, Norway
Tel.: +47 735 979 52
E-mail: trondaal@idi.ntnu.no

Fabien Duchateau
LIRIS, UMR5205
Université Claude Bernard Lyon 1
Lyon, France
Tel.: +33 472 445 825
E-mail: fduchate@liris.cnrs.fr

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

Joffrey Decourselle
LIRIS, UMR5205
Université Claude Bernard Lyon 1
Lyon, France
E-mail: jdecours@liris.cnrs.fr

Nicolas Lumineau
LIRIS, UMR5205
Université Claude Bernard Lyon 1
Lyon, France
E-mail: nluminea@liris.cnrs.fr

1 Introduction

Memory institutions are responsible for offering access to the vast body of resources that constitute our collective memory and the creation of high quality metadata is a key service to fulfill this mission. In libraries worldwide, most metadata has been recorded using different implementations of the MARC format such as MARC 21 or UNIMARC. However, these formats have shown many limitations in terms of meeting the requirements of modern information management including interoperability, reuse, and information disambiguation [19]. The Functional Requirements for Bibliographic Records (FRBR) [13] and its updated version Library Reference Model (LRM) [41] are designed to provide a sound and more explicit semantics for bibliographic metadata and enable innovations and enhancements such as improved navigation and enrichment features [4,10,20]. However,

more than twenty years after the original specification of FRBR, the model is still not widely implemented in library systems [14]. A major obstacle is the migration of legacy catalogues, which consists of interpreting data found in existing records and representing it in formats based on the FRBR semantics (e.g., the semantic web vocabulary for the RDA data elements[1], LD4L ontology[2] and BIBFRAME[3] [28]). The main challenges of this complex process (a.k.a FRBRization) is to achieve best possible quality in the resulting catalogue, reduce cost and effort by automatic tuning and provide state-of-the-art interfaces to facilitate user validation.

In the last decades, the development of FRBRization tools has demonstrated many enhancements to improve the process and quality of results e.g. clustered deduplication and exploitation of added entries [18]. Unfortunately, many of these efforts are rarely reused since it is difficult to evaluate and compare the tools. First, experiments are hard to reproduce because implementations and datasets rarely are available to others. Catalogue excerpts may be provided, but they do not reflect all the variations and challenges found in real world catalogues and are only selected for illustrating specific cases [3]. Last but not least, metrics used in these projects are ad hoc and not intended to evaluate all possible aspects that a transformation ideally should consider. In closely-related fields, the existence of a benchmark led to numerous improvements in that field, for instance in information retrieval [6], ontology matching [22], schema mapping [5] or in entity matching [26]. Thus, we believe that the lack of a common FRBRization benchmark is an obstacle to the migration of legacy data and large scale adoption of FRBR implementations.

In this paper, which extends a previous version [16], we propose BIB-R[4], 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 completeness of a tool by testing all possible patterns and issues that ideally should be addressed by the FRBRization. The second dataset BIB-RCAT is extracted from the catalogues of libraries and can be used for comparing or experimenting with the data quality that is typically found in real world catalogues. The assessment of the process relies on a set of metrics which can be used at different stages. Before FRBRization, a user is interested in determining whether a tool requires tuning or not for a given catalogue, or predict the outcome and assess the feasibility of the process. During processing, performance may be a bottleneck and it is crucial to be able to estimate the execution time cost e.g. for processing very large collections. Finally, when the FRBR catalogue has been produced, the evaluation deals with the quality of the new catalogue. To summarize, the main contributions of the BIB-R benchmark are:

- The definition of new metrics that (i) estimate the quality of a FRBRization process, (ii) assess the efficiency of the process, and (iii) evaluate the quality of a FRBRized catalogue. The latter metrics are presented using a formal notation to be understandable by both librarians and computer scientists
- The release of two public datasets (open license CC-BY-NC[5]) for allowing a fair evaluation of FRBRization tools. The expected FRBR result is provided with each dataset
- An experimental study with three recent FRBRization solutions, showing the benefits of the proposed metrics and datasets

In the rest of the paper, we present the background in Section 2 and an overview of our benchmark BIB-R in Section 3. Evaluation metrics for pre-FRBRization, FRBRization and post-FRBRization are presented respectively in Sections 4, 5 and 6. The datasets are described in Section 7. The experimental study is detailed in Section 8. Section 9 concludes and outlines future work.

2 Background

The feasibility of transforming legacy data into a representation based on the FRBR model has been demonstrated in many projects such as WorldCat FictionFinder, Aust-Lit, OpenCat and `datos.bne.es` [23,27,24,29,45,46]. Consequently, the FRBRization process has been widely studied and improved during the last decade and more important, our understanding of the opportunities and challenges have matured. In this section, we present the patterns of bibliographic entities and relationships that need to be taken into account when migrating a typical library catalogue, and discuss specific challenges that complicate the process. Finally, we describe different FRBRization tools with a particular focus on how evaluation has been performed.

2.1 Bibliographic patterns and FRBRization issues

Although the FRBR model introduces a limited set of entities interrelated by basic relationships, the cardin-

[1] http://www.rdatoolkit.org/
[2] http://www.ld4l.org/ontology
[3] http://loc.gov/bibframe/
[4] http://bib-r.github.io/
[5] http://creativecommons.org/licenses/by-nc/2.0/

ality of the relationships – as well as the large number of optional relationships that may exist [40] – leads to a big variety of graph structures that may be extracted from bibliographic records [3]. Within this variety, we can generalize and identify a set of frequently occurring subgraphs that we refer to as *bibliographic patterns*. These patterns will also be the building blocks of more complex structures and include:

- Core pattern
- Augmentation pattern
- Derivation pattern
- Complementary pattern
- Aggregation pattern

Each pattern can be recorded in existing MARC records in various explicit or implicit ways, with an impact on the FRBRization [3,40,47].

The **Core** pattern forms the skeleton that most bibliographic records will include, with a single Work and a single Expression embodied in a Manifestation – and Person[6] as creator of the Work. Figure 1 depicts a schematic representation of this pattern with dotted lines indicating optional parts. Its FRBRization is relatively easy, unless the pattern is associated with data quality issues in catalogues. With a single Work and Expressions, the core structure is given by the model and the challenge is mainly to establish the proper identity of entities. The subject relationship from Work to Concept[7] is additionally included as part of the core pattern. Subject entries are important in bibliographic information, but in many cases this pattern is trivial to extract because of designated fields that clearly identifies the "has subject" relationship as well as the type of the Concept.

The **Augmentation** pattern is defined for cases where the record describes additional content related to an existing (main) Work, but the additional content is considered as subordinate and does not influence the main Work. Figure 2 illustrates the augmentation pattern and common examples are illustrations or forewords. When it comes to FRBRization, this is a somewhat optional pattern depending on the need for proper documentation of augmentations and their associated Persons as individual entities in the resulting catalogue. If this pattern is included, the proper structure should be that of a Manifestation embodying Expressions of both the augmentation and the main Work. A relationship between the two Works is needed to fully describe

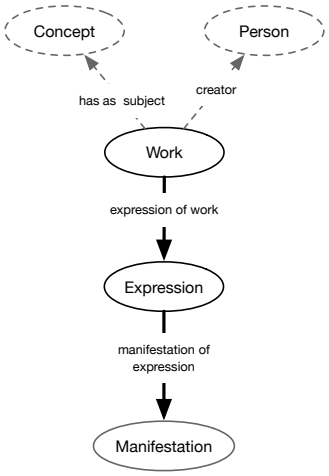


Figure 1: A schematic representation of the core pattern (including a related concept)

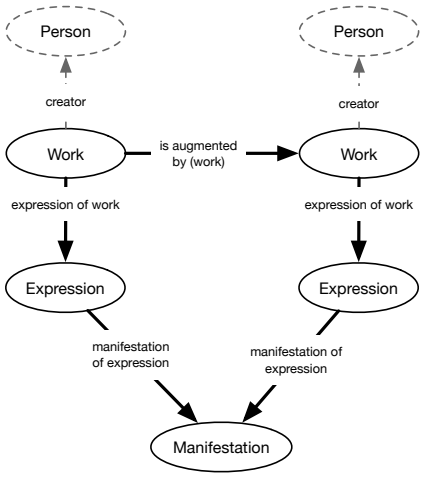


Figure 2: A schematic representation of the augmentation pattern (with a relationship between Works)

the semantics of the pattern. Additionally, agents responsible for the creation of the additional content will in most cases be created. Typical challenges for identifying this pattern from records is to detect whether the pattern is present and identifying the entities from textual descriptions or relator codes. Augmentations are rarely described with titles or other identifying information and need to be identified by the Work it augments and the creator of the augmentation.

The **Derivation** pattern means that one Work is the modification of another Work. This includes a variety of Work-to-Work relationship types where each Work typically has an established identity as a bibliographic entity, such as adaptations (e.g. the movie Lord of the Rings, based on the novel) and imitations (e.g. Bored of the Rings is a parody of Lord of the Rings). A schematic

[6] We use "Person" in our examples for the sake of readability. The initial FRBR model also includes a Corporate Body entity type. In the revised Library Reference Model the proper supertype is "Agent".

[7] "Concept" is used as a categorical supertype for anything that can be the subject.

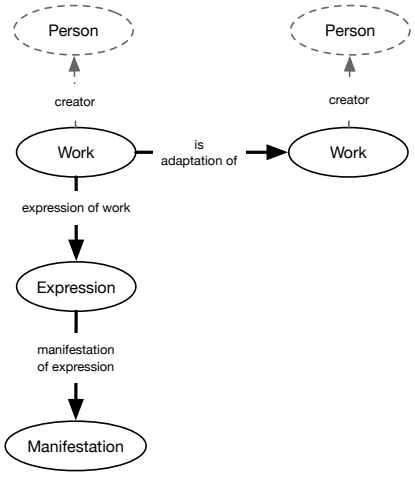


Figure 3: A schematic representation of the derivation pattern (as an adaptation)

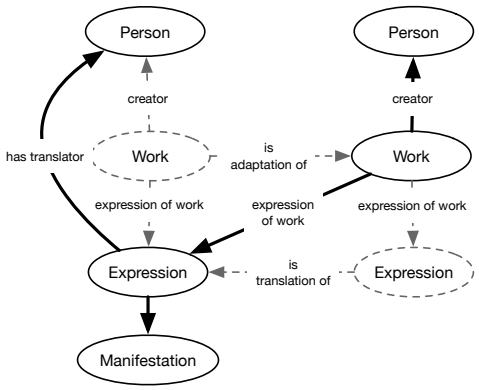


Figure 4: A schematic representation of the derivation pattern (as a translation)

representation of derivation as an adaptation is shown in Figure 3. The FRBRization of such patterns usually implies the creation of an additional related Work and its optionally associated Agent(s). Common challenges in finding this pattern in library records is the variety of practices used to describe the relationship and the source of the derivation. It may be described in textual notes or in added entries. The type of the relationship may be indicated or not using subfields for this purpose or designated fields which is the case for UNIMARC, and if missing it may be hard to distinguish derivations from other patterns. Translations, which represent the classical FRBR example, should formally be considered as a variety of the derivation pattern. A difference, however, is that the translation Work is less important for end users as an individual entity. Due to the frequent occurrence of this pattern, it is more convenient to use the semantically equivalent pattern illustrated in Figure 4 which also appears to be more in line with how end users view and understand a translation [37]. The same would be the case for other versions of a Work that do not have a strong individual identity as intellectual contribution e.g. abbreviations and narrated version. The pattern can be identified in different ways from the information in bibliographic records. It differs from the core pattern by the presence of a relationship to the translator which can be explicitly identified from e.g. relator codes or terms in notes and responsibility statements.

The notion of **Aggregation** is commonly described as a whole-part relationship. It includes three different categories of structures at the Work level. The first one is usually referred to as ensemble, and it is close to the *"parent Work"* concept, in which the parent Work is as important as its parts. The Lord of the Rings is an ensemble composed of three Works. The second category is entitled aggregating Work, a grouping of Works in which the aggregation itself usually is of lesser interest as a standalone Work. Novels or short stories of different authors may appear in the same published book, or different pieces of classical music may be on the same album. The main entities of interest as Works in these cases would be the individual parts, whereas the aggregation itself is primarily of interest as a Manifestation that embodies multiple expressions. The third category is related to various collections such as periodicals (e.g. journals, magazines), publication series and library collections (e.g. "science-fiction pocket book"). Figure 5 depicts an aggregation pattern. The FRBRization of aggregations mainly results in the creation of relationships between Works and their aggregating or parent Works. New Agents may also be created. The pattern can be difficult to discover in records because of the many ways it can be described: added entries indicated to be analytical entries, use of subtitles or part numbering, series entries, notes and other information.

