# 68th IFLA Council and General Conference
# August 18-24, 2002

| | |
|---|---|
| **Code Number:** | **053-133-E** |
| **Division Number:** | **IV** |
| **Professional Group:** | **Cataloguing** |
| **Joint Meeting with:** | **-** |
| **Meeting Number:** | **133** |
| **Simultaneous Interpretation:** | **Yes** |

## Data mining MARC to find: FRBR?

**Knut Hegna**
University of Oslo Library
Oslo, Norway
E-mail: Knut.Hegna@ub.uio.no

**Eeva Murtomaa**
Helsinki University Library
Helsinkio, Finland
E-mails: Eeva.Murtomaa@helsinki.fi

## *Abstract:*

*This paper summarizes a project where MARC data from two national bibliographies was analysed in the light of the data model presented in the FRBR study from IFLA. During the project we found that even though the information in the MARC records holds attributes relevant for identifying the work, expression and manifestation entities, the accuracy and formal syntax are too simple to be properly handled by programs. Some of the results may be used to present better hit lists in OPACs. The project presented two suggestions for an OPAC user interface based on the ideas of the FRBR study and on the results of the project.*

The complete report is located here:
http://folk.uio.no/knuthe/dok/frbr/datamining.pdf

### FRBR and MARC

The first question we asked ourselves was: Can we find the FRBR structure in the existing MARC records? If we look at a single record, we see that there is information about the work, the expression and manifestation:

```
*008880325 no esp
*02000 $a 82-991075-2-0 $b h. $c Nkr 60.00
*04110 $a espnor *08200 $a 839.822[S]
*10010 $a Ibsen, Henrik $d 1828-1906
*24510 $a Puphejmo (1879) $c Henrik Ibsen ; tradukis:
          Odd Tangerud ; lingve kontrolita de Esperantista
          Verkista Asocio (EVA)
*26000 $a Hokksund $b Eldonejo Odd Tangerud $c 1987
*26900 $a [Drammen] : Tangen-trykk
*30000 $a [1], 57 s. $c 24 cm
*50000 $a Originaltittel: Et dukkehjem. -
          Originalutgave: København : Gyldendal, 1879
*99100 $a Tangerud, Odd
```

This is a record from the Norwegian national bibliography describing a document containg an Esperanto version of the the famous play *A doll's house* by Henrik Ibsen. We find the work title in the 500 note field, the relation to the creator of the work in the 100 field. This information may identify the work in this record.

Furthermore, we find information about the expression through the language codes in the fields 008 and 041, and a relation to the person responsible for the translation partly in the statement of responsibility and in the field 991.

Information about the manifestation we find in the fields describing the document in hand: 245, 260, 269, 300 and others.

So the answer to our question is:

**Yes**: WHY? Because a bibliographic record may

- describe both the work and the manifestation
- contain traces of the expression
- contain some relations in the added entries, notes and subject descriptions

Elements of the FRBR model are to some extent present in the MARC record!

But, unfortunately, the answer is also **No**. WHY?

- Because the cataloguing rules are well suited to the card catalogue and printed bibliography, not to the FRBR model
- central information is often recorded in a way more suitable for the human mind and eye, than a computer.

Looking at one single record is of no use. We decided to analyse sets of records generated as hit lists searching in the Finnish and Norwegian national bibliographies. Our scope was limited to author search.

We put a question mark at the end of the project title. We did not know what results we would gain. We felt the investigation should be open-ended. At least we expected to identify some problems with the cataloguing rules and the MARC format in this respect.

We started out looking at the tables in the FRBR study showing which attributes of the entities are important for identifying the work, the expression and the manifestation. We then tried to map information from the MARC records to these attributes.

| FRBR attribute | FRBR value | NORMARC | FINMARC | Selected |
|---|---|---|---|---|
| title of work | high | | 241 $a | yes |
| | | 500 $a | 500 $a | yes |
| | | 505 $a | 505 $a | yes |
| | | | 248 $h | yes |
| | | 240 $a | 240 $a | yes |
| | | 245 $a | 245 $a | yes |
| relation to person responsible | high | 100 $a | 100 $a $h | yes |
| | | 70010 $a | 70010 $a $h | |
| intended termination | high | ? | ? | no |
| form of work | moderate | interpretation of Dewey | 008 pos.24-27, 29-30, 33-34 | no |

Table 1: *Attributes identifying the work entity in the MARC formats*

To get the work title we look for uniform titles or original titles. We first check if there is a 241 field (original title). The Norwegian national bibliography does not use this field, but the Finnish does. If no 241 field is present, we check the 500 and 505 fields for the text *Original title* in the two languages: *Originaltittel*:, *Orig.tit.:* (abbreviated) or *Originaltitler*: (plural) in Norwegian records and *Alkuteos*: or *Alkuteokset*: (plural) in the Finnish records. The original titles always follow this text. If no original titles are found in the 50X fields on the basis of this test, we move on to other title fields in this sequence: 248 $h (title proper of a part in Finnish records) and 240 $a. The last possible solution is to pick the original title(s) from the field 245.

