# Improving the presentation of library data using FRBR and Linked data

*When a library end-user searches the online catalogue for works by a particular author, he will typically get a long list that contains different translations and editions of all the books by that author, sorted by title or date of issue. As an attempt to make some order in this chaos, the Pode project has applied a method of automated FRBRizing based on the information contained in MARC records. The project has also experimented with RDF representation to demonstrate how an author's complete production can be presented as a short and lucid list of unique works, which can easily be browsed by their different expressions and manifestations. Furthermore, by linking instances in the dataset to matching or corresponding instances in external sets, the presentation has been enriched with additional information about authors and works.*

By Anne-Lena Westrum, Asgeir Rekkavik and Kim Tallerås

*2012, Code4lib Journal, Issue 16*

## Introduction

After years of delay, it seems like Oslo is finally getting its new public library. At least there are concrete plans for a new, large and innovative building. The plans, of course, primarily concern the physical space: How can the building contribute to modern information services, and which features of a modern public library should it enable and encourage? In connection to such questions, many interesting discussions are taking place. One of them deals with the traditional axis point of the library: The document collection.

While there is a rapid development going on in how we think about library buildings, their means and objectives, we are also witnessing a parallel *digital* revolution, pushing forward new thoughts and solutions on collection development and distribution. With the expansion of the Web, online catalogues have a new context that involves both opportunities and challenges (Coyle, 2010). The opportunities are related to the effective infrastructure for sharing and dissemination. Among other things, the challenges relate to the existing library standards for document description – metadata. These standards were made in a different technological era.

## The Pode project

The plans for a new library building and the discussions about library services affected by these plans have given the Oslo public library an indirect opportunity to examine how their metadata can be used in new contexts and in ways that contribute to better services. The independent *Pode* project[1], funded by *ABM-utvikling*[2], but located at the Oslo Public Library, has done exactly that. The project has, during the last few years, been experimenting with descriptive metadata related to mash ups, reference models such as FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998), new generations of OPACs and Linked Data. This work has led to at least one central insight: One cannot create better services, based on already existing metadata, than what the quality of the metadata will support. In this article we describe a subproject of Pode dealing with (NOR)MARC[3] records describing manifestations related to the Norwegian authors Knut Hamsun and Per Petterson.

## Finding the way through library hit lists

Knut Hamsun (1859-1952) is Norway's most prominent novelist and one of three Norwegian Nobel laureates in literature. His literary production includes about 30 novels, a few plays and collections of short stories, one collection of poetry and some non-fiction and biographical writings. Altogether Hamsun's production counts a total of 40 works; it is a bibliography that a library user should be able to browse easily. However, the image that meets the library user is quite different. In the online catalogue at Oslo Public Library, a qualified search for "Hamsun, Knut" as author will produce a list of 585 hits (as of November 11th, 2011). This is of course way too many hits to provide for an author who wrote 40 books. Notice that this is the result of an advanced qualified search. A more typical simple search, which is what most library users would try, provides an even longer list.

The problem is that the online catalogue doesn't distinguish between an author's different works and different versions of one work. In our list of 585 hits, as many as 63 correspond to different representations of one novel: *Hunger*. These would be different editions, different formats and translations into different languages. The users must of course be able to choose whether they want the book, the audio book or the movie, and they must be able to choose what language they want to read the book in, but to most users it is more disturbing than useful when the OPAC makes them choose between more than 20 different editions of *Hunger*in the Norwegian language[4].

## Library standards

The library catalogue has traditionally focused on describing physical objects. Each manifestation of a book is represented by a separate record, and there are no functional connections between records that describe manifestations of the same work. A library user, who searches the online catalogue for a particular title might therefore sign up on a waiting list to borrow one particular edition of a classic novel without realizing that numerous other editions of the same book are already available. Another user might end up not getting the book at all if he accidentally picked an edition that no longer has available copies. The Pode project has based its experiments on a hypothesis that library users are typically interested in finding a particular title, not a particular edition of that title, and that this interest is especially typical for fictional works, where different editions usually have identical content. From the perspective of the library system, the user should ideally be able to make a reservation for a title without having to choose between different editions. Of course, those who do care about editions should still have the opportunity to specify this, but why should everyone else be forced to pick?

As many have pointed out, the present library standards were developed prior to the web and the present infrastructure for production, distribution and utilization of metadata[5]. In addition, the standards that introduced library data to the electronic sphere were developed years before the invention of Entity Relationship (ER) models and relational databases (Thomale, 2010). This presents some challenges with implementing reference models like FRBR (that separates editions from works), which is based on ER-analysis and relationships not implemented in or between MARC records.