The **Complementary Works** pattern aims at modelling a relationship between Works having the same importance (contrary to augmentations). Two main categories are covered by this pattern: sequels stand for an ordered sequence of Works (e.g. the Harry Potter series where individual novels are successors/predecessors to each other) and accompanying works which represent individual Works of different nature but with a strong dependency between (e.g. a manual of exercises and the associated book of solutions). The complementary Works pattern is illustrated in Figure 6. The FRBRization of complementary works mainly results in the creation of relationships between Works and the successful identification of this pattern relies on the proper iden-

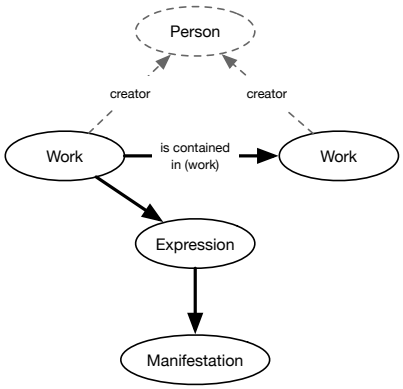


Figure 5: A schematic representation of the aggregation pattern (the whole and its parts having the same creator)

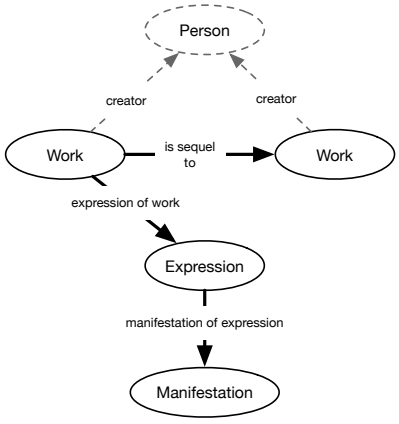


Figure 6: A schematic representation of the complementary works pattern (as a sequel having the same author)

tification of the related work and the relationship.

The challenge of interpreting library records to discover and extract bibliographic patterns increases significantly when combined with data quality issues found in many library catalogues. Although the MARC formats are not intended for recording structured information at the level that FBRR requires, there are many elements that if used would increase the quality in terms of explicit representation of entities and relationships including systematic use of original titles and indicators to flag the semantics of an entry [2]. A specific set of requirements for FRBRization were identified in the TelPlus project [32]. Records that do not satisfy these requirements need to be automatically enriched, cleaned or manually corrected. Six of these requirements can be perceived as major issues for successful FRBRization. Note that the most extreme issues, such as *missing proper title*, are rarely encountered in real-world catalogues of reasonable quality, and thus have been discarded. The **missing publication date** and **missing record identifier** issues deal with a lack of information about the publication. These data are not crucial when FRBRizing, but they can help disambiguating at the Expression and Manifestation levels. For instance, without an ISBN number, the publication date is useful to retrieve the correct edition. The **missing uniform title** and **missing original title** issues are related to the identity of works, and mainly have an impact on the derivation and aggregation patterns respectively. If such titles are missing, false positive Works and relationships will be created. When a record has a **missing relator code** issue, it means that an agent is associated to an entity, but his or her role is not specified (e.g. translator, drawer). This makes it difficult to select the correct type of relationship in FRBR, and can result in an incorrect or weaker semantics. Finally records can have the **missing authoritative responsibility** issue. In this case, agents (and consequently their roles) cannot be clearly identified causing duplicate entities for the same person to be created, which again may introduce duplicate works and add other noise to the data.

The presented set of patterns is intended as a complete list of cases for analyzing difficulties and issues related to FRBRization, and is based on results in previous work [3,40]. The patterns reflect a certain level of semantic abstraction or grouping of cases. New patterns can be created by deriving from these patterns or by combining them.

2.2 Tools

Grouping-record tools were the first techniques used to perform FRBRization by grouping records (seen at the Manifestation level) to deduce entities at a higher level of abstraction (i.e., the Expression and the associated Work). Among these tools are the **Work-Set algorithm** [25] and PRIMO [38]. However, grouping-records tools cannot be used for extracting complex structures [3,24] and in the last decade rule-based tools capable of creating FRBR have emerged such as LC Display Tool [43], VFRBR [39], Extensible Catalog [9], TelPlus [32], FRBR-ML [44], LibFRBR [11], X3ML [33] or Cervantes [42]. Due to the large number of FRBRization tools and projects, we refer to a recent survey for an exhaustive list [18], and focus on three rule-based tools that are publicly available for experiments. The first one, VFRBR, developed in the context of the Variations project, aims at FRBRizing catalogues with a

focus on the music domain [39,35]. Since music cataloguing makes frequent use of added entries to record information, VFRBR's strategy is to interpret these added entries as separate entities. Rules in VFRBR are hardcoded in Java, but a description of the rules is available online in textual documents[8]. The online catalogue Sherzo is the proof of concept that lets users explore musical works, composers and related entities extracted from 185,000 MARC records. Extensible Catalog (XC) is an open-source project for a complete Integrated Library System, which includes the FRBRization module Metadata Service Toolkit [9]. The latter is based on the XC Schema, which extends the FRBR model (e.g. with the MARC 21 Holding concept). XC exploits added entries to extract advanced relationships (but limited to those of the first FRBR group). Rules in XC are hardcoded in Java, which means that they are difficult to update for non-IT people. The third tool is FRBR-ML [44], which is based on Aalberg's approach [1]. The rules are written in an external XML file. The authors also discuss the possible techniques for structuring the FRBR output catalogue. Finally, the tool provides enhancements to disambiguate some complex cases by exploiting other catalogues or Linked Open Data knowledge bases. An excerpt of rules from FRBR-ML is shown in Figure 7 related to the core pattern. This also serves as a generic example of what information that typically makes up a rule for interpreting FRBR from MARC.

```
<entity type="c:Person" templatename="MARC21-100-Person">
  <anchor tag="100"/>
  <attributes>
    <datafield tag="100">
      <subfield code="a" type="a:nameOfPerson"/>
      <subfield code="c" type="a:titleOfPerson"/>
      <subfield code="d" type="a:dateAssociatedWithPerson"/>
      <subfield code="u" type="a:affiliation"/>
    </datafield>
  </attributes>
  <key order="1">
    <element>*:datafield/*:subfield[@*:type= 'a:nameOfPerson']</element>
    <element>*:datafield/*:subfield[@*:type= 'a:titleOfPerson']</element>
    <element>*:datafield/*:subfield[@*:type= 'a:dateAssociatedWithPerson']</element>
  </key>
  <relationships>
    <relationship condition="*:subfield[@code = '4'] eq 'aut')"
          reltype="a:authorOf"
          inverse="w:author">
      <target entity="MARC21-240-Work"/>
    </relationship>
  </relationships>
</entity>

<entity type="c:Work" templatename="MARC21-240-Work">
  <anchor tag="240"/>
  <attributes>
    <datafield tag="240">
      <subfield code="a" type="w:titleOfWork"/>
      <subfield code="d" type="w:dateOfWork"/>
      <subfield code="f" type="w:dateOfWork"/>
      <subfield code="k" type="w:formOfWork"/>
      <subfield code="m" type="w:mediumOfPerformance"/>
      <subfield code="n" type="w:numberingOfPart"/>
      <subfield code="o" type="w:otherDistinguishingCharacteristics»/>
      <subfield code="p" type="w:numberingOfPart"/>
      <subfield code="r" type="w:key"/>
    </datafield>
    <datafield tag="380">
      <subfield code="a" type="w:formOfWork"/>
    </datafield>
  </attributes>
  <key order="2">
    <element>*:relationship[@type = 'aut'])/@href</element>
    <element>*:datafield/*:subfield[@*:type = 'w:titleOfWork']</element>
    <element>*:datafield/*:subfield[@*:type = 'w:dateOfWork']</element>
    <element>*:datafield/*:subfield[@*:type = 'w:formOfWork']</element>
    <element>*:datafield/*:subfield[@*:type = 'w:numberingOfPart']</element>
    <element>*:datafield/*:subfield[@*:type = 'w:mediumOfPerformance']</element>
    <element>*:datafield/*:subfield[@*:type = 'w:key']</element>
    <element>*:datafield/*:subfield[@*:type = 'w:otherDistinguishingCharacteristics']</element>
  </key>
  <relationships/>
</entity>
```

Figure 7: Excerpt of rules from FRBR-ML

2.3 Evaluation

The FRBRization process has traditionally been evaluated based on its output (the produced FRBR catalogue). This evaluation is possible in various forms: the most frequent option requires a ground truth or gold standard, i.e., an expert FRBRized catalogue as explored in the TelPlus project [32]. The comparison between the expert FRBRized catalogue and the automatically produced catalogue indicates whether the tool is able to perform an acceptable transformation. A main issue with this approach is the manual construction of the expert catalogue. The FRBR-ML approach avoids the demanding construction of a gold standard by converting the FRBRized catalogue back to a MARC catalogue [44]. The evaluation is performed between the initial MARC catalogue 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 catalogue. Besides, the rules for the reverse transformation have to be created as well. To evaluate a process, metrics are required. In the TelPlus project, an aggregation metric was 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. To the best of our knowledge, the result of the FRBRization, at the end of the process, is the only aspect that so far has been assessed, and other steps are currently not evaluated (for instance, tuning and performance, see Section 3).

2.4 Discussion

The bibliographic patterns that frequently occur in bibliographic descriptions have been well identified, and many of the challenges for successful FRBRization of catalogues caused by format limitations and cataloguing

[8] Variations VFRBR rules available at http://www.dlib.indiana.edu/projects/vfrbr/projectDoc/

practices are well documented. However, most tools have an ad-hoc view on what patterns to identify and what challenges to address. Furthermore, tools are only tested against private datasets, whose characteristics are not clearly defined. Available metrics either assess the deduplication quality or compare results from a roundtrip conversion. In particular, there is no well defined set of metrics for evaluating the quality of a generated FRBR collection in terms of coverage of bibliographic patterns. Other aspects such as support for adapting and tuning the tools have never been evaluated. Contrary to other research domains such as record linkage [26], ontology matching [22] or information retrieval [6], there is no benchmark for one of the most crucial challenges in the management of bibliographic information. Additionally, understanding the weak and strong points of the FRBRization process tends to promote novelty and enhancements in the future implementations and common datasets and evaluation metrics enable a fair comparison between tools.

3 Overview of the benchmark

Figure 8 depicts an overview of the FRBRization process [1,18], which typically is composed of the three main steps pre-FRBRization, FRBRization and post-FRBRization. Pre-FRBRization includes preparing the input catalogue and tuning the tool. This step is optional, but allows for correcting errors or inconsistencies in the catalogue that can be detected in advance and corrected automatically or manually. Examples include correcting inconsistent use of relator codes, removing empty records, etc. The tuning task is about adapting the process to the format and cataloguing practice by editing rules or adapting the implementation of the tool. Additional tasks could be to configure parameters such as setting decision threshold for deduplication. Next, the FRBRization starts using a prepared catalogue and a customized set of rules. The transformation of each record produces a set of entities and relationships according to the rules applied. The 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 produced FRBR collection [18]. We only mention validation and enrichment. The former enables expert to verify and correct the generated FRBR catalogue while the latter refers to the task of adding information from external sources.

Most approaches only evaluate the output from the core FRBRization, but evaluations made in the initial step are equally important and will have a strong impact on the final result in terms of quality and performance. Our benchmark BIB-R provides metrics and datasets for evaluating all aspects of the FRBRization process. It focuses on the initial FRBR model and its revised version (LRM), and does not take into account the specificities and the complexity of alternative implementations of the model such as FRBRoo.

