# Towards a Pattern-based Semantic Enrichment of Bibliographic Entities

Joffrey Decourselle

# Towards a Pattern-based Semantic Enrichment of Bibliographic Entities

Joffrey Decourselle*
joffrey.decourselle@liris.cnrs.fr

LIRIS, UMR5205, Université Claude Bernard Lyon 1, Lyon, France

**Abstract.** The semantic enrichment of cultural heritage data is a crucial step for memory institutions to provide enhanced search services and to reduce the costs for managing resources. Yet, using information systems based on semantic web technologies implies to migrate and enrich the legacy data according to the dedicated semantic models. In the library community such task can be very complex when handling large catalogs of records stored using old and flat models. This paper focuses on the challenges related to metadata migration process and semantic enrichment of bibliographic data. We present the contributions and lessons learned from the FRBRisation study and we provide preliminary discussions and ongoing work about the enrichment of FRBRised entities.

**Keywords:** semantic enrichment, metadata migration, FRBRisation, FRBR, record interpretation, metric, model, tool

## 1 Introduction

The emergence of Semantic Web technologies makes people who seek cultural heritage data expecting more efficient and simplified ways to get information [6]. With catalogs of cultural data growing more and more, memory institutions have to improve their information systems and reinvent their practices [28, 17]. This thesis focuses on a major challenge for cultural institutions: *how to improve the search and enrichment of information while reducing the cataloguing efforts?*

In the digital library community, the widely used data models like the Machine-Readable Cataloging formats (MARC) have shown their limitations for reusing and enriching the semantics of data [16]. Such models, based on a flat structure and coded metadata, have been derived for years and suffer from many different cataloguing practices. Several national libraries promote the adoption of new semantic models with new cataloguing rules such as principles from the Function Requirements for Bibliographic Records (FRBR) [27], and more recently, the Library Reference Model (LRM) [22]. Those recommendations enhance the representation of bibliographic data and its enrichment leveraging on an entity-relationship formalism and vocabularies with more explicit semantics [5].

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

However, only few libraries already use FRBR, mainly because the transformation of the legacy data to the new model is still a very challenging task [1]. When migrating thousand of records, the process must be done semi-automatically using dedicated rules. In the context of MARC-based catalogs, the FRBRisation (i.e., the metadata migration from MARC to FRBR) may imply hundreds of migration rules due to the deep differences between the flat model of data in input and the graph-based structure of FRBR. The Figure 1 shows the typical three-steps workflow of the FRBRisation process. The *Tuning* task aims at writing the migration rules, the *Extraction* step consists in applying all rules on each record to generate FRBR entities and relationships and the *Normalization* task aims at cleaning and merging the extracted entities to build a FRBR catalog. Quality of such migration progress, when done automatically, can be significantly altered due to inconsistencies in data, duplicate resources or even specific cataloguing practices. Thus, the FRBRisation requires time to analyze a catalog, to write the migration rules and to validate the results of the process.

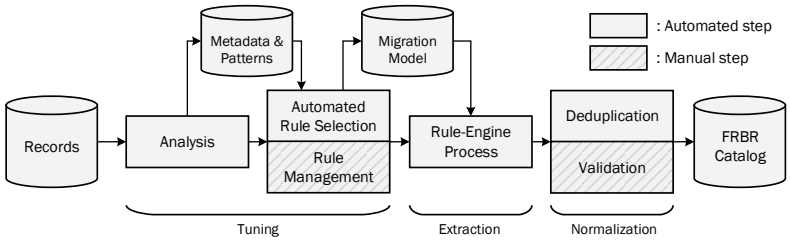


**Fig. 1.** Example of a semi-automated workflow of the FRBRisation process

Even if the obstacles to the Semantic Web adoption are already known, the evolution of digital catalogs remains a necessary step for librarians to move forward [31]. That is why they must be accompanied with the relevant tools to facilitate and automate as much as possible their migration process [20, 11]. This paper presents our contributions and perspectives related to the semantic enrichment of bibliographic data. First, we detail new improvements for the FRBRisation process in terms of effectiveness (e.g., by enhancing the interpretation of records in input) and in terms of efficiency (e.g., by using automatically assigned rules). Secondly, we introduce our benchmark for the FRBRisation in which we provide different metrics, dedicated datasets and evaluation results done on recent FRBRisation solutions. Finally, we present initial works done for the enrichment of FRBRised entities using external repositories.