The MARC format embodies technical inscriptions and logic from the card catalogue, which it was developed to automate. Metadata in card catalogues were read and interpreted by humans, a feature that is continued in the MARC format, which dictates the making of (separate) records, largely consisting of human-readable text strings. This is of course a simplified description of library metadata practices. The process of making a MARC record is characterized by a complex interaction with cataloguing rules like AACR2 and ISBD, and most of the motives behind the text strings are to be found in such rules. In a Web context, we want machines to process the data and interpret them for us. Text strings must be absolutely consistent in order for machines to accurately interpret the data and create a useful presentation of that data.

In relational database and linked data environments the best practice doctrine is to avoid disambiguation by providing unique identifiers – respectively using primary/foreign keys and URIs (Berners-Lee, 2006; Codd, 1970). MARC records are lacking such explicit identifiers that would have helped our indexing tools and search engines to separate two authors with the same name, or maybe to merge these authors if they are likely to represent one

person on the basis of concurrent relationships to the same uniquely identified works. These kinds of challenges will of course continue when we want our machines to identify relations between documents, authors and different FRBR entities. Which books represent manifestations of a particular expression or a work? How do we identify works like short stories, when they are only described textually in a MARC note field? For most books published after 1970 we have ISBN numbers that identify manifestations, and we have names and titles that we can get computers to reason over, but without consistent catalogues and machine-readable identifiers, this is still a tough job. It takes at least a proper cleanup.

## FRBRization of MARC records

The Pode project used a tool developed by the Norwegian University of Science and Technology for automated FRBRization of MARC records (Aalberg, 2006). The tool uses XSL transformations on catalogue exports in the MarcXchange format to sort the data within bibliographic records based on which FRBR entities the data applies to.

Library catalogue records are mainly descriptions of manifestations; the individual record describes one particular edition of a published work, with manifestation specific information such as time and place of publishing, physical description, ISBN, etc. But the record will also contain information that applies to the expression and work this manifestation is related to. For example, the data that contain the name of the author and the original title of a book are pieces of information that describe the work entity, while information about the document's language and format say something about the expression. The FRBRization tool will identify which fields apply to which FRBR entities, and use this to divide each record into a work part, an expression part and a manifestation part, with FRBR relations between them. If the tool outputs identical work descriptions for two different MARC records, the records are assumed to describe two different embodiments of the same work. The tool will also produce group 2 and 3 entities, such as agents and subjects.

Initially the project ran the selected corpus of records (908 records describing manifestations of Hamsun and Petterson) uncleansed through the automated FRBRization system. The first attempt gave results that were far from perfect. This was mainly due to missing information in the MARC records and inconsistent cataloguing practice. To progress further in the experiment, the project had to clean up a considerable number of records in order to get data that were sufficiently expressive and consistent.

## The clean up process

In brief, the clean up process mainly consisted of identifying and adding original and uniform titles to records where this was missing, adding information about individual works and short stories collected in one volume, and setting indicators to distinguish between significant work titles and non-significant titles. In addition some time was spent correcting typos and errors. Altogether approximately 60 hours was spent cleaning up the productions of Knut Hamsun and Per Petterson, including the time used to determine the rules for correction.

One of the main jobs in the clean up process was to identify and improve records for translated works that either lacked the Norwegian original title completely, or only provided the original title in human-readable notes. Another job dealt with the adding of uniform titles in cases where the titles of non-translated titles differed from the original title. Due to Norwegian spelling reforms, several of Hamsun's works have been released with different titles at different times. For example, the short story *Paa tourné* (original title) has also been released with the spellings *På tourné* and *På turné*.

Another task was setting indicators to separate significant real work titles from non-significant titles. Non-significant titles are typically found in collections of an author's work, where the content has been put together and the publication has been entitled by someone else than the author himself. Publications such as these should not be listed as works in a FRBRized bibliography, although the individual novels or short stories contained in the collections should[6].

Examples:

a) 245 10 ‡aGrowth of the soil ‡cKnut Hamsun ; translated from the Norwegian by W. Worster

(English translation of an original Hamsun work. First indicator set to one means this is a significant work title.)

b) 245 00 ‡aTales of love and loss ‡cKnut Hamsun ; translated by Robert Ferguson

(English collection of short stories that were not originally published together. First indicator set to zero means this is a non-significant title.)

See the Appendix for more information about the cleanup process.

## Outcomes

While the first attempt at FRBRization identified 149 works by Hamsun, the list was further reduced to a number of 84 after cleaning up the MARC records. In the case of Per Petterson, we saw an increase in the number of works from 14 to 41, due to the adding of titles of individual short stories and essays to some catalogue records. The resulting lists of works corresponded almost exactly with the actual bibliographies of the two authors. The only exception being one work by Hamsun which was listed by the Norwegian national library's Hamsun bibliography, but missing in Oslo public library's collection. The clean up process thereby unintentionally provided us with a method for identifying missing works in the library collection. This is otherwise a tedious procedure when you are dealing with long lists of hundreds of manifestations.