4 Pre-FRBRization evaluation

Prior to performing the actual processing of records, there is a need to gain insight into the characteristics of the catalogue. Data about the information in the records is needed to prepare the catalogue for processing and to develop or adapt the set of rules that will govern the interpretation. Records in a catalogue may describe a range of bibliographic patterns and records may have been created over many years with changing cataloguing practice. Another motivation for analyzing the catalogue is to determine characteristics that inherently will have an impact on the output. Catalogues may be based on different implementations of the MARC specification. MARC 21 and UNIMARC differ significantly in the use of fields which necessarily leads to differences in what can be extracted. Catalogues may represent different selections of resources and for this reason follow different cataloguing rules and practices. Such aspects will influence the FRBRization. Some issues may be resolved when setting up the transformation, others are inherent to each catalogue and will cause major differences in what can be extracted from different catalogues.

Our pre-FRBRization metrics include measures to describe the characteristics of a catalogue in terms of opportunities for extracting specific patterns, as well as threats that may lead to erroneous results. Such metrics can be calculated as a percentage by using the number of records concerned divided by the total number of records. These measures do not necessarily relate to the actual processing or quality of the tools, but gives valuable context for interpreting the final result and represent a methodology for comparing the results. These metrics also enable experts to give a priority to the most important rules that need to be written (e.g. rules for interpreting translations are crucial when one knows that the catalogue contains a large number of records with a translation). Additionally, we define a set of measures for describing the correspondence between the catalogue and the set of rules applied in the processing. Conditions for rules may e.g. be based on indicators or relator codes and the completeness of transformation depends on a rule set that covers all possible

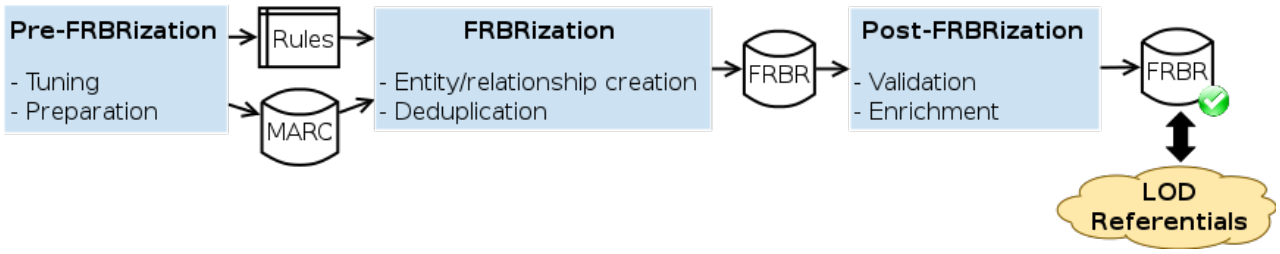


Figure 8: Overview of the FRBRization process

conditions. Measures related to the rules are mainly relevant for comparing different rule sets developed for the same tool, since what constitutes a rule will be different between tools.

Statistics about the usage of specific fields and field values can be used to characterize the potential for identifying and extracting specific bibliographic patterns from a catalogue. The core pattern is of course the most frequent, and since we can assume that it is found in all records there is no need to have a metric for this pattern. Measures for other patterns can be computed by counting records with fields or values that are known to reflect a specific pattern. The **metric AUG** describes the percentage of records having an augmentation pattern. The most reliable indication of augmentations will be the usage of specific relator codes in both MARC 21 and UNIMARC, but the occurrence of specific terms in the statement of responsibility or notes may be used as indications as well. The **metric DER** stands for the derivation pattern where adaptations typically are identified by specific fields in both UNIMARC and MARC 21 as well as in added entries in MARC 21 which do not have an indicator for analytical entry. Other conditions have to be used for the second category of derivations, such as translator relator codes indicating a translation. The **metric AGG** deals with the aggregation pattern (i.e., a whole-parts relationship). It measures the percentage of records exhibiting the aggregation pattern. This can be identified using a number of indications such as the use of subtitles and numbering of parts, series entries or specific whole-part linking fields such as in UNIMARC. At last, the **metric COW** computes the percentage of records containing a complementary work, i.e., a relationship between Works of the same importance. Again, this can be detected by the presence of specific fields in both MARC 21 and UNIMARC as well as by specific terms or phrases in note fields. Numbers calculated for these metrics will not be the same as we will get by counting the actual occurrence of the pattern in the final FRBRized result due to the deduplication process that is performed.

In addition to identifying possible occurence of various bibliographic patterns, we can express the threats to FRBRization quality by using metrics for data quality issues. The TelPlus project established six requirements for FRBRization [32], that in this context can be seen as errors in the initial records. According to this, the following metrics can be defined: **MID**, **MPD**, **MUT**, **MOT**, **MRC** and **MAR** which respectively compute the percentage 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 cataloguing 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*. The former occurs when the link to the authoritative title is incorrect while the latter stands for a broken authoritative link to an Agent (only used in UNIMARC, for instance fields 700$3). 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). Inconsistent practices, within or across catalogues, makes it more difficult to FRBRize, and human intervention is usually required to indicate how to interpret such fields. Once the practice is known, writing rules to extract information from these ”coded values” is rather simple and beneficial to the FRBRization of the record. The **metric CPN** deals with these inconsistent *cataloguing practices and norms*, which otherwise would represent useful information.

The specific rules that are implemented to govern a FRBRization process will have a crucial impact on the results and the systematic analysis of the rules is a

method that can be exploited to develop and improve upon the rules. The following metrics are intended to provide feedback in the development process by measuring and describing the correspondence between the rule set and the data in the records that the rules will be applied on. The **metric MR** is for missing rules, i.e., a field or condition that is not processed by a rule, thus causing loss of information. Missing rules needs to be detected systematically to inform towards completeness of the processing. Measures can be calculated as percentage of records or fields not processed and can be related to specific patterns and their metrics such as **MR-AUG** to calculate possible augmentations not covered by the rules. In the same fashion, we define the metrics **MR-DER** for derivations, **MR-AGG** for aggregations, **MR-COW** for complementary works, and **MR-CPN** for cataloguing practices. The **metric UR** identifies rules that are not used for a given FRBRization of a specific catalogue. Essentially of interest for rule sets which are intended to be reused across projects and will indicate reuse characteristics and determine what can be removed if e.g. the number of rules will have an impact on performance. The **metric CR** is for *conflicting rules*. This issue may occur if the set of rules has been developed over time, because the rule definitions are too complex to manage manually, or because rules that successfully have been applied on one collection turns out to conflict with other rules when applied on another catalogue. Ideally this is a situation that should not occur and the metric is intended more as a final checkpoint.

Tools have different characteristics in terms of the manual effort that is needed to install and adapt the tool, create or tune the processing rules, as well as other tasks that require manual intervention. The **metric TAC** stands for *tool application cost* and it aims at evaluating the cost for preparing the tool. Adapting an existing FRBRization tool to a specific collection can be defined as a configuration task, a programming task, or a combination of these. The best measure for this metric is to use man hours because of the variations in how and what manual work that needs to be performed. For tools where all processing rules are hard coded into the tool, such as VFRBR which is implemented in Java, the adaptation of the tool to a different collection is an implementation task that requires programming skills. For tools where processing rules are parametrized and described separately, such as in FRBR-ML where the rules are coded separately in an XML file, the task is better described as configuration and skills required to edit and create rules will depend on the format of the rule specification. A lesson learned in various projects is that setting up a FRBRization process is a highly iterative process consisting of implementing rules, running test, inspecting the results and editing or adding rules. The previously described metrics (e.g. AUG, MID, MR) not only provides statistics on the catalogue and the set of rules, but they aim at reducing the TAC score with simple identification of the problems inherent to the inputs. Alternative measures that can be considered e.g. for doing a more detailed study on the cost of adapting or working with a tool, is to use measures from software engineering [8] such as lines of code (LOC) for implementation or configurations that resembles implementations. If the configuration is performed in a graphical user interface, it is possible to explore user actions as a measure [34].

All predictive metrics are summarized in Table 1. The specifications of each metric (i.e., input, output and overview of the algorithm) are available online[9].

Metric	Related pattern/issue
AUG	Augmentation
DER	Derivation
AGG	Aggregation
COW	Complementary works
MID	Missing record identifier
MPD	Missing publication date
MTF	Missing type and form
TLE	Title linkage error
MUT	Missing uniform title
MOT	Missing original title
RLE	Resp. linking error
MRC	Missing relator code
MAR	Missing auth. resp.
CPN	Cataloguing practices and norms
MR	Missing rules
UR	Unused rules
CR	Conflicting rules
TAC	Tool application cost

Table 1: List of pre-FRBRization metrics

5 Evaluating FRBRization

The core task in any FRBRization project is the use of a software tool to interpret a collection of records and produce output in a format that implements the FRBR model. However, tools have different characteristics in terms of efficiency when processing records. Most implementations of the FRBRization process report on two separate subtasks in the automatic processing of

[9] http://bib-r.github.io/specifications-metrics.txt

records. The first is to extract entities and relationships from each of the records. The second is to perform deduplication of entities – sometimes referred to as the merging step. This two-step process is also found in grouping tools, where the first step is to extract grouping keys from each of the records and the second step is to deduplicate and merge keys. These two steps are best evaluated separately, if possible, because they are likely to have different time complexity. Otherwise, a combined metric can be applied. For both metrics we are mainly interested in how an increase in the number of records will affect the processing time, because this indicates the scalability of the tool and is a measure that can be used to compare different tools. The overall processing time for a given set of records is also relevant, but primarily indicates whether a tool is able to perform the processing within reasonable time.

Execution time cost of the entity/relationship extraction (ETC): applying a set of rules to a collection of records in a sequential way is typically a process with theoretical time complexity of N, meaning that the processing time will increase in a linear way as the number of records increases[10]. Each processed record adds more or less the same amount of time to the overall processing time. Different tools, and even different rule sets, will show differences in the average processing time per record, but this is of marginal interest as long as the processing time for the collection is within a reasonable limit. The purpose of using this metric is to document the performance characteristics of the extraction part of the process. The measure should be a graph showing how the processing time increases as the size of the collection increases, which allows for comparing tools. Additionally, the total processing time of a certain number of records can be specified. Time complexity of N is the assumed performance, but better performance can e.g. be achieved if the tool uses parallel programming or distributed processing [30].

Execution time for deduplication (ETD): deduplication of entities is the second subprocess of the FRBRization process. This is potentially an expensive processing task because each entity has to be compared with all other entities of the same type to find equivalent entities and perform the merging [12,21]. Theoretical worst case scenario is N^2 which means that the processing time increases with quadratic time as the number of records increase. Also this is a characteristic that is best described using a graph showing how the processing time increases as the size of the collection increases. The problem of deduplication cost has been addressed in different FRBRization research works, but solutions are not reported in ways that allows for comparing the tools. The TelPlus tool provides an optimization (a clustered deduplication) to reduce the execution time [32]. The descriptive keys from OCLC are useful to group records and improve the detection of duplicates [24]. Also developers of XC have discussed the need for improving performance of the FRBRization[11] with respect to execution time of the deduplication.

Metric	Definition
ETC	Execution time for creating entities and relationships
ETD	Execution time for deduplication

Table 2: List of FRBRization metrics

Table 2 summarizes the list of FRBRization metrics for computing the execution performance of tools. Characteristics should be reported using plots showing how the processing time increases as the number of records grow. The actual timing for a given number of records will be different between tools but major difference will be easy to spot when comparing the graphs. The suggested metrics are intended to support the comparison of tools, but will also be relevant as documentation for a specific tool. The specifications of each metric (i.e., input, output and overview of the algorithm) are available online[9].

6 Post-FRBRization evaluation

Post-FRBRization evaluation is the qualitative assessment of the output from the FRBRization process. As a context for our metrics, we have chosen the evaluation method which is based on an expert FRBRized catalogue which serves as a gold standard data set for the result. This is a well known methodology in related fields and also considered to be a most reliable form of evaluation [31]. It can be used to directly assess the quality in terms of conformance to the FRBR model and supports any kind of structure of entities and relationships. For the metrics in this section, we use a formal notation to establish a precise description of the measures.

