Digital services in a new digital library – new ways of presenting the library catalogue

Asgeir Rekkavik
Oslo Public Library, Deichmanske bibliotek
Oslo, Norway
E-mail: asgeir.rekkavik@kul.oslo.kommune.no

Kim Tallerås
Oslo and Akershus University College of Applied Sciences
Oslo, Norway
E-mail: kim.talleras@hioa.no

Anne-Lena Westrum
Oslo Public Library, Deichmanske bibliotek
Oslo, Norway
E-mail: anne-lena.westrum@kul.oslo.kommune.no

Meeting: 80 — *Inspired moments in cataloguing* — Cataloguing

Abstract:

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 applied a method of automated FRBRizing based on the information contained in MARC records.

The project also experimented with an RDF representation of MARC records to demonstrate how an author's complete production can be presented as a short and lucid list of unique works. The outcome was an inspiration to pursue a fully converted RDF representation of the library catalogue.

Introduction

The plans for a new library building in Oslo, Norway 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 Pode 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.

The Pode project ended a year ago, but the results and findings have now been taken further in a project to create public services available for the library users. The project is based on the RDF representation of the MARC records and the goal is to create an interactive station that presents the library catalogue in a new way. The project will connect the library catalogue with external data sources and present it in a visual and intuitive way through a touch screen based interface. We are also working on a semantic collection of book recommendations connected to the RDF catalogue records.

The Pode project

The Pode project worked with new ways of using metadata found in library catalogues, in particular ways of mashing the catalogue content with other "mashable" resources. In addition, the project also looked into the possibilities and challenges the web technologies provide in relation to today's systems and practices.

The project used traditional protocols provided by the ILS vendors as well as converting and finetuning data for frbrization in order to provide Linked library data "with a meaning".

The project also experimented with an RDF representation of MARC records to demonstrate how an author's complete production can be presented as a short and lucid list of unique works, which in hand 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.

Our second central insight was: An RDF representation of our library catalogue gives us a more workable dataset with lots of more opportunities to create services for our users than the traditional MARC datasets.

FRBR and FRBRization – 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 568 hits (as of May 7th, 2012). 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 568 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.

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. 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.

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. 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 another Norwegian author, Per 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[1].

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[2] 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[3] 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 complete crosswalk at http://www.bibpode.no/blogg/?p=1573.

[1] 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.
[2] 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
[3] http://rdvocab.info/

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[4].

The RDF representation of the two authors instantly gave us a more workable dataset. The experience was so positive, that we decided to convert the whole library catalogue to see if this could be the answer to our fundamental wish: To create better services for our users.

Linked data and RDF as a basis for new services

In connection with the "Book recommendations" and "Active shelves" projects at the Oslo public library, a lot of work has been put into converting the library catalogue into the RDF format. The effort we have put into this field at this time is far more detailed and thorough, and provides us with far more possibilities than the experiments we did back in the Pode project. The catalogue at Oslo Public Library consists of approximately 420 000 records, dating back from the late 1970s.

RDF is a format that makes library catalogue data truly machine readable. Where machine readability in connection with the MARC format means that computers can read, store and process the *characters* in a catalogue record, machine readability in RDF means that a computer can read the actual *meaning*, or semantics, of our data. In brief, converting library data into RDF is about "translating" catalogue records into sets of simple statements, where we use unique identifiers for both those things we want to say something about, as well as the information we want to give about those things

Are we able to think new thoughts without forgetting our old skills? We are not the first library in the world to convert library data into RDF. But while a lot of the work that has been done in this field focus on extracting and converting a small subset of essential information from the catalogue records, we have tried to take "everything" with us. The effort of library cataloguers is thorough and priceless, and anybody who has worked in a library will know that it is critical that we can access more information about our documents than title, name of creator, subject and ISBN. We have tried to keep as much of the information from the MARC records as possible, and to express it as precisely and correctly as possible. We have systematically worked our way through the Norwegian NORMARC standard, field by field, and have made our desicions on how to interpret the data and how the contents could be expressed in RDF.