## RDFization

The FRBRized datasets were converted to RDF, using XSLT[7] and a crosswalk between MARC fields and RDF predicates that was developed by the project. The crosswalk mainly used well-known vocabularies and ontologies like *Dublin Core metadata terms*, *Bibliographic ontology, Core FRBR, FOAF* and *SKOS*. But the project also constructed several more specific sub-properties to express our data more exactly than these vocabularies allow for. Later the project discovered that the RDA vocabularies[8] contain predicates that cover a lot of the more library specific information that needed to be expressed. In later revisions of the crosswalk some of these have replaced our own predicates.

See the example in the table below, or the complete crosswalk at http://www.bibpode.no/blogg/?p=1573.

| MARC | | RDF | | |
|---|---|---|---|---|
| **Field** | **Subfields** | **Subject Type** | **Predicate** | **Object** |
| 001 | | bibo:Document | pode:titleNumber | xsd:integer |
| 008/35-37 | | bibo:Document | dct:language | lvont:Language |
| 019 | a | bibo:Document | dct:audience | dct:AgentClass |
| 019 | b | bibo:Document | dct:format | pode:PhysicalFormat |
| 019 | d | bibo:Document | pode:literaryFormat | pode:LiteraryFormat |

| | | | | |
|---|---|---|---|---|
| 020 | a | bibo:Document | bibo:isbn | rdfs:Literal |

Once the data were converted to RDF, they were enriched with links to other datasets. Works were linked to instances in DBpedia and Project Gutenberg, while persons were linked to DBpedia and VIAF. In order to be able to sort a list of works chronologically, the project added information about date for first edition to the work instances. This information is not easily extracted from a MARC record, if at all contained. With this new and enriched dataset, the project was able to develop a web application that allowed an end-user to browse through the library's complete collection of these authors' books by choosing from a short list of works instead of searching through a flat list with hundreds of manifestations. Furthermore the application could give the end-user relevant information about authors from DBpedia, as well as links to digital full text versions in Project Gutenberg.

A simple web application was developed that allowed end-users to browse this part of the library collection, clustered as FRBR entities, with additional information provided from external sources made available through the linking of data[9].

## Summary

The increasing literature on metadata quality deals with criteria and principles that aim to maximize the utilization of metadata. Consistency (and coherence) are among such criteria (Bruce & Hillmann, 2004). Flexible characteristics, such as extensibility, modularity, the ability of refinements and multilingualism, are others (Duval, Hodgins, Sutton, & Weibel, 2002). Some also emphasize machine processability as an independent principle (Nilson, 2010). This literature has moreover in common the desire to create interoperability and metadata quality across communities (e.g. Chan & Zeng, 2006; Haslhofer & Klas, 2010;Hillmann & Phipps, 2007). This implies a generalization of the quality concept and discussions on public benefit beyond the circumstances of the certain metadata environments and standards.

In light of the development of the Web and modern principles of metadata quality, it can be tempting to immediately convert library data to a more linked data-friendly format like RDF. On the basis of what we have – the MARC records – it's not yet that simple. Nilsson (2010) presents two models of metadata harmonization. Vertical harmonization increases interoperability within a given set of standards. Horizontal harmonization contributes to interoperability between various standards not given in advance. The Pode projects tells us that a conversion of existing traditions to new standards can be problematic in relation to both models of harmonization. They also tell us

that a “vertical clean-up” helps a lot in order to make it easier to achieve horizontal effects.

The experiments have been performed on a limited set of data. It is hard to estimate how comprehensive the challenges would be when conducting conversions of larger data sets at the Oslo Public Library. The experiments conducted have shown, however, that a conversion in accordance with a reference model such as FRBR and principles for linked data provides some immediate benefits in the form of cleaner and more “browseable” results lists, and in opportunities to utilize external quality data – the metadata quality increases. These are experiences that may be important in new projects, particularly in light of the plans for a new, forward looking and modern library.

## About the authors

Anne-Lena Westrum has a degree in multimedia design and is working as a project manager at Oslo Public Library. She is currently studying interaction design and user behavior at Gjøvik University College, while managing two new user centered projects at the library.

Asgeir Rekkavik is a librarian at Oslo Public Library. He has been involved in several RDF and linked data related projects.

Kim Tallerås is a doctoral student at the Department of Archivistics, Library and Information Science at Oslo and Akershus University College of Applied Sciences. He is working with issues related to metadata and semantic interoperability.

## Acknowledgements

The Pode project would not have been possible without the support of the Norwegian Archive, Library and Museum authority and the contribution from Oslo University College in cooperation with the Norwegian Library Laboratory.

The article summarizes a 3 year long project with work efforts from an interdisciplinary team at Oslo Public library. The team wishes to extend a special thanks to Associate Professor Trond Aalberg at Norwegian University of Science and Technology and Magnus Enger at Libriotech for important contributions making the end result possible.