We have identified nine metrics to compare the FRBRized catalogue or collection $\mathcal{T}$ and the expert FRBR catalogue $\mathcal{E}$. These metrics are useful to understand

[10] Note that the extraction can have a higher complexity in specific cases, such as when records contain references to other record(s) which needs to be looked up during the extraction.

[11] Presentation at code4lib 2011 about improving the performance of eXtensible Catalog's deduplication module, http://www.extensiblecatalog.org/learnmore/publications

the weak and strong points of a tool and the rules that it applied in the FRBRization process. The calculated values will provide information about what rules should be added to increase the quality of the transformation, and can be used to estimate the manual effort needed to improve the results. For all the post-FRBRization metrics, a zero value means that the tool successfully manages to solve a task or issues. Conversely, a 100% score to one of these metrics indicates that a tool completely fails for handling an issue. The specifications of each metric (i.e., input, output and overview of the algorithm) are available online[9].

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 catalogue $\mathcal{E}$ is missing in the catalogue $\mathcal{T}$ which is produced by the tool. 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 linking (e.g. to referential authorities or to the Linked Open Data). Either the link does not exist in $\mathcal{E}$ or it has a different value in $\mathcal{T}$ for the same external target. The metric calculates false discovery rate [7], 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 catalogues (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 $\mathcal{T}$. For these four metrics, a score equal to 0 indicates that the tool has perfectly handled all the data while error rate equal to 100% means a complete failure.

The second set of metrics deals with patterns. A pattern is a complex structure of minimum two entities with a relationship in between, but many patterns are more complex and include additional entities and relationships. Most patterns will have a main structure such as the Work and the relationship that makes up the main part of a Derivation. A complete pattern implies the main part as well as entities and relationships that can be characterized as secondary, such as Persons associated with the Works in an instance of the Derivation pattern. Only part of a pattern may be incorrect and the evaluation should reflect this. The **metric FPND** (full pattern not detected) measures the discovery of complete patterns. More precisely, it means that all elements of a pattern need to be included and structured correctly in the result (i.e., the main entity, the main relationship, and all secondary elements of the pattern). For instance, a derivation pattern (for a translation) is fully detected when the new Expression, its relationship to the Work and possible translator entities and translation relationships are found. This metric is obviously very strict, and although it provides an overview about the capabilities of a tool, it is not sufficient to understand the reason of a failure. Thus the next metrics are more specific with regards to the discovery of a pattern. 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 (of a pattern) that are not present among all main entities. 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 a *"is a translation"* relationship. The metric MRND computes the percentage of main relationships (of a pattern) that have been incorrectly detected among all main relationships. 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, the *translation* relationship) are missing or incorrect. The metric ESE computes the percentage of incorrect secondary elements among all secondary elements. For these pattern-related metrics, a score equal to 0 means a total success while an error rate equal to 100% stands for a failure.

Table 3 provides first-order logic notations for the post-FRBRization metrics. During evaluation, we compare two collections, $\mathcal{T}$ which is produced by a tool and the expert collection $\mathcal{E}$. This comparison depends on the type of data. Consider the data $e \in \mathcal{E}$ and $t \in \mathcal{T}$. When dealing with entities, the type of entity and the value of its main label (e.g. title, name) needs to be verified:

$$e \equiv t \iff type_e = type_t \wedge value_e = value_t$$

For relationships, the checking is performed based on the type of relationship and the two linked entities:

$$e \equiv t \iff type_e = type_t \wedge entity_e^1 = entity_t^1 \wedge entity_e^2 = entity_t^2$$

Finally, the properties are compared according to their type, their owner (entity) and their value:

$$e \equiv t \iff type_e = type_t \wedge entity_e = entity_t \wedge value_e = value_t$$

To support metrics related to parts of patterns we need annotations in the expert collection[12]. Thus we define the set $\mathcal{E}' \in \mathcal{E}$ which includes all main elements of a pattern (i.e., the main entity and the main relationship) and the set $\mathcal{E}'' \in \mathcal{E}$ which contains all secondary elements of a pattern. In Table 3, the formal notation for missing data (MD) states that a datum e appears in the expert collection $\mathcal{E}$ but has no equivalence in the created collection $\mathcal{T}$. As described above, it can be redefined for each type of data. An incorrectly added data (IAD) is defined as a datum t in the created collection $\mathcal{T}$ which has no equivalence in the expert collection $\mathcal{E}$. Similarly to MD, the metric IAD can be redefined according to the data type. A data linkage error (DLE) is a property t referring to an external link for which an equivalent property exists in $\mathcal{E}$ for the same data source or knowledge base (e.g. VIAF), but their values are different (e.g. on VIAF value *76382712* for Terry Pratchett). Semantic mismatch data (SMD) are defined as data with a different semantics (usually a subsumption noted $t \subset e$) in both collections. A pattern is considered as not completely detected (FPND) when either one of its main elements in $\mathcal{E}'$ (entity or relationship) or any secondary elements in $\mathcal{E}''$ are not found in the collection $\mathcal{T}$. A main entity of a pattern e' is not detected (MEND) when it does not have a corresponding entity t in the created collection $\mathcal{T}$. Note that the metrics MRND and ESE have a similar definition as for MEND due to our generic notation.

Lastly, it is possible to define a balanced metric that combines the five main quality issues related to FRBRization (namely missing data, incorrectly added data, linkage error, semantic mismatch and pattern detection). The **metric OQF** stands for Overall Quality of FRBRization and it is defined as:

$$\frac{\alpha_1 MD + \alpha_2 IAD + \alpha_3 DLE + \alpha_4 SMD + \alpha_5 FPND}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5}$$

where α_1, α_2, α_3, α_4 and α_5 are weights whose sum equals 1. Each weight indicates the importance of its associated quality issue for the overall quality of the transformation. A weight can be set to 0 when a quality issue should not be considered in the evaluation. The OQF metric computes scores in the range [0, 1], where a minimum score indicates that the FRBRization is as expected (without errors). Although such global metric is useful for comparing tools or alternative transformations, it is very limited for deeper analysis of the results compared to individual metrics.

[12] Our expert collections include specific annotations for each element of the patterns, else it would not be possible to compute the metrics MEND, MRND and ESE.

7 Datasets

Similarly to existing benchmarks in related disciplines, we provide within BIB-R two datasets that allow the assessment of FRBRization tools. In our context, a **dataset** is a set of collections. Each **collection** contains a set of MARCXML records, available as both MARC 21 and UNIMARC, and has an associated expert FRBRized version of the collection. The expert collections have been manually created and verified by a librarian and three digital library researchers. All records included in the collections are produced from original MARC records. The expert FRBR collections are in RDF using the RDA linked data and Semantic Web representations vocabulary[13], which takes into account the new concepts from LRM [36] and has good support for the data elements found in MARC records.

The records have been extracted from real-world catalogues, and modified when needed. To evaluate a tool, the procedure is to run it with an input collection in the MARC format the tool supports and to compare the produced FRBRized collection against our expert result collection using the post-FRBRization metrics (see Section 6). The first dataset is called **T42** and is intended for testing specific cases. The second dataset named **BIB-RCAT**[14] simulates a real-world catalogue. The rest of this section describes these two datasets, which are detailed in a report [17] and available with an open license CC-BY-NC at `http://bib-r.github.io/`.

7.1 Dataset T42

The objective of the dataset T42 is to evaluate whether a FRBRization tool is able to handle the set of patterns that have been defined, including the issues in the data that typically may complicate the interpretation process. We define a **test** as the combination of a pattern and an optional 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

[13] `http://www.rdaregistry.info/`

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

Metric	Related issue	Formal notation
MD	Missing data	$e \in \mathcal{E}, \forall t \in \mathcal{T}, t \not\equiv e$
IAD	Incorrectly added data	$t \in \mathcal{T}, \forall e \in \mathcal{E}, e \not\equiv t$
DLE	Data linking error	$t \in \mathcal{T}, t \rightsquigarrow$ 'external link' $\wedge$ $(\exists e \in \mathcal{E} \wedge$ $e \rightsquigarrow$ 'external link' $\wedge\ value_t \neq value_e \wedge$ $source_t = source_e)$
SMD	Semantic mismatch data	$e \in \mathcal{E}, t \in \mathcal{T}, (t \subset e) \vee (e \subset t)$
FPND	Full pattern not detected	$e' \in \mathcal{E}', e'' \in \mathcal{E}'', \forall t \in \mathcal{T}, \forall t' \in \mathcal{T}, t \not\equiv e' \vee t' \not\equiv e''$
MEND	Main entity not detected	$e' \in \mathcal{E}', \forall t \in \mathcal{T}, t \not\equiv e'$
MRND	Main relationship not detected	$e' \in \mathcal{E}', \forall t \in \mathcal{T}, t \not\equiv e'$
ESE	Error(s) in secondary elements	$e'' \in \mathcal{E}'', \forall t \in \mathcal{T}, t \not\equiv e''$
OQF	Global/Combined/Overall Quality of FRBRization	$\frac{\alpha MD+\beta IAD+\gamma SMD+\delta FPND}{\alpha+\beta+\gamma+\delta}$

Table 3: Formal notation of post-FRBRization metrics

deals with specific missing information. For a few tests, a correct solution can only be found when the tool performs a cleaning or preprocessing step before running the FRBRization (e.g. to remove specific characters or to create keys between records for facilitating the discovery of a bibliographic pattern). The dataset contains 42 tests which are crucial for testing specific aspects of FRBRization. Table 4 provides the complete list of the tests with their main features. For instance the *test 1.0* contains records with the core pattern and without any 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. Table 5 provides overall statistics for the dataset T42 (second column). For example, this dataset includes records in three languages (English, French, German), eight media types (e.g., books, movies, articles, audio) and there is an average of ten records per test. All tests can be downloaded at the benchmark URL and the records included in the test can be visualized in a graph, as illustrated by Figure 9.

7.2 Dataset BIB-RCAT

The BIB-RCAT dataset simulates a real-world catalogue in which various bibliographic patterns and issues may be found. It contains records in MARC 21 and UNIMARC, and an expert FRBR collection for these records. It is composed of records from various existing library catalogues (e.g. a public French library, a Swiss hospital library). The size of this catalogue (560 records) is smaller than ordinary catalogues found in libraries since the expert collection requires a time-demanding effort to be manually produced and verified, but it can be extended over time. Table 5 provides global statistics for the dataset BIB-RCAT (third column). The expert collection contains 1922 entities with more than 9500 property values.

7.3 Discussion

The use of the benchmarking datasets is intended to be straightforward, but a few challenges needs to be addressed. First, we assume that a tool produces output in RDF - or some format that can be transformed directly to RDF without loss of the structure and semantics created by the tool. Different vocabularies may be used for FRBR in RDF, and even when using the RDA vocabulary there will be differences in the types used in the data that is created by the tool and the data in the expert collection (because of the intricate subtype hierarchy as well as the canonical and lexical type system, for instance *P30011* and *dateOfPublication* in RDA). To tackle this issue, a mapping file[15] has been created to group equivalent types from various FRBR vocabularies. If a tool makes use of an unsupported type vocabulary, the mapping file needs to be updated accordingly.

Implementations which simplify the FRBR specification (such as LD4L[2], in which the Expression level does not exist) are obviously penalized for some metrics (for instance Missing Data). Specific enhancements offered by the tools (e.g. enrichment of entities) may provide correct content but lead to a decrease in terms of quality. Indeed, properties and relationships that are not recognized in the gold standard expert collection are considered as incorrectly added data.