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

See our complete conversion script that converts binary MARC to RDF by YAML mapping at https://github.com/digibib/marc2rdf

At the same time it is important that an RDF version of the catalogue doesn't simply become another dataset that says exactly the same, only in a different language. Cataloguing rules and the MARC format were designed several decades ago, to achieve the best results possible within certain limitations, and some of these limitations aren't as relevant today as they were when these standards were made. Furthermore, many of the rules are characterized by the dual function of both describing documents as well as localizing them. If the old rules and the old ideas are brought into the new format, the old limitations will accompany them. The RDF format gives us possibilities that MARC lacks, and it is important that we take advantage of this.

One particular example: What does it mean when a catalogue records states that Karl Marx is the author of a "The Communist Manifesto", while Friedrich Engels is the co-author? This distinction between functions doesn't really have anything to do with the actual functions of the two persons. Actually we don't really have an author field in MARC at all; there is the "Main entry" field. The function of this field is to say something about where a document should be located in a collection. A physical document can only be in one place, therefore it can only have one main entry. Marx' name is listed before Engels' name on the title page; therefore he alone gets the honour of being stated as author of the book, while Engels' part is reduced to that of "co-author".

When we express this information in RDF, we don't have to care about this kind of artificial and non-functional distinctions. Karl Marx is an author of the book and so is Friedrich Engels. There is no conflict between these statements.

Another example is when information in catalogue records is ambigious and context dependant. Take this added entry as an example:

```
*700 0 $a King, Stephen
       $d 1947-
       $e forf.
       $j am.
       $t Rita Hayworth and Shawshank Redemption
```

What is the connection between the novella "Rita Hayworth and Shawshank Redemption" and the catalogued document in this case? That all depends on the context. If the catalogued document is a book, the statement most likely means that the novella is part of the document. If the catalogued document is a DVD disc, it probably means that the movie is based on the story told in the novella. Correct interpretation of the data demands that one know something about texts and movies and how they can relate. In other words, it demands a human reader.

Both examples tell us something about limitations to what we can express in the MARC format. The first example says something about how concepts such as main entries and added entries aren't primarily something that has been designed to describe books, movies or music. They are designed to organize physical collections and choosing ways a library user should be allowed to look up in a card catalogue. The second example has to do with the degree of machine readability. When correct interpretation of the data depends on human common sense knowledge, the data can hardly be said to be truly machine readable.

SPARQL is the query language we use to access RDF data. We can do any ordinary catalogue search with the SPARQL language, in addition we get a lot extra.

The CCL library search provides us with many ways to combine queries to make advanced and specific searches. But there is one limitation you can never escape: No matter how you construct your search, what you get in return will be catalogue records. You can ask any question you like, but only as long as it starts with the words "Which books…"

You cannot ask for a list of topics covered in documents in a particular language, or a list of authors that have written about a particular topic. To answer questions like these, you would have to find those *books* that the library has in that language or on that topic, and then perform the tedious process of going through that list to identify the different topics or authors.

Another limitation with traditional library searches is that they can't contain unknown entities. For example you cannot construct a CCL search so that it returns all books by the author of a given title. You have to look up the known title, find the name of the author and then make another search for books by that author. This might not be very demanding, but what if the information you need depends on more than one unknown entity? A question such as "Which novels can I find in the library, that I can also find an adaptation of on DVD?" is hopeless to answer with the traditional library search tools, unless you have plenty of time. With the aid of SPARQL you can ask:

Give me all pairs of documents, doc-1 and doc-2, so that:

doc-1 is a movie on dvd
doc-2 is a book containing a novel
doc-1 is based on the work X
doc-2 is a manifestation of the work X

Closing remarks

We will continue our work with the RDF representation of our catalogue. The next step will be to implement a search to see if the results will be any better than we get from our OPAC at this time. You can follow our progress at http://digital.deichman.no/

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.

References

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