The remainder of the paper is organized as follows: Section 2 presents motivating challenges and summarizes the state of the art, Section 3 describes the contributions on the migration process. Section 4 provides an open discussion about preliminary results. Section 5 concludes the paper with future directions.

## 2 Motivating challenges

The Semantic Web principles promote the use of graph-based models (e.g., RDF) to represent information and to facilitate its linkage and enrichment. In the context of cultural heritage data, many digital catalogs are still based on flat models (e.g., MARC format) and remain isolated from other sources (e.g., national repositories). Hence, the main challenges for the evolution of bibliographic catalogs are both related to the migration of legacy records towards new semantic models and to the enrichment of the migrated entities with other sources. This section presents important challenges from both tasks which should be considered for a successful Semantic Web adoption.

**Selecting a migration tool.** Metadata migration is a process which aims at offering a more expressive way to represent information by transforming the model of a data collection into another while preserving the completeness and the nature of the data [29]. The FRBRisation is a well-known migration process since it is a privileged way to adopt the new standards for representing bibliographic data [19, 15]. The process consists in interpreting bibliographic records to extract the relevant FRBR entities and relationships to generate a FRBR-based graph. Yet, after decades of discussions and contributions on this subject (see these recent surveys about FRBRisation [30, 8]), it is still complicated to select the relevant FRBRisation tool for a given catalog and to configure it properly.

**Tuning of the migration.** When FRBRisation is done automatically or semi-automatically, a tuning task is necessary to write the set of rules which should be applied on input records for extracting the FRBR entities (e.g., a rule may specify that the presence of the proper title field in a MARC record should trigger the creation of a Manifestation entity from FRBR). Yet, the writing task of rules becomes much more complex when dealing with records based on a flat structure of fields (i.e., key-value pairs) without clear semantics. For instance in MARC, each field does not necessarily contain atomic values, requiring sometimes to filter the relevant data during the migration. Moreover, some fields may be used regardless of the MARC specification forcing the creation of dedicated (and not reusable) rules. Those aspects, related to cataloguing practices or potential errors, make it impossible for a generic FRBRisation tool to deal with any catalog. Thus, the challenge of the tuning step is to better anticipate the specificities of an input catalog to ease the writing task of all migration rules.

**Building the migration model.** Another challenging aspect of the FRBRisation relates the models and systems used to build the migration rules. Existing FRBRisation solutions have proposed migration models (e.g., with XSL files), which generally consist in mappings between MARC fields and FRBR entities. The X3ML solution [18] proposes, for instance, a framework to build rich mappings between different models used for cultural heritage. Nevertheless, entity-centric mappings are sometimes not sufficient to express the rich knowledge that a bibliographic record may contain. The latter may include complex patterns

not only representing the description of a work but also its context and its place in larger bibliographic families [21]. For instance, a single work in a MARC catalog can appear in different records to express different realizations (translations, adaptations), different embodiments (digital, paperback), and several links to related works. Hence, while the migration rules of FRBRisation should deal with catalogs specificities, it must also ease the handling of complex bibliographic patterns to generate richer FRBR output [1].

**Evaluating the migration.** The evaluation of the FRBRisation is also a crucial challenge. Although specificities of each catalog require a manual validation of the migration results, it remains necessary to be able to evaluate any tools with criteria representing real-world cases. The TelPlus project [15] provided quantitative statistics of extraction and aggregation and also results of a survey about the search service. Some performance results were also available from the eXtensible Catalog experiments [3]. However, few metrics have been designed to evaluate the quality results of the migration. Furthermore, the datasets used in most attempts of FRBRisation have not been published making it impossible to reproduce the experiments. To tackle this lack of data, FRBR-ML [25] proposed to evaluate the FRBRisation by transforming back FRBR into MARC. Considering that some complex FRBR patterns can be implicitly expressed in MARC, we advocate that the domain still needs relevant metrics and datasets to ease the detection of weaknesses in the different tools.