## Notes

[1] For more information about the Pode project, visit http://bibpode.no/?q=node/9

[2] ABM-utvikling is the former Norwegian authority for libraries, archives and museums

[3] NORMARC is the Norwegian dialect of MARC. Web edition:http://www.nb.no/fag/kompetansesenter/kunnskapsorganisering/dnk/normarc

[4] Pisanski & Žumer (2010) indicate that users prefer a more overall orientation into a bibliographic universe.

[5] See http://www.loc.gov/marc/transition/news/framework-103111.html for an interesting and important announcement from Library of Congress advocating a similar message addressing the MARC standard especially.

[6] Although relationships between MARC fields are implicitly expressed through a record, it is not easy to unambiguously and explicitly express relations between fields in cases where a document consists of several works.

[7] In later work of this type, we have used Ruby scripts with a YAML mapping file instead of XSLT. For details, see: https://github.com/bensinober

[8] http://rdvocab.info/

[9] http://bibpode.no/linkedauthors/

## References


Aalberg, T. (2006). A Tool for Converting from MARC to FRBR. *ERCIM News*, (66). Retrieved fromhttp://www.ercim.eu/publication/Ercim_News/enw66/aalberg.html

Berners-Lee, T. (2006). Linked data. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting.*Metadata in practice* (s. 203-222). Chicago: ALA Editions.

Chan, L. M., & Zeng, M. L. (2006). Metadata interoperability and standardization?: A study of methodology: Part I. *D-Lib Magazine*, *12*(6). doi:10.1045/june2006-chan

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*,*13*(6), 377-387. doi:10.1145/362384.362685

Coyle, K. (2010). Library data in a modern context. *Library Technology Reports*, *46*(1), 5-13.

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, *8*(4). Retrieved from http://www.dlib.org/dlib/april02/weibel/04weibel.html

Haslhofer, B., & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys (CSUR)*, *42*(2), 7. doi:10.1145/1667062.1667064

Hillmann, D. I., & Phipps, J. (2007). Application profiles: Exposing and enforcing metadata quality.*Proceedings of the 2007 International Conference on Dublin Core and Metadata Applications: Application profiles: theory and practice* (p. 53–62). Retrieved from http://portal.acm.org/citation.cfm?id=1344591.1344601

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: Final report*. München: K.G. Saur. Retrieved fromhttp://archive.ifla.org/VII/s13/frbr/frbr.pdf

Nilson, M. (2010). *From interoperability to harmonization in metadata standardization: Designing an evolvable framework for metadata harmonization*. Stockholm: Kungliga Tekniska Högskolan.

Pisanski, J., & Žumer, M. (2010). Mental models of the bibliographic universe. Part 1: mental models of descriptions. *Journal of Documentation*, *66*(5), 643-667. doi:10.1108/00220411011066772

Thomale, J. (2010). Interpreting MARC: Where's the bibliographic data? *Code4Lib Journal*, *2010*(11). Retrieved from http://journal.code4lib.org/articles/3832

## Appendix – Clean Up and Corrections

The main job was to ensure that all the translated records (both due to foreign languages and Norwegian language reforms) contained the original work title in the 240 field. Where the 240 was missing, it was added automatically based on information in note field 574, a NORMARC-specific field containing information about original title. Where this note field was missing or the automatically conversion failed, the title was applied manually.

Another time-consuming part of the job was to identify and determine actual (original) titles of significant works in the 245 and 740/700‡t fields and set the appropriate indicator according to the corrections rules.

Dealing with the second indicator in the 740 field, the project chose to use the value 2 only when the field contains a significant work title. For all other titles the indicator is set to 0, even if they are analytic entries. Even if this practice does not conform the convention of using 700‡a + ‡t for analytical title entries, this was decided upon to be the most efficient approach to detect analytical work published in an unambiguous way.

Based on the results from the first attempt of FRBRization, corrections in the 908 records for Hamsun and Petterson included:

- Correcting the language code in 008 in 5 records
- Added uniform title (or “original title” in precise accordance with NORMARC terminology) in 240 fields in 85 records and correcting typos of existing 240 fields in 24 records
- Correcting typos in 245‡a (or wrong ISBD syntax) in 6 records
- Correcting the first indicator in 245: Before the correction 137 records had indicator 1 = 0 or blank, while indicator 1 = 1 was used in 774 records. After correction the distribution was 263 – 651
- Correcting the 700 fields. In the original records, we found 948 700‡a and 545 700‡t fields. In the corrected records the numbers are reduced to respectively 917 of 700‡a and 481 of 700‡t. The change is due to a more systematic use of 740 fields in all records that have the same author (which is registered in 100).
- Changing the second indicator in the 700 field in order to clarify whether an entry is a unique work or not.