[15] http://bib-r.github.io/mappings.xml

	Core (1.x)	Augmentation (2.x)	Derivation (3.x)	Aggregation (4.x)	Complementary work (5.x)
Basic	1.0	2.0	3.0	4.0	5.0
Missing publication date	1.1	2.1	3.1	4.1	5.1
Missing record ID	1.2	2.2	3.2	4.2	5.2
Missing type or form of material	1.3	2.3	3.3	4.3	5.3
Title linkage error	1.4	-	-	-	-
Missing uniform title	1.5	2.5	-	4.5	5.5
Missing original title	-	-	3.6	-	-
Responsibility linkage error	1.7	-	-	-	-
Missing relator code	1.8	2.8	3.8	4.8	5.8
Missing authoritative responsibility	1.9	2.9	3.9	4.9	5.9
Use of cataloguing practices	1.10	2.10	3.10	4.10	5.10

Table 4: List of main features of each test from T42

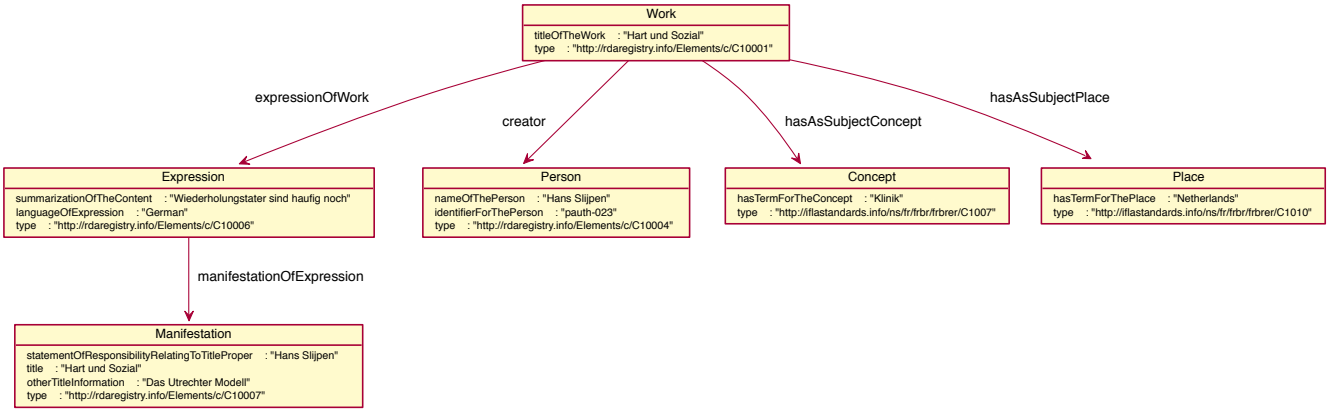


Figure 9: Extract from the vizualisation display for the records included in the *core* category (only one incomplete record shown)

Feature	T42	BIB-RCAT
Number of 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 5: Statistics for datasets T42 and BIB-RCAT

Some bibliographic patterns in the test set may appear controversial, e.g. due to disagreements between communities and different interests. For instance, a journal regularly publishes issues. A first possible interpretation is to consider the journal as a Work and each issue as Manifestations linked to this Work. Another interpretation may give more importance to issues by representing them as individual Work. In that case, the journal can be seen as a "super-Work" on the top of these issues. The BIB-R datasets reflect a single interpretation and a tool may therefore produce a correct interpretation that could result in a low quality if another interpretation is expected in the gold standard. A possible solution is to provide different expert FRBR collections to take into account the various interpretations, but this would complicate the management of datasets. Hopefully these ambiguous cases are not so frequent and should not have a high negative impact on the quality.

Finally, datasets in our benchmark only considers one scenario, the full migration (i.e., transforming a whole catalogue into FRBR). However, there exist other scenarios that are of great interest for the librarian community with different challenges. A library may decide to migrate a subset of its catalogue, for instance for testing a FRBR version of their catalogue (partial migration). Another scenario is synchronization, in which two libraries, whose catalogues are based on different models (MARC and FRBR), need to share or synchronize their data. Finally, the adoption of FRBR in libraries will cause a transition period during which MARC records and FRBR collections will co-exist. Libraries with an FRBR catalogue may still frequently receive MARC records that need to be integrated in their catalogue, thus implying an online migration. The main challenge

in this scenario deals with scalability: searching for duplicate entities in a potentially large catalogue requires efficient algorithms to avoid latency for end-users.

We acknowledge that the datasets are subject to various interpretations and that they focus on a single scenario. However, we explain in the next section why these issues do not have a significant impact on the experimental validation.

8 Experiments

The benefits of our benchmark BIB-R for the evaluation of FRBRization has been demonstrated by an experiment using three tools which are publicly available: FRBR-ML[16], Extensible Catalog[17] (XC) and Variations VFRBR[18]. All tools are within the rule-based category and have been presented in Section 2. The experiments conducted with these tools cover: evaluating the strengths and weaknesses of FRBRization tools, comparing tools in a real-world FRBRization scenario and facilitating the tuning of a tool. Only the most relevant results are presented, but additional data and plots are publicly available in an online appendix [15].

8.1 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 evaluated using the post-FRBRization metrics. In other words, each tool has produced a result for each test, and these results have been compared to the expert collection provided in the benchmark. A basic set of rules is available for each tool. For equity reasons, we have not tuned the tools (by updating their set of rules), although they have been developed for different purposes. Recall that the post-FRBRization metrics produces scores between 0 and 100%, with the value 0 standing for a perfect result. We present in this paper the results for a subset of tests[19], namely `1.0`, `2.1`, `3.2`, `4.3` and `5.5` (check the appendix for other tests [15]).

[16] FRBR-ML tool, previously named marc2frbr

[17] Extensible Catalog

[18] Variations VFRBR tool (adjusted version, only to facilitate compilation)

[19] The tests have been chosen according to a sequential order (remind that test `5.4` does not exist). The analysis of the results is however not limited to this subset.

8.1.1 Missing data

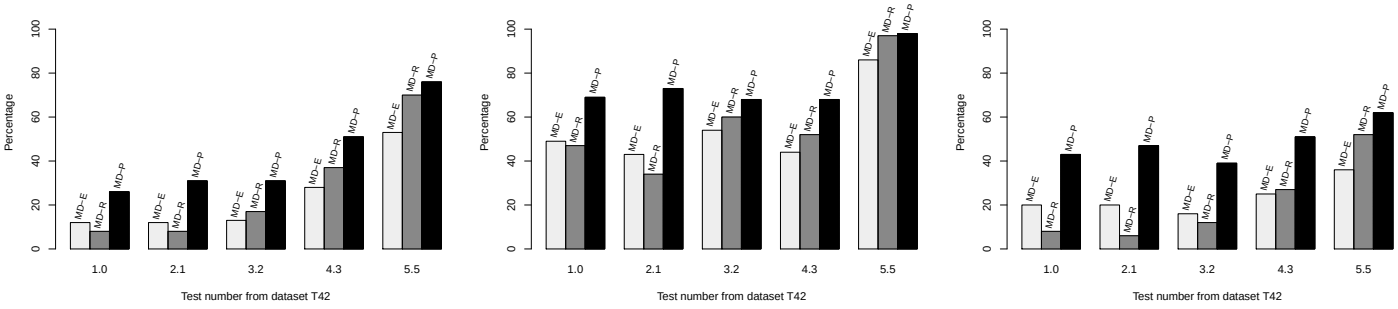
A first finding is that none of the tools completely transform all information contained in the original MARC records, thus some elements do not appear in the output collection. Figure 10 illustrates this trend by showing the missing elements in terms of entities (MD-E), relationships (MD-R) and properties (MD-P) for the selected tests. The missing relationships are correlated with missing entities: when an entity is not created, its relationship to the rest of the collection is also missing. The properties are the most elements in the test that are most missing in the results. Similarly to relationships, properties are dependent on an entity since an entity that is missing will also exclude its property values from the result. The more complex the record becomes (bibliographic pattern and/or issue), the more is missing in the output. Even with a simple case (test 1.0, core pattern without any issue), the tools may miss entities such as Concepts.

For FRBR-ML, missing data is limited for simple patterns and issues that are trivial to account for in the rules. With more complex patterns or issues, the results show more missing data. The scores of VFRBR for missing data are strongly impacted by the fact that it does not create Work entities (thus decreasing the MD-R and MD-P too). XC is the tool that shows the most stable results on each test[20]. The values for missing properties suffer from the fact that this tool is implemented to merge some properties, such as making a single title string by combining the main title with related subtitles.

8.1.2 Incorrectly added data

This study concerns data that is added in the FRBRization, but is not found in the expert collection. Misplaced data is also covered by this category i.e., properties which should have been associated with another entity or used as a link between other entities. Figure 11 depicts the results achieved by the three tools in terms of incorrectly added entities (IAD-E), incorrectly added relationships (IAD-R) and incorrectly added properties (IAD-P). Neither of FRBR-ML and VFRBR adds any incorrect data (for all tests). It also means that they are able to put relationships and properties at the correct position. The tool XC does not add any incorrect entities, however, it generates incorrectly added data in terms of properties and relationships (roughly between

[20] Note that XC does not create Agent and Concepts entities, but it rather adds properties within the main Work or Expression. Our evaluation takes this specificity into account and XC is not penalized when a property and its associated value correctly represent the Agent or the Concept.



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

Figure 10: Experiment results for evaluating missing data (MD-E, MD-R, MD-P)

10% to 20% for a given test). These additional data are in fact misplaced data (e.g. the abstract property is attached to the Expression rather than the Work).

8.1.3 Semantic mismatch data

The third experiment is for evaluating data which is correctly placed, but with a different semantics than expected. For instance, an entity *Agent* could be used instead of the more appropriate subtypes *Person* or *Corporate body*. Similarly, a relationship *contributed* could replace an expected relationship *created*. The metrics semantic mismatch entity (SMD-E), semantic mismatch relationship (SMD-R) and semantic mismatch property (SMD-P) reflect this issue, as shown in Figure 12. We found that all tools produce semantic mismatch, but not significantly. Less than 10% of the relationships have a different semantics than in the expert collection. Besides, this issue is only present at the relationship level: it occurs in 36 tests for FRBR-ML, 24 tests for VFRBR and 21 tests for XC (out of 42 tests). Finally, we note that the majority of relationships are associated with the more complex patterns, and this result is consequently dependent on the ability of the tool to detect a variety of patterns.

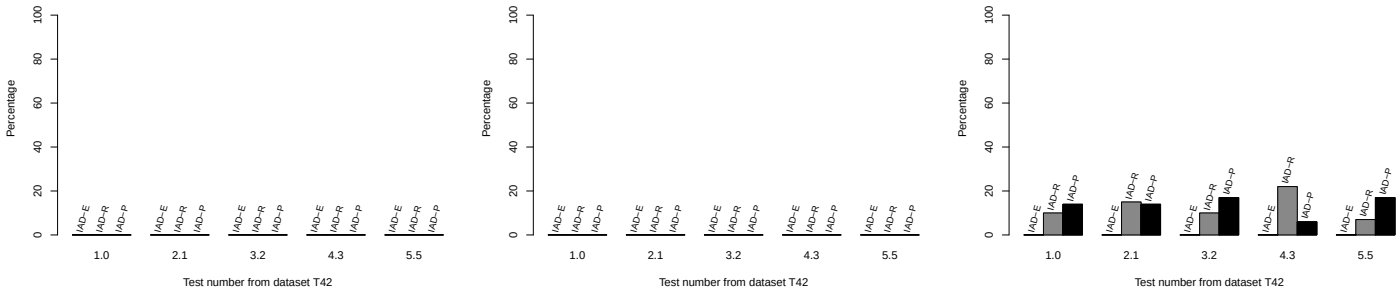
8.1.4 Detection of bibliographic patterns

To measure the tools ability to detect patterns, our metrics measure the number of patterns in which the main entity has not been detected (MEND), in which the main relationship has not been detected (MRND) and in which any secondary elements of a pattern raise an error (ESE). Figure 13 depicts the scores obtained by the three tools for these metrics. Broadly speaking, it appears that most tools do not perform well (many plots have values close or equal to 100% of elements not detected). This is not surprising since pattern detection is one of the biggest challenges in FRBRization. As a reminder, the experiments also use the default set of rules provided with the tools, which may not have been prepared for processing complex patterns. FRBR-ML obtains good results for detecting the core pattern (test 1.0), and it is able to discover half of the main entities in derivations, aggregations and complementary works but none in augmentations. It cannot identify the main relationship in complex patterns (even when the main entity is found). However, it is successful in detecting the secondary elements of a pattern, except for derivations[21]. VFRBR is not able to detect most patterns, even for the core pattern. This is mainly because this tool was designed with music records in mind and this cataloguing practice does not always give correct interpretations if applied to e.g. monographs. However it manages to discover a few secondary elements for some pattern categories. XC achieves quality scores which strongly depends on the category of patterns. For the core pattern, it is able to detect main entities and main relationships effectively, but only half of the secondary elements. With augmentations, it fails for extracting the main entity, but not the secondary elements. For derivations and aggregations, the difficulty lies in the detection of the main relationships while half of the main entity and secondary elements are identified. The behaviour for complementary works is influenced by the accompanying issue of cataloguing practice, but half of the patterns are usually detected.