**Entity Linking challenge.** The next phase, after the metadata migration of records, involves an enrichment step which aims at bringing new information to the migrated entities [13]. In the context of cultural heritage data, important challenges relate the linking task of entities with data from external repositories (e.g., Linked Open Data) [10, 12] and the deduplication task to detect and merge equivalent entities [4]. Initiatives like Europeana [23] raised important challenges and issues in the linking domain like dealing with multiple languages or managing all various forms of cultural items. Furthermore, other branches of research try to perform the enrichment task with larger datasets and multiple repositories, sometimes different than Linked Open Data to generate richer information [24]. Unfortunately, many enrichment experiments limit their action into making links between entities whereas cultural institutions expect real solutions for enriching their entities according to the bibliographic wealth. In the context of MARC data, the FRBR model, close to the triple-based standards of Semantic Web, is a real asset to enrich the patterns of bibliographic data [26].

**Data matching and merging.** The fusion of extracted data is a well known challenge in the enrichment task. It implies to get the correspondences between different entities extracted from different sources to be able to merge them. Yet, while major sources from LOD already provide *sameAs* links between them, finding correspondences between isolated repositories may require an extra Ontology Matching process [2]. This becomes even more important when involving

new unstructured sources (e.g., Websites, Blogs) to get more fresh information. Considering that bibliographic families become increasingly rich and complex, the fusion process has to evolve by taking into account the emergence of new semantic models to improve quality and to benefit from graph structures [9].

**Evaluating the enrichment.** The enrichment task with external sources should also be evaluated to confirm the relevancy of aggregated data, especially for the tuning of new semantic search engines. However, it can be really complex and long to evaluate aggregation tasks done on large amounts of data. Recent works from Europeana showed the efforts needed to evaluate a semantic enrichment process done on thousands of resources [14]. Moreover, crowd-sourcing becomes an interesting solution to evaluate both the relevancy of enriched information and the graphical interfaces of semantic search engines. Yet, considering the domain of cultural heritage, further experiments have to been done for evaluating the impact of new cataloguing models in integrated systems.

## 3 Pattern-based migration model

To deal with the different challenges presented in the previous Section, we propose to consider the migration and enrichment processes in a case-oriented way. A case represents a fragment pattern of the whole FRBR model we want to extract or enrich. In the context of digital libraries, such cases should be inspired by the different bibliographic patterns formulated for years by the community (e.g., adaptation in motion picture, aggregation of poems, addition of appendix).

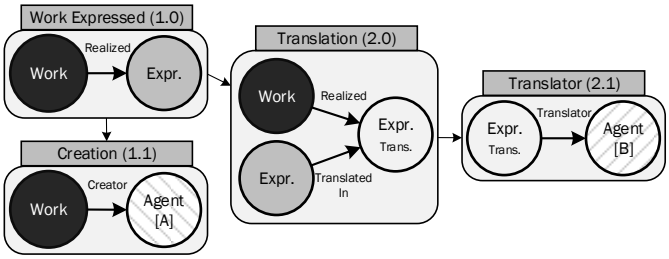


**Fig. 2.** Conceptual representation of the Case-Oriented Model with the bibliographic pattern of Translation in example

### 3.1 Characteristics of the model

The Figure 2 shows a conceptual view of a bibliographic pattern using a case-oriented representation. Each named box (e.g., Creation 1.1) represents a fragment pattern we want to obtain in our FRBR output of the migration. In this example, all the fragment patterns describe a bigger one which is the Translation pattern. Since the latter can be more complex than others patterns, the model

should be flexible enough to allow its design in sub-fragments. Now we present a list of recommendations for building the model:

- **The model should be an oriented graph**. It means that a node of the graph is a distinct case and an edge represents a relationship between cases. Each record to be migrated will follow the path of cases starting by root cases. Since each case can have its own conditions to be satisfied (e.g., presence of variant language to activate the case *Translation (2.0)*), a record may skip a whole branch of cases if the first case is not validated. For instance in Figure 2, if the case *Work Expressed (1.0*) is validated, then both cases *Creation (1.1)* and *Translation (2.0)* will be evaluated.
- **Each case should be built independently**. To improve readability, some distinct cases may involve a same entity to design a complete pattern. An example in Figure 2 shows the same Work (black circle) reused in three different cases, *1.0*, *1.1* and *2.0*. Yet, to prevent any deduplication error, the rule engine will only deal with one unique entity during the migration.
- **Cases should be specialization of other cases**. When a case is linked to another case, it means that the latter is a specialization of the first case. In such way, the specialization case inherits from the conditions and actions of the previous case. For instance, the case *Translator (2.1)* which deals with the translator Person entity will be evaluated only if the case *Translation (2.0)* is validated.