Usually the titles mentioned in the 505 field are repeated in the title added entry fields (745 $a in Finland, 740 $t in Norway), but there is no way to decide whether the data in these fields are original titles or not. Some times the fields contain the original titles, some times other title information. Neither 745 nor 740 contain information qualifying the title. In many cases these 74X fields are added entries concerning another work than the work in question, but are related to it in some way.

Together with relation to the creator of the work we felt we could identify the work or works in each record. On this basis, with these data, we could collocate the identical works found in several records and differentiate them from other works. This done, we used other data to differentiate/collocate different expressions, identified by language code and translator. Other data was used to identify the manifestations.

Some results:

| **Author** | **Number of records** | **Number of work id-s** | **Number of unique work id-s** |
|---|---|---|---|
| Ibsen, Henrik (n) | 744 | 914 | 220 |
| Wassmo, Herbjørg (n) | 149 | 159 | 19 |
| Gaarder, Jostein (n) | 237 | 237 | 14 |
| Solstad, Dag (n) | 92 | 93 | 35 |
| Kunnas, Mauri (f) | 130 | 133 | 41 |
| Jansson, Tove (f) | 576 | 595 | 92 |
| Linna, Väinö (f) | 168 | 168 | 20 |

Table 2: *Results of using the work reduction procedure on records from the Norwegian (n) and the Finnish (f) national bibliographies*

We see that there were 744 records where Henrik Ibsen appears as author or as a person added entry. We were able to extract 914 work identifiers from these records and from these 914 identifiers we identified 220 unique works. This looks nice, but Ibsen has written 26 plays and some collections of poems. Why then, 220 works?

There were records which lack information or the information is wrong: collections and selections of plays where the original titles or uniform titles were not identifiable even though they were present (12); 33 records where subtitle was included in the title proper, giving rise to 8 unique works; 3 records with misprints in the original title; 14 records with more or less modern spelling differing from the original title; 10 records with no original titles at all resulting in work headings with foreign language titles.

These problems are discussed in our report.

**The user interface**

We found some structure. The second question we then asked ourselves was: how could we benefit from this structure when presenting the hit lists to the user?

We first looked into the FRBR tables to find out which attributes were important for selecting between like entities and mapped those attributes to MARC tags.

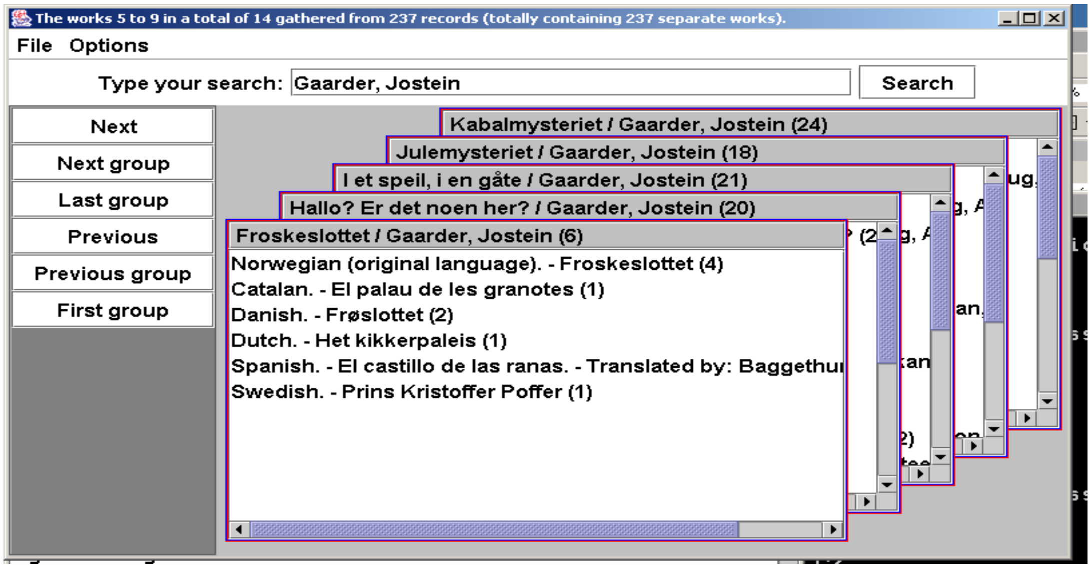
To make selection easier it is important also to decide what kind of ordering is most appropriate. The sorting (filing) principle should be easily identifiable by the user and should vary according to which entity level is presented.

We believe that the hit list in a search should appear according to the search performed and the results themselves. When you search for a distinct person, the hit list should consist of his or her works in some order, alphabetical perhaps or chronological or by a list of the different functions he plays related to the entities (author, illustrator, translator).

In the first interface the hit list is presented by overlapping cards. The top of cards may all be seen as one horizontal axis going into the screen of a two axis system and the card face as another - vertical - axis differentiating the information under the top heading of each card.

The two axis might be *authors - works*, *works - expressions*, *expressions - manifestations*, *subjects - works*. This would depend on the search performed. If the user searches for a specific author and there are several hits, the horizontal axis will list the person`s names, and the vertical axis the works of each person. If, on the other hand, the search results in only one hit, the horizontal axis will contain the works of the author, and the vertical axis the various expressions.