8.1.5 Overall analysis

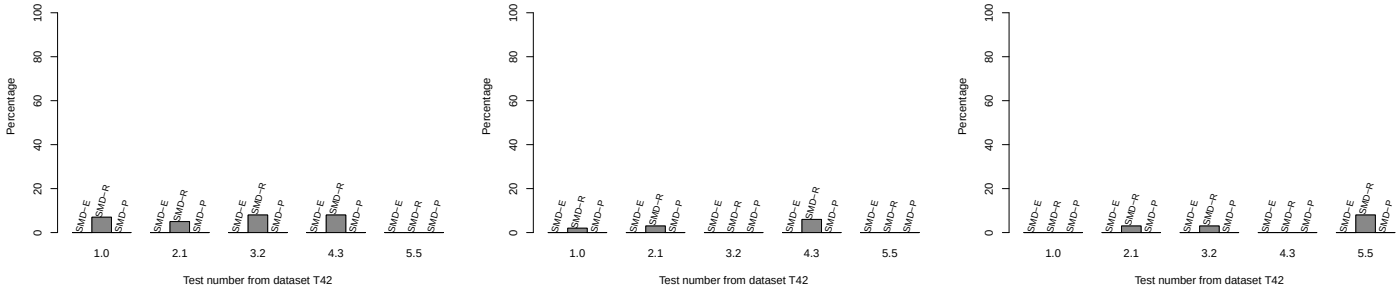
A last experiment about strengths and weaknesses enables a broader view on the quality obtained by the tools. First, we use more generic metrics to evaluate

[21] Note that the category patterns 4.x (aggregations) and 5.x (complementary works) do not have secondary elements and all tools achieve a 0% ESE score for these tests.



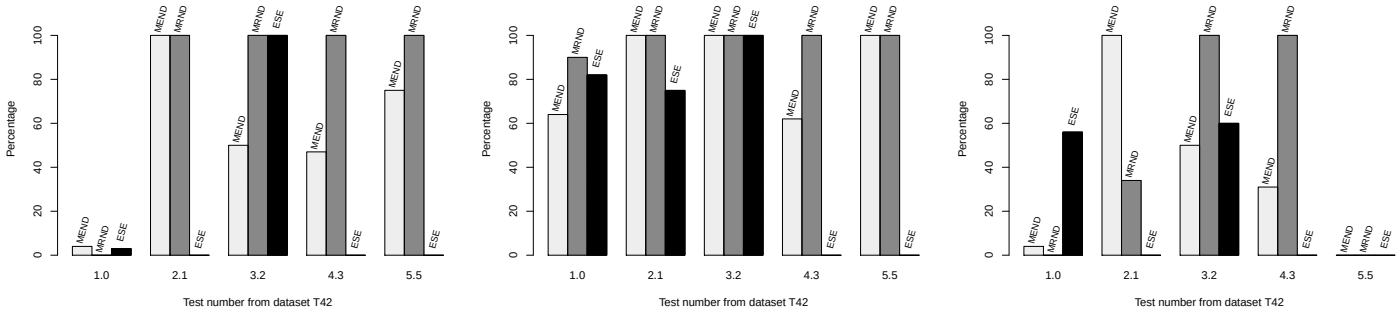
(a) FRBR-ML (b) VFRBR (c) XC

Figure 11: Experiment results for evaluating incorrectly added data (IAD-E, IAD-R, IAD-P)



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

Figure 12: Experiment results for evaluating semantic mismatch data (SMD-E, SMD-R, SMD-P)



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

Figure 13: Experiment results for evaluating bibliographic patterns (MEND, MRND, ESE)

the tools: MD for all types of missing data, IAD for all types of incorrectly added data, SMD for all types of semantic mismatch and FPND for any error in detected patterns. For missing data, the global trend is that it slowly increases according to the complexity of the dataset. FRBR-ML misses around 25% in test 1.0 and it gradually reaches 80% in tests 5.x. VFRBR results ranges from 70% up to 90%. XC starts with results around 40% which increases to 65% in the last tests. The incorrectly added data is not relevant for FRBR-ML and VFRBR, which both have scores equal to 0%. We already have discussed the additional data included by XC for relationships and properties. In average, this tool achieves IAD scores between 10-15%. Similarly, the semantic mismatch is not important in this scenario. All tools obtain a score equal or close to 0%, which

indicates that all tools are mainly able to handle correctly such issue. Last but not least, the detection of complete patterns using the metric FPND: contrary to previous metrics, this one is not a weighted average of more detailed metrics. For a pattern to be complete, it needs to have its main entity, its main relationship and all secondary elements. Consequently, achieving acceptable scores for this strict metric is not an easy task. FRBR-ML is able to detect almost all patterns in the core category (6%, an error due to a single record with two authors in added entries which are not correctly linked to the Work). For the other patterns, FRBR-ML totally fails (100%), mainly because of a forgotten entity or relationship. VFRBR follows the same behaviour as FRBR-ML: patterns are fully detected for the core category only (100% for tests above 2.0). But even for the simplest category, VFRBR is limited to the discovery of a few patterns (88%). These poor results are again explained by the fact that the tool only creates a Work and an Expression when a uniform title is found (a situation which occurs in two records). XC is not able to detect any full pattern in tests 1.0 up to test 4.10, mainly due to the absence of creation of Agent entities. In the complementary works category, XC correctly detects a few complete patterns (67%).

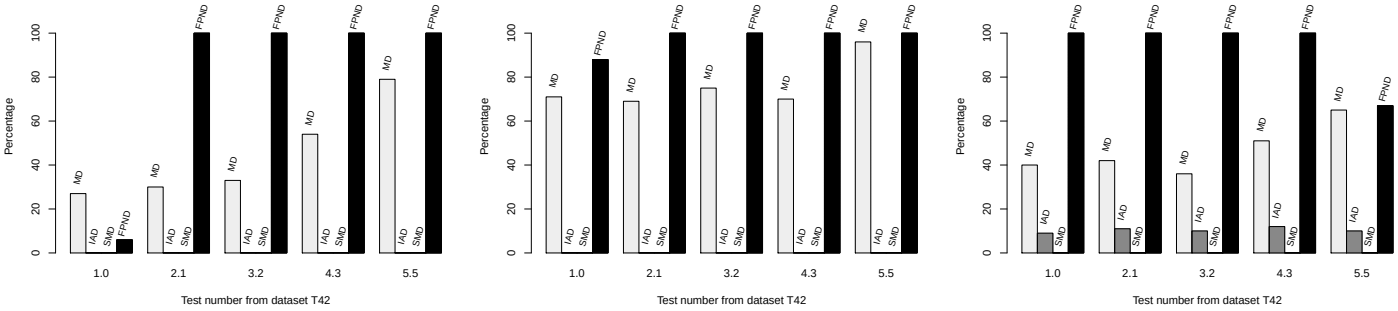
Another means of presenting global results is to aggregate them for each category of pattern (Core, Augmentation, etc.), which takes into account possible cataloguing errors and other quality issues. Table 6 indicates, for a given tool and a given category of pattern, the metrics with a high score (average above 90% for all tests in a pattern category). For instance, FRBR-ML obtains a high score for the IAD metric in all tests of the core category of pattern. Thus, we can consider that this tool is reliable for this metric in this context. The same applies for low scores (average below 10% for all tests in a pattern category) and the results are presented in Table 7. For instance, the FRBR-ML tool obtains a low score for the ESE metric in the core category. Concerning the successful metrics, we note that SMD is always above 10%, whatever the tool or the pattern category. The benchmark does not include many cases for which the semantics could be ambiguous. FRBR-ML and VFRBR also manage to avoid the incorrect addition of data (metric IAD in all categories). In some categories, FRBR-ML and XC may obtain good scores for a pattern-related metric (e.g. MEND or MRND for the core category using XC), but not VFRBR. The secondary elements of a pattern (ESE) can be well managed even for complicated patterns (aggregation, augmentation), which shows that the tools are able to extract individual information from a record (e.g. the name of a translator) but not able to derive the complete pattern. This last assumption is also supported by the table of the most failed metrics, which are all related to the bibliographic patterns. The main relationship of a pattern (MRND) is almost never detected correctly by any tool (except in the core category). FRBR-ML and XC share the same behaviour (low scores for the same metrics and same pattern category) while VFRBR fails with more metrics.

Finally, we present an overview of the results using 3D plots. Figure 17 depicts the scores of the three tools for each test case and each metric. The X-axis stands for the 42 test cases while the Y-axis represents the metrics (IAD stands for IAD, IAD-E, IAD-R and IAD-P, PAT stands for FPND, MEND, MRND and ESE, etc.). Note that the objective of these plots is not to point out a specific value but only to provide the global shape of the results. FRBR-ML and XC both have a similar shape and color with lower quality for detecting the patterns. VFRBR has a similar shape, but it also achieves a lower quality because of missing data, which explains that its plots includes more green ”peaks” and red ”summits”.

At the end of Section 7, we discussed open issues related to our datasets. These issues do not have a significant impact on the presented results for several reasons. First, all unknown properties produced by the tools were detected, manually verified and added in the mapping file if needed so that completeness is guaranteed for their implementations. Besides, specific features such as enrichment from external data sources in FRBR-ML were disabled to avoid a penalty caused by incorrectly added data. To the best of our knowledge, only two records contain a controversial interpretation (case of a journal and its issues, in the tests Aggregation and Complementary Works). With an average of 10 records per test, each controversial record could decrease the quality in a few tests by about 10% if not correctly detected. To summarize, our dataset T42 and post-FRBRization metrics are useful for understanding the failures of a tool and for selecting a tool according to its capabilities. The results can be presented and studied according to several dimensions.

8.2 Comparing tools in real-world context

The objective of this second experiment is to compare FRBRization tools in a real-world context using both FRBRization and post-FRBRization metrics. All tools rely on their basic set of rules (no tuning). Note that the FRBRization metric TAC is not presented in this experiment, because we have chosen not to configure



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

Figure 14: Experiment results for a global evaluation (MD, IAD, SMD, FPND)

	Core	Augmentation	Derivation	Aggregation	Complementary work
FRBR-ML	IAD, SMD, MEND, MRND	IAD, SMD	IAD, SMD	IAD, SMD	IAD, SMD
VFRBR	IAD, SMD	IAD, SMD	IAD, SMD	IAD, SMD	IAD, SMD
XC	SMD, MEND, MRND	ESE, SMD	SMD	SMD	SMD

Table 6: Metrics with high score for each FRBRization tool and category

	Core	Augmentation	Derivation	Aggregation	Complementary work
FRBR-ML	ESE	MEND, MRND	MRND, ESE	MRND	MRND
VFRBR	MRND, ESE	MEND, MRND, ESE	MEND, MRND, ESE	MRND	MEND, MRND
XC	ESE	MEND,MRND	MRND, ESE	MRND	MRND

Table 7: Metrics with low score for each FRBRization tool and category

any tool. Detailed execution times are not available for the three evaluated tools, thus we provide the overall execution time for FRBRizing the dataset BIB-RCAT in the bottom row of Table 8, which corresponds to the sum of execution times for interpreting entities and for deduplication (ETC and ETD). The post-FRBRization metric DLE is not given either, since there exists too many authority files or knowledge bases (e.g. Linked Open Data) and the expert FRBRized collection cannot include a link for each of these sources.

Table 8 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 incorrectly detected patterns. 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 cataloguing practice challenges, which complicate the interpretation of the records. Finally, some additional entities (e.g. Concept) are not interpreted and created. The basic set of rules are not sufficient for achieving an acceptable quality.

The execution times (ETC + ETD) are acceptable with this rather small dataset (less than one minute for all tools). The performance results are representative for very small collections, but do not reflect the performance of a full migration from a MARC collection to FRBR. Large scale migration and continuous update with new FRBR records requires an efficient indexing of the FRBR collection to quickly identify potential duplicate entities.