### 3.2 Expected benefits from the model

**Enhanced expressivity of the output.** Broadly speaking, this approach aims at facilitating the design of complex rules which are needed to handle any expected pattern in the migration, while keeping them easily readable. In practice, it consists of an abstraction layer that provides guidelines to build the whole FRBR model we expect without limitations. The modelling task can be done on two phases. First, the FRBR model is divided into bibliographic patterns then in fragment patterns (example in Figure 2). Once the graph of fragment patterns is set, the second phase consists of creating the mappings and conditions on each branch of the graph to fit the catalog specificities and to extract the relevant data to feed the FRBR properties. With this structured way to build the rules, we expect FRBRisation experts to implement more and more interesting bibliographic patterns rather than just being focused on the simplest cases (i.e., in FRBR, one Work related to one Expression, related to one Manifestation).

**More efficient tuning.** Our proposal is also an asset to reduce efforts for writing the migration rules (i.e., mappings and conditions upstream the fragment patterns). One of the most time consuming task for the expert of the migration consists in analysing the catalog in input and understanding its specificities. Hopefully, since the scope of each fragment pattern can be described clearly in our model, it becomes easy to write a set of functions to automatically detect if a specific case should be involved for a certain part of a catalog. For instance, a

MARC record with different languages, variant titles or translator relator code should lead to the activation of the Translation case (and all related fragment patterns) in our migration model. Thus, the already written cases can be reused for different catalogs thanks to such automated pre-analysis process which detects each case to use while interpreting the records.

### 3.3 Evaluation of the Migration process

As migration becomes easier with such case-oriented model, it remains necessary to evaluate the whole process. Yet in the context of FRBR, there are still few metrics and datasets available to perform a complete evaluation of a FRBRisation tool. Since both the pre-analysis of the input and the modelling task of FRBR output have a strong impact on the migration's quality, we need both *pre* and *post* FRBR metrics to evaluate the migration. By leveraging on our study on bibliographic patterns and catalog inconsistencies we now provide both datasets and metrics to perform a complete Benchmark of a FRBRisation solution [7]. We also did experiments with this Benchmark, BIB-R, on three recent tools.

## 4 Preliminary results and lessons learned

In this Section we present the results obtained so far via our FRBRisation prototype named CoM3T (for Case-oriented MARC Metadata Migration Tool). This tool implements our case-oriented approach presented in the previous Section.

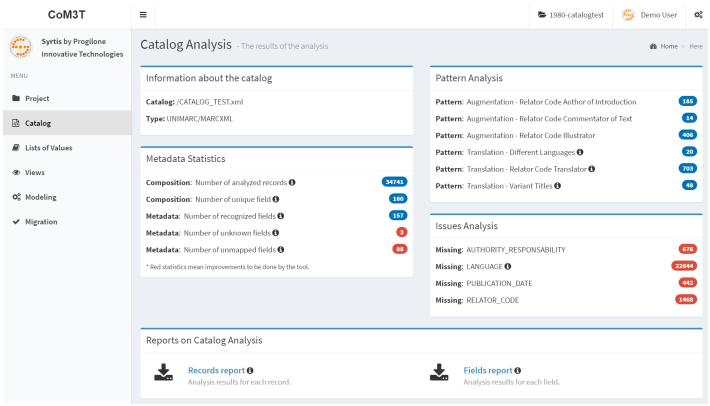


**Fig. 3.** Screenshot of the catalog analysis view in CoM3T

The tuning task is handled by an automated process which analyzes the input catalog of MARC/XML records. This analysis is based on the pre-FRBRisation metrics to extract all the specific interpretations that should be handled by rules. It also provides statistics on the catalog, details on the bibliographic pattern

detected and detailed reports for each distinct metadata used in the records. The figure 3 shows a web interface provided by our tool to ease the access to the analysis results.

