FRBRization: using UNIMARC link fields to identify *Works*

Manolis Peponakis,
Michalis Sfakakis and
Sarantos Kapidakis
Laboratory on Digital Libraries and Electronic Publishing
Department of Archive and Library Sciences
Ionian University
Kerkira (Corfu), Greece

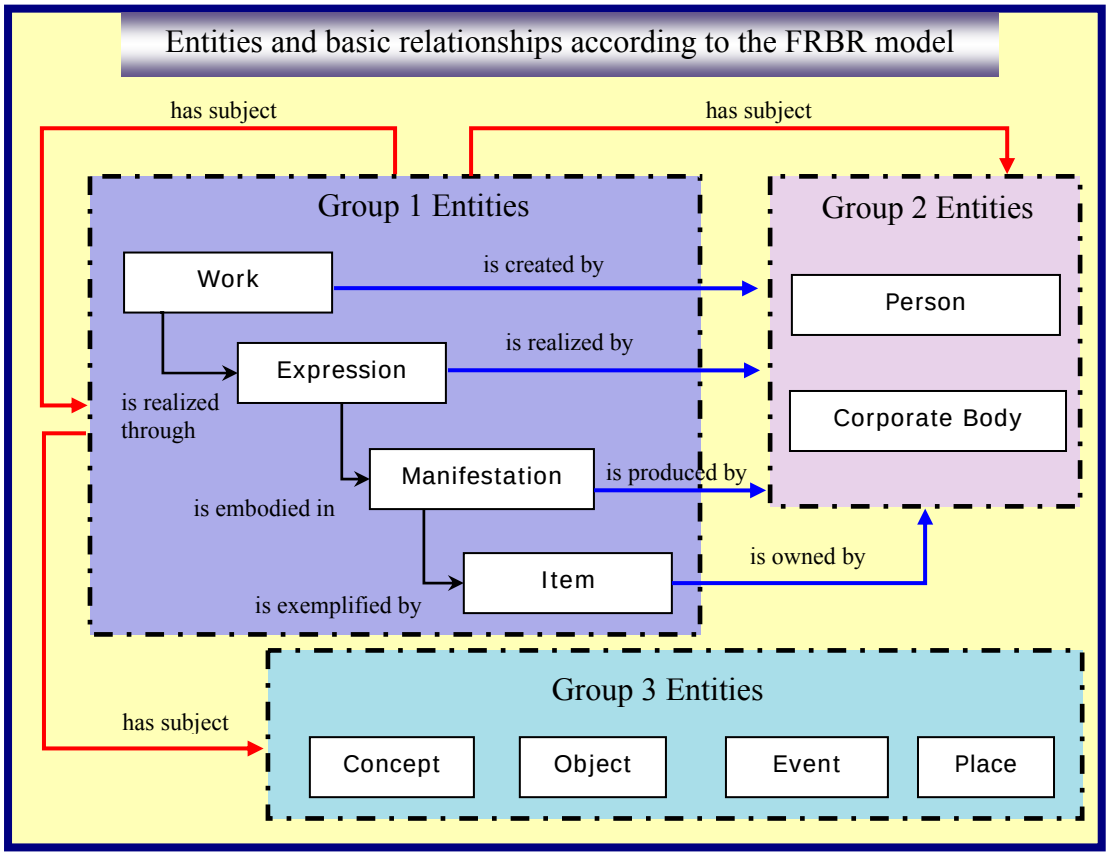
Meeting: **187 — *Advancing UNIMARC: alignment and innovation* — IFLA UNIMARC Programme (UNIMARC)**

Abstract:

The main objective of this study is to amalgamate the MARC 21 FRBRization practices with UNIMARC format semantics and to highlight some differences between them in the context of FRBRization. The main focus is to examine the possibility of using the UNIMARC link fields in order to identify the Functional Requirements for Bibliographic Records (FRBR) Work entities. In our approach we suggest that all records linked with 45X fields may belong to the same Work with the record which contains these fields. As a test set of this approach we used a sample of records of ancient Greek authors from the Union Catalogue of Hellenic Academic Libraries.

FRBR

Functional Requirements for Bibliographic Records (FRBR) is a conceptual, entity - relationship model developed by IFLA. Figure 1 below shows a graphical representation of the basic relationships between the entities.



FRBR is "an entity-relationship model as a generalized view of the bibliographic universe, intended to be independent of any cataloguing code or implementation" (Tillet, 2004). They are neither a metadata schema nor cataloguing rules. The Resource Description and Access (RDA) rules which are the successor of the AACR are a cataloguing code which implements the FRBR.

Since the focus of this paper is on the first group of entities, Figure number 2 attempts to give a simple example of their meaning.

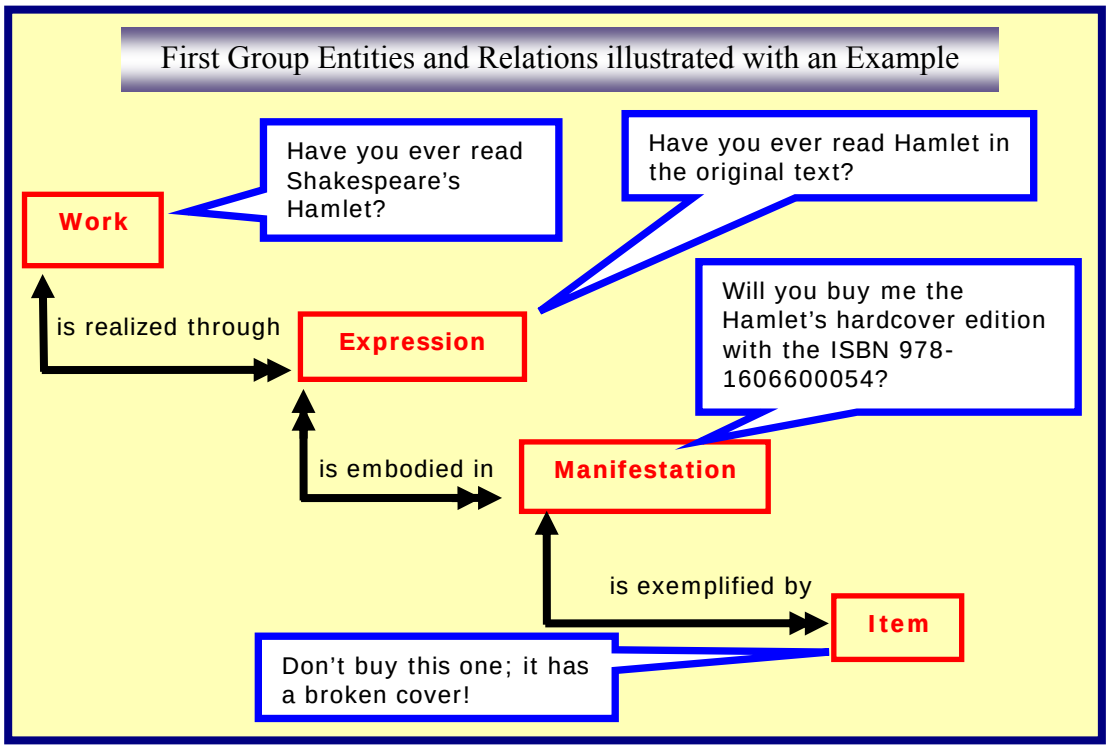


Figure 2: First Group Entities and Relations illustrated with an Example (Peponakis et al, 2010)

FRBRization

It is widely accepted that traditional catalogues have reached their limits, and, as suggested in Yee (2