To conclude, this experiment showed that our dataset BIB-RCAT and associated metrics are useful to compare tools in a real-world context. For performance evaluation, much larger datasets are needed.

8.3 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

Metric	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%
ETC + ETD	2.8s	44.9s	2.8s	3.4s

Table 8: Results of FRBR-ML, VFRBR and XC for the dataset BIB-RCAT

8, the results of FRBR-ML for dataset BIB-RCAT can be improved. To provide insight to the expert, we compute predictive scores for the basic set of rules on the dataset BIB-RCAT. The predictive scores are shown in Figure 15. The white bar (FRBR-ML) stands for the results with the basic set of rules. For instance, we note that 37% of the records contain cataloguing practice challenges (metric CPN) and that 40% of the records contain an augmentation pattern. On the contrary, only a few records include aggregation (AGG equal to 4%) or miss the authoritative responsibility (MAR equal to 6%). This information is crucial because it helps prioritising the rules that should be added to obtain a high impact on the quality. The predictive metrics also provides information about the set of rules. The basic set of rules provided with FRBR-ML contains many rules that are not used for the dataset BIB-RCAT (score of UR equal to 85%) and 24% of rules are missing to take into account all fields from the dataset BIB-RCAT. Some of the rules are conflicting (CR equal to 4%). Finally, the metrics for specific patterns indicate that 100% of the rules needed to tackle derivations (metric MR-DER) are missing, augmentations (metric MR-AUG) and aggregations (metric MR-AGG).

Based on these predictive scores, an expert has enhanced the basic set of rules of FRBR-ML. This update corresponds to the tool application cost (TAC metric). It took 4 hours in total[22], which can be decomposed into:

- About 1 hour for correcting minor changes such as type-corrections and adding rules for missing subfields. Most of this time is used to identify the correct type (i.e., checking the mapping to the RDA vocabulary, looking up the RDA registry);
- About 1 hour to implement functions that simplifies the rules with respect to setting the right type of relationship according to relator codes, as well as coding proper handling;
- About 1 hour to add new templates (e.g. for parent work when the field 245$p is encountered, Concepts from field 650, or templates to create augmentation works);
- About 1 hour for testing and debugging.

The enhanced set of rules has been tested with the prediction metrics, and it appears with black bars in Figure 15 (FRBR-ML_tuned). Now, only 7% of the rules are missing to process all fields, and a few rules not in use have been deleted (metric UR down to 77%). The most significant enhancement deals with the pattern detection: all rules to identify both augmentations and 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 8 (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.

9 Conclusion

During the last decades, the growing interest for adopting a semantic model such as FRBR in library catalogs and systems has led to the development of many FRBRization tools. These tools should not only transform legacy records by merely adding semantic annotations to each data element, but they also need to identify implicit bibliographic patterns and represent them correctly in terms of entities and relationships. Besides, the specificities and different cataloguing practices have to be taken into account when interpreting records. Addressing these challenges requires both gold standard datasets created by experts and relevant metrics to al-

[22] Note that the expert had knowledge about the proposed metrics, and the given time may increase for people who need to understand the concepts behind these metrics.

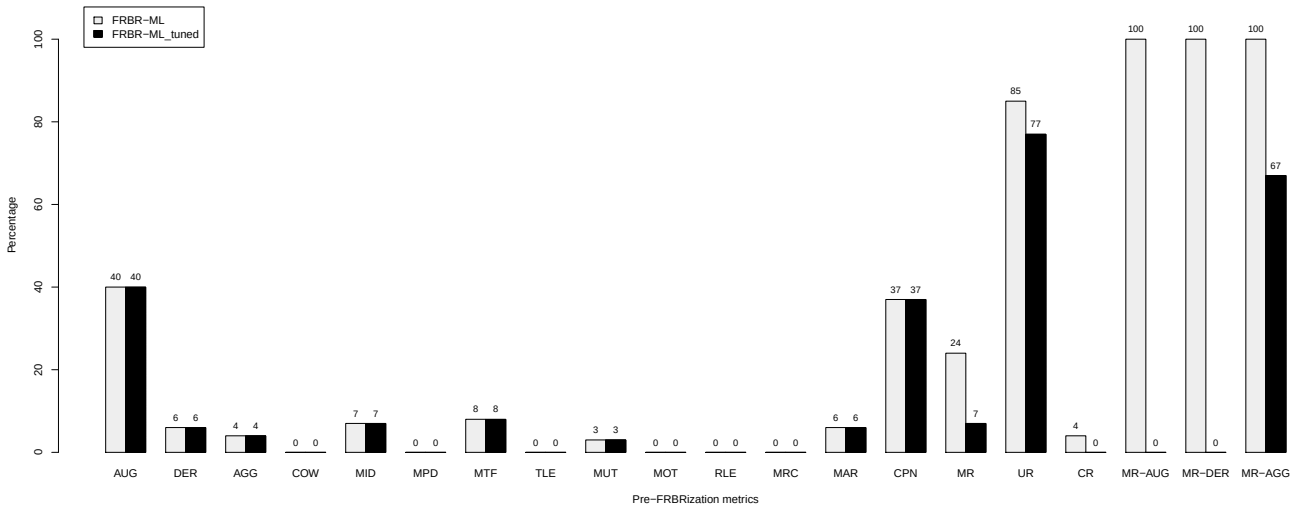


Figure 15: Applying predictive metrics on BIB-RCAT for FRBR-ML basic rules and tuned rules

low any tool to be fairly compared and evaluated.

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 publicly available datasets (T42 and BIB-RCAT). Our metrics enable the evaluation of each stage of the FRBRization process. Before running the transformation, they provide statistics and prediction about the possible outcome given a set of rules. During the process, mainly performance metrics are computed. At the end of the transformation, it is about evaluating the quality of the produced catalogue. To facilitate the use of all proposed metrics, a guidance diagram is proposed in Figure 16 so that users may quickly identify the relevant metrics. To promote fair evaluation, our datasets can be used either for testing a specific feature (detection of a bibliographic pattern and/or a cataloguing issue) or for simulating the FRBRization of a real-world collection. Extensive experiments have been performed with three recent tools (FRBR-ML, Variations VFRBR and Extensible Catalog) to demonstrate the possibility to identify their strengths and weaknesses. We have also shown how the pre-FRBRization metrics are useful for determining the rules that need to be added or updated.

The release of this benchmark brings different perspectives. First, the datasets could be improved and a complete list of statistics provided (e.g., number of expected augmentations in each test). Adding more records in both datasets would be beneficial, but the main challenge is to update the FRBR expert collection. In its current version, BIB-R has no support for the new concepts from FRBR Library Reference Model [41]. Although these new concepts are rather limited (e.g. Timespan), the benchmark should include new data in the original records to enable the use of these recent concepts. Open issues were previously described in Section 7: the different interpretations of the FRBR specification and the focus on a single scenario (full migration). The former problem needs to be solved by librarians. Otherwise, the benchmark needs to be flexible to integrate gold standard for each case. Designing datasets for other scenarios such as a partial migration or a synchronization is another perspective. About the metrics, we support the idea that the FRBRization metrics should be implemented internally by FRBRization tools to enable evaluation of the process. Graphical user interfaces should be developed and metrics for assessing the efficiency of the user interaction design would help select a tool which facilitates the user validation task. The metrics about missing data could be refined according to the importance of the data (a record identifier is rarely used while an original title is crucial). In the mid-term, the tools will enable semantic enrichment of the collections. In such case, the metrics about incorrectly added data needs to be refined to distinguish between an incorrect data and a correct data which was integrated from external resources. The metric in charge of computing data linking errors is difficult to set up since it requires links to a possibly large set of external resources. A more realistic solution could consist in participating in Information Retrieval challenges such as the Knowledge Base Population challenge[23]. Links to the most common knowledge databases such as

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

Linked Open Data (VIAF, DBpedia, Geonames) could also be semi-automatically detected for the entities of our benchmark.

Acknowledgements 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).

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). DOI 10.1007/11931584_31
2. Aalberg, T., Merčun, T., Žumer, M.: Coding FRBR-Structured Bibliographic Information in MARC, pp. 128–137. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). DOI 10.1007/978-3-642-24826-9_18. URL http://dx.doi.org/10.1007/978-3-642-24826-9_18
3. Aalberg, T., Žumer, M.: The Value of MARC Data, or, Challenges of FRBRisation. Journal of Documentation **69**, 851–872 (2013)
4. 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)
5. Alexe, B., Tan, W.C., Velegrakis, Y.: STBenchmark: towards a benchmark for mapping systems. Proceedings of the VLDB **1**(1), 230–244 (2008)
6. Bailey, P., Hawking, D., Krumpholz, A.: Toward meaningful test collections for information integration benchmarking. In: Proceedings of IIWeb (2006). URL http://es.csiro.au/pubs/bailey_iiweb.pdf
7. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological) pp. 289–300 (1995)
8. Boehm, B., Abts, C., Chulani, S.: Software development cost estimation approaches — a survey. Annals of Software Engineering **10**(1), 177–205 (2000). DOI 10.1023/A:1018991717352. URL http://dx.doi.org/10.1023/A:1018991717352
9. Bowen, J.: Moving Library Metadata Toward Linked Data: Opportunities Provided by the eXtensible Catalog. International Conference on Dublin Core and Metadata Applications (2010)
10. Buchanan, G.: FRBR: Enriching and Integrating Digital Libraries. In: Proceedings of Joint Conference on Digital Libraries, pp. 260–269 (2006). DOI 10.1145/1141753.1141812
11. Chang, N., Tsai, Y., Dunsire, G., Hopkinson, A.: Experimenting with implementing FRBR in a Chinese Koha system. Library Hi Tech News **30**, 10–20 (2013)
12. Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In: Quality Measures in Data Mining, pp. 127–151. Springer (2007)
13. Committee, S., Group, I.S.: Functional Requirements for Bibliographic Records: final report, vol. 19. K. G. Saur (1998)
14. Coyle, K.: FRBR, Twenty Years On. Cataloging & Classification Quarterly pp. 1–21 (2014)
15. Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., Lumineau, N.: Appendix: Benchmarking and Evaluating the Interpretation of Bibliographic Records. Tech. rep., LIRIS, NTNU (2016). URL http://liris.cnrs.fr/~fduchate/docs/appendix/appendix-BIB-R.pdf
16. Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., Lumineau, N.: BIB-R: a Benchmark for the Interpretation of Bibliographic Records. In: Theory and Practice of Digital Libraries (TPDL). Hannover, Germany (2016). URL https://hal.archives-ouvertes.fr/hal-01324529
17. 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)
18. Decourselle, J., Duchateau, F., Lumineau, N.: A Survey of FRBRization Techniques. In: Theory and Practice of Digital Libraries, pp. 185–196 (2015). URL https://hal.archives-ouvertes.fr/hal-01198487
19. Denton, W.: FRBR and the History of Cataloging. Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools (2007)
20. 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)
21. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on knowledge and data engineering **19**(1), 1–16 (2007)
22. Euzenat, J., Rosoiu, M.E., Trojahn, C.: Ontology matching benchmarks: generation, stability, and discriminability. Web Semantics: Science, Services and Agents on the World Wide Web **21** (2013)
23. Hickey, T., Vizine-Goetz, D.: Implementing FRBR on large databases. Dublin, Ohio: OCLC (2002)
24. Hickey, T.B., O'Neill, E.T.: FRBRizing OCLC's WorldCat. Cataloging & Classification Quarterly **39**, 239–251 (2005)
25. Hickey, T.B., Toves, J.: FRBR Work-Set Algorithm (2.0). OCLC (2009). URL http://www.oclc.org/research/activities/frbralgorithm.html?urlm=159780
26. Ioannou, E., Rassadko, N., Velegrakis, Y.: On generating benchmark data for entity matching. Journal on Data Semantics **2**(1), 37–56 (2013). DOI 10.1007/s13740-012-0015-8
27. Kilner, K.: The AustLit gateway and scholarly bibliography: A specialist implementation of the FRBR. Cataloging & Classification Quarterly **39**, 87–102 (2005)
28. Kroeger, A.: The road to bibframe: the evolution of the idea of bibliographic transition into a post-marc future. Cataloging & classification quarterly **51**(8), 873–890 (2013)
29. Le Bœuf, P.: Customized OPACs on the Semantic Web: the OpenCat prototype. IFLA World Library and Information Congress pp. 1–15 (2013)
30. Leopold, C.: Parallel and Distributed Computing: A survey of Models, Paradigms and approaches. John Wiley & Sons, Inc. (2001)
31. Leroy, G.: Gold Standard and User Evaluations, pp. 131–137. Springer London, London (2011). DOI 10.1007/978-0-85729-622-1_9. URL http://dx.doi.org/10.1007/978-0-85729-622-1_9
32. Manguinhas, H.M.A., Freire, N.M.A., Borbinha, J.L.B.: FRBRization of MARC Records in Multiple Catalogs. In: J. Hunter, C. Lagoze, C.L. Giles, Y.F. Li (eds.) JCDL, pp. 225–234. ACM (2010)
33. Minadakis, N., Marketakis, Y., Kondylakis, H., Flouris, G., Theodoridou, M., Doerr, M., de Jong, G.: X3ml framework: An effective suite for supporting data mappings. In: Workshop for Extending, Mapping and Focusing the CRM—co-located with TPDL (2015)