The tool has been experienced on two real world catalogs of MARC records, 100,000 records from a public library and 400,000 records from a university hospital library. Each of catalog statistics of analysis were first generated, then the records were FRBRised using our tool and finaly, the results were loaded in an Integrated Library System, based on the FRBR model, to ease the navigation in generated entities. Those experiments helped us to make observations on the benefit of our approach:

- The pre-analysis results are a reliable source of decision to build specific rules or to clean part of the catalog before FRBRisation. It helps for communicating on the FRBRisation model and to reduce lacks of errors.
- The automated activation of cases from pre-analysis results allows us to provide a FRBRisation tool were the migration step can be launched without human intervention. Hence, with an automated deduplication phase, the whole FRBRisation can be done fully automatically. This is crucial when dealing with periodic flows of records.
- The case-oriented model, implemented in a web application, allows to draw overviews of the FRBRisation rules. Once again it eases the communication with actors of the migration.
- The processing time for applying the complete migration model on each record (even on large catalogs) is negligible using recent Java parallel processing features. The deduplication remains the most time consuming task.

These results showed us the limits and improvements to be done with such approach. Although the case-oriented model can be managed easily, some specific needs of the migration could not be handled by this approach. For instance, creating links between FRBR entities generated from different catalogs may imply an additional process, done independently of the extraction step. Improvements are also expected for the deduplication phase in which the quality of the matching task is hardly dependant on the completeness of input data and also from the way the migration model extracts such data. Hence, the fully automation of the deduplication can sometimes be impossible since it always requires specific adjustments. The web application for managing the FRBRisation should be evaluated in terms of usability to involve more non-IT specialists in the configuration of the process. Finally, the major expected contribution relates to the enrichment of FRBRised data with external repositories. Since many challenges related to semantic enrichment are well-known by the community, the case-oriented approach should be an interesting way to bring qualitative improvements in the context of bibliographic entities. The knowledge of patterns representing complex relationships between entities should refine the way the external sources are requested and should ease the fusion of extracted information.

## 5 Future directions

In this paper, we introduce a case-oriented model for the metadata migration and enrichment of bibliographic data. This approach brings new perspectives to ease the FRBRisation and enrichment of bibliographic data. We also present concrete implementation of our approach for the FRBRisation process and preliminary results of experiments. Thanks to our benchmark for FRBRisation tools, we promote improvements in migration process to create richer relationships between entities and to improve the search and navigation of users in cultural heritage collections. In the next steps of our research we plan to extend or approach to the extraction of new information from external repositories to complete or discover bibliographic patterns. The main idea is to improve the generation of queries and the selection and normalization of results by leveraging on the acquired knowledge of bibliographic patterns from FRBRisation. Our future work includes a theoretical formulation of our case-oriented approach and new experiments of knowledge extraction from different databases to enrich FRBRised data.

## References


1. Aalberg, T., Žumer, M.: The value of MARC data, or, challenges of frbrisation. Journal of Documentation 69(6), 851–872 (2013)
2. Bernstein, P.A., Madhavan, J., Rahm, E.: Generic schema matching, ten years later. Proceedings of the VLDB Endowment 4(11), 695–701 (2011)
3. Bowen, J.B.: Moving Library Metadata toward Linked Data: Opportunities Provided by the eXtensible Catalog. In: International Conference on Dublin Core and Metadata Applications. pp. 44–59 (2010)
4. Christen, P.: Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media (2012)
5. Coyle, K.: FRBR, twenty years on. Cataloging & Classification Quarterly 53(3-4), 265–285 (2015)
6. Crupi, G.: Beyond the Pillars of Hercules: Linked data and cultural heritage. JLIS.it 4(1) (2013)
7. Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., Lumineau, N.: BIBR: a Benchmark for the Interpretation of Bibliographic Records. In: Theory and Practice of Digital Libraries (TPDL) (2016)
8. Decourselle, J., Duchateau, F., Lumineau, N.: A Survey of FRBRization Techniques. In: Theory and Practice of Digital Libraries (2015)
9. Dong, X.L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., Zhang, W.: From data fusion to knowledge fusion. Proceedings of the VLDB Endowment 7(10), 881–892 (2014)
10. Dutta, A., Meilicke, C., Stuckenschmidt, H.: Enriching structured knowledge with open information. In: Proceedings of the 24th International Conference on World Wide Web. pp. 267–277. International World Wide Web Conferences Steering Committee (2015)
11. Faith, A., Chrzanowski, M.: Connecting RDA and RDF: Linked Data for a Wide World of Connected Possibilities. Pennsylvania Libraries 3(2), 122 (2015)
12. Haslhofer, B., Momeni, E., Gay, M., Simon, R.: Augmenting Europeana content with linked data resources. In: Proceedings of the 6th International Conference on Semantic Systems. p. 40. ACM (2010)