This is shown with the works of *Jostein Gaarder* along the horizontal axis and the expressions along the vertical axis under each work. The number of expressions identified for each work is given in parenthesis at the end of the work heading. The number of manifestations for each expression is found at the end of the expression information.


The works 5 to 9 in a total of 14 gathered from 237 records (totally containing 237 separate works).

Type your search: Gaarder, Jostein

Kabalmysteriet / Gaarder, Jostein (24)

Julemysteriet / Gaarder, Jostein (18)

I et speil, i en gåte / Gaarder, Jostein (21)

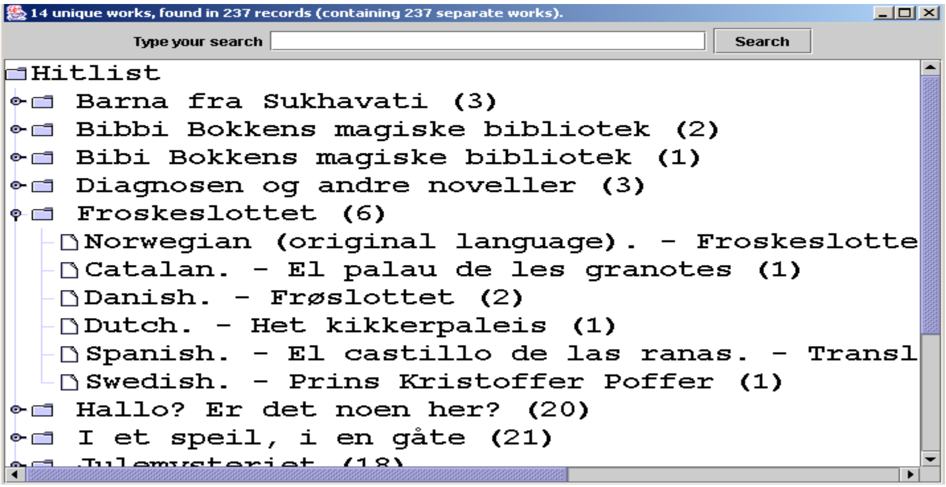
Hallo? Er det noen her? / Gaarder, Jostein (20)

Froskeslottet / Gaarder, Jostein (6)

Norwegian (original language). - Froskeslottet (4)
Catalan. - El palau de les granotes (1)
Danish. - Frøslottet (2)
Dutch. - Het kikkerpaleis (1)
Spanish. - El castillo de las ranas. - Translated by: Baggethui
Swedish. - Prins Kristoffer Poffer (1)

The user might select a specific expression thereby initiating a separate window containing all the manifestations under this expression. The manifestations of the chosen expression are sorted chronolocially, and the publishing year is presented first for each record. The window contains buttons with various functions such as *print*, *save* and *order* and the window does not disappear unless the user explicitly closes it. This means that the users can keep as many sets of manifestations as they want, all in separate windows.

The second user interface presents the works as nodes or branches of a hit list in a tree-like structure. The works appear sorted alphabetically according to the original title and the number in the end of the title indicates how many expressions are identified under this work.


14 unique works, found in 237 records (containing 237 separate works).

```
Hitlist
   Barna fra Sukhavati (3)
   Bibbi Bokkens magiske bibliotek (2)
   Bibi Bokkens magiske bibliotek (1)
   Diagnosen og andre noveller (3)
   Froskeslottet (6)
     Norwegian (original language). - Froskeslotte
     Catalan. - El palau de les granotes (1)
     Danish. - Frøslottet (2)
     Dutch. - Het kikkerpaleis (1)
     Spanish. - El castillo de las ranas. - Transl
     Swedish. - Prins Kristoffer Poffer (1)
   Hallo? Er det noen her? (20)
   I et speil, i en gåte (21)
   Julemysteriet (18)
```

The work nodes are expandable with the expressions as leaves. The expressions are sorted in the same way as for the interface previously described, original language on the top and then languages alphabetically.

These leaves are active, a click initiates a manifestations window as before.

**Summary**

One of the main reasons for the noise we have experienced through our experiments stems from the unqualified use of the 700 field and the way the systems index this information together with the 100 field regardless of the real function the person has.

We feel that function in the 700 field should be mandatory in the cataloguing rules and that the list of functions should also be expanded. The systems must use function to present more structured hit lists for the end user.

This would make it possible for systems to present all functions a person might have in the database, making it possible for the end users to choose either all or some of the functions in their bibliographical navigation.

Our investigation also showed that it would be an advantage for the analysis if original titles could be entered in a more consistent way, in separate, repeatable fields.

Language codes are one of the most important attributes to identify the different expressions of a specific work. It is also important in order to identify whether a manifestation is a translation or not.

Normalization of data is a linking device in itself. Using authority files in the bibliographic environment helps to establish the navigational structure, both by collocating entities and differentiating between them. The most common authority files used are the names of persons and corporate bodies. Our work shows that there is also a need for work authorities to be able to collocate the same work under one heading.

Thank you for your attention.