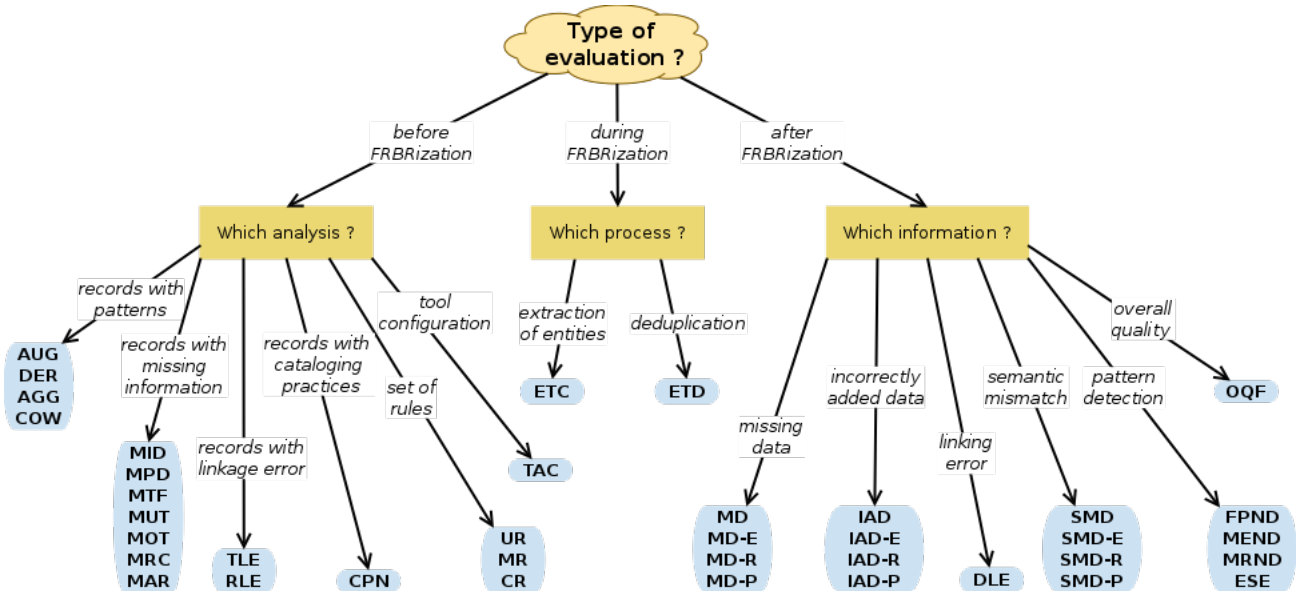
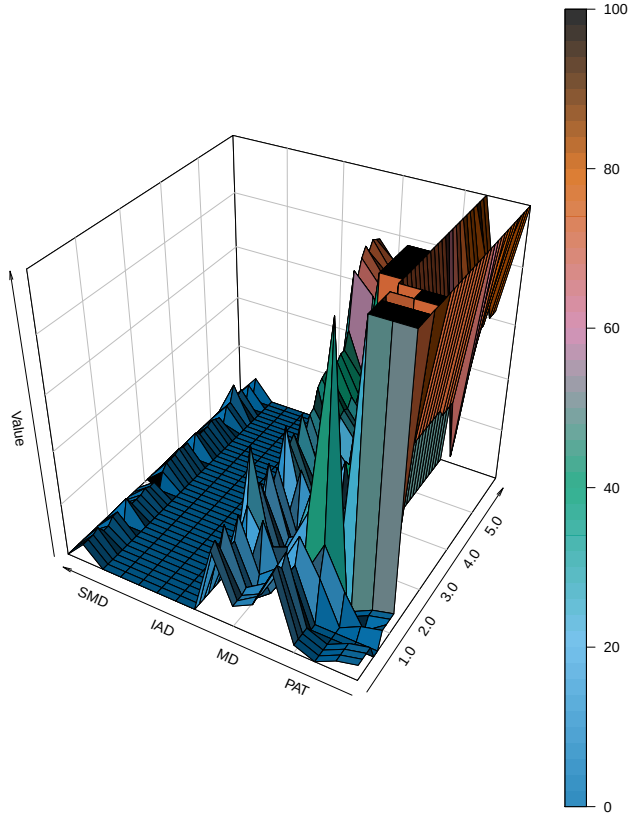
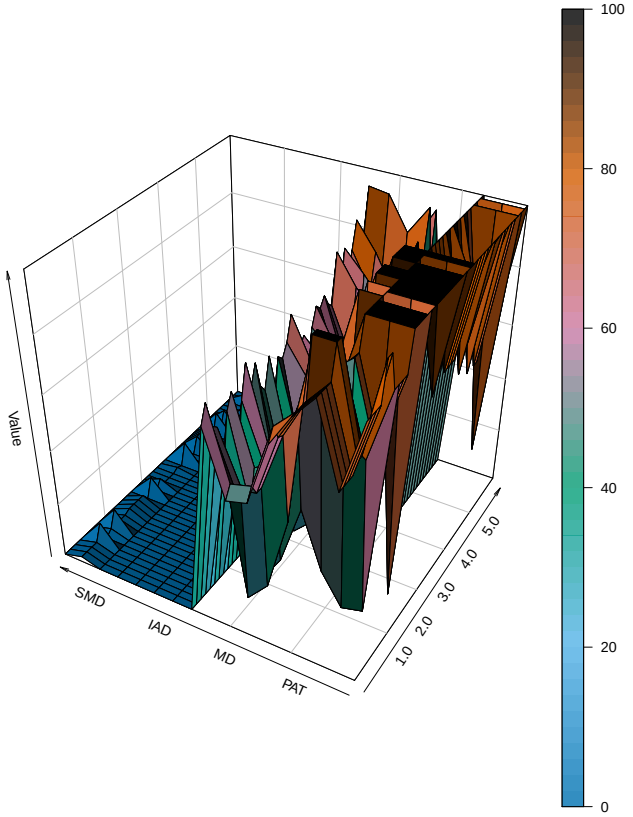


Figure 16: Guidance diagram for all metrics included in BIB-R. The diagram is read from the top and it enables the identification of relevant metrics (blue leaf nodes) by selecting desired criteria (questions in yellow internal nodes and answers on the edges). Pre-FRBRization metrics are fully described in Section 4, FRBRization metrics are presented in Section 5 and post-FRBRization metrics are detailed in Section 6

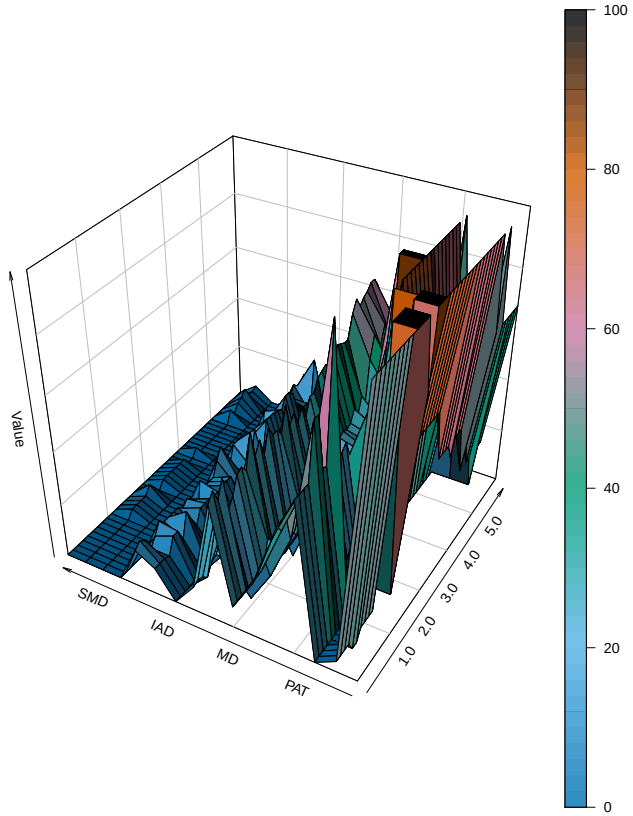
34. Norman, D.A.: The design of everyday things: Revised and expanded edition. Basic books (2013)
35. Notess, M., Dunn, J.W., Hardesty, J.L.: Scherzo: A FRBR-Based Music Discovery System. In: International Conference on Dublin Core and Metadata Applications, pp. 182–183 (2011)
36. Phipps, J., Dunsire, G., Hillmann, D.: Building a platform to manage rda vocabularies and data for an international, linked data world. Journal of Library Metadata **15**(3-4), 252–264 (2015). DOI 10.1080/19386389.2015.1099990. URL http://dx.doi.org/10.1080/19386389.2015.1099990
37. Pisanski, J., Žumer, M.: User verification of the frbr conceptual model. Journal of Documentation **68**(4), 582–592 (2012). DOI 10.1108/00220411211239129
38. Putz, M., Schaffner, V., Seidler, W.: FRBR: The MAB2 Perspective. Cataloging & Classification Quarterly **50**, 387–401 (2012)
39. Riley, J.: Enhancing Interoperability of FRBR-Based Metadata. International Conference on Dublin Core and Metadata Applications (2010)
40. Riva, P.: Mapping MARC 21 Linking Entry Fields to FRBR and Tillett's Taxonomy of Bibliographic Relationships. Library resources & technical services **48**(2), 130–143 (2013)
41. Riva, P., Le Boeuf, P., Žumer, M.: FRBR-Library Reference Model. Tech. rep., IFLA FRBR Review Group (2016). URL https://www.ifla.org/files/assets/cataloguing/frbr-lrm/frbr-lrm_20160225.pdf
42. Romero, G.C., Esteban, M.P.E., Such, M.M., Carrasco, R.C.: Transformation of a Library Catalogue into RDA Linked Open Data. In: Theory and Practice of Digital Libraries (TPDL), pp. 321–325 (2015). DOI 10.1007/978-3-319-24592-8_26. URL http://dx.doi.org/10.1007/978-3-319-24592-8_26
43. Schneider, J.: FRBRizing MARC records with the FRBR Display Tool (2008). URL http://jodischneider.com/pubs/2008may_frbr.html
44. Takhirov, N., Aalberg, T., Duchateau, F., Žumer, M.: FRBR-ML: A FRBR-based framework for semantic interoperability. Semantic Web Journal **3**, 23–43 (2012)
45. Vassallo, V., Piccininno, M.: Aggregating Content for Europeana: A Workflow to Support Content Providers, pp. 445–454. Springer Berlin Heidelberg (2012). DOI 10.1007/978-3-642-33290-6_50. URL http://dx.doi.org/10.1007/978-3-642-33290-6_50
46. Vila-Suero, D., Villazón-Terrazas, B., Gómez-Pérez, A.: datos. bne. es: A library linked dataset. Semantic Web **4**(3), 307–313 (2013)
47. Zhang, Y., Salaba, A.: Implementing FRBR in libraries: key issues and future directions. Neal-Schuman Publishers (2009)



(a) FRBR-ML



(b) VFRBR



(c) XC

Figure 17: Experiment results with 3D plots (all test cases, all post-FRBRization metrics)