13. Lacasta, J., Nogueras-Iso, J., Falquet, G., Teller, J., Zarazaga-Soria, F.J.: Design and evaluation of a semantic enrichment process for bibliographic databases. Data & Knowledge Engineering 88, 94–107 (2013)
14. Manguinhas, H., Freire, N., Isaac, A., Stiller, J., Charles, V., Soroa, A., Simon, R., Alexiev, V.: Exploring comparative evaluation of semantic enrichment tools for cultural heritage metadata. In: International Conference on Theory and Practice of Digital Libraries. pp. 266–278. Springer (2016)
15. Manguinhas, H.M.Á., Freire, N.M.A., Borbinha, J.L.B.: FRBRization of MARC records in multiple catalogs. Joint conference on Digital libraries p. 225 (2010)
16. Meehan, T.P.: What's wrong with MARC? Catalogue and Index 174, 33–42 (2014)
17. Mercun, T., Svab, K., Harej, V., Zumer, M.: Creating better library information systems: the road to FRBR-land. Information Research 18(3) (2013)
18. 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: 19th International Conference on Theory and Practice of Digital Libraries, Poznań, Poland (2015)
19. Mitchell, E., McCallum, C.: Old data, new scheme: An exploration of metadata migration using expert-guided computational techniques. Proceedings of the American Society for Information Science and Technology 49(1), 1–10 (2012)
20. Pandey, S.R., Panda, K.C., Others: Semantic solutions for the digital libraries based on semantic web technologies. Annals of Library and Information Studies (ALIS) 61(4), 286–293 (2015)
21. Riva, P.: Mapping MARC 21 linking entry fields to FRBR and Tillett's taxonomy of bibliographic relationships. Library resources & technical services 48(2), 130 (2004)
22. Riva, P., Žumer, M.: Introducing the FRBR Library Reference Model. In: Ifla Wilc 2015. pp. 1–7 (2015)
23. Stiller, J., Petras, V., Gäde, M., Isaac, A.: Automatic enrichments with controlled vocabularies in europeana: Challenges and consequences. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, pp. 238–247. Springer (2014)
24. Suchanek, F., Weikum, G.: Knowledge harvesting in the big-data era. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 933–938. ACM (2013)
25. Takhirov, N., Aalberg, T., Duchateau, F., Žumer, M.: FRBR-ML: A FRBR-based framework for semantic interoperability. Semantic Web 3(1), 23–43 (2012)
26. Takhirov, N., Duchateau, F., Aalberg, T.: Linking FRBR entities to LOD through semantic matching. In: Theory and Practice of Digital Libraries. Springer (2011)
27. Tillett, B.: What is FRBR? A conceptual model for the bibliographic universe. The Australian Library Journal 54(1), 24–30 (2005)
28. Tillett, B.: RDA and the Semantic Web, Linked data environment. JLIS.it 4(1), 139–145 (2013)
29. Walkowska, J., Werla, M.: Advanced Automatic Mapping from Flat or Hierarchical Metadata Schemas to a Semantic Web Ontology. In: Theory and Practice of Digital Libraries, pp. 260–272. Springer (2012)
30. Zhang, Y., Salaba, A.: Implementing FRBR in Libraries: Key Issues and Future Directions. Neal-Schuman Publishers (2009)
31. Zhang, Y., Salaba, A.: What Do Users Tell Us about FRBR-Based Catalogs? Cataloging & Classification Quarterly 50(5-7), 705–723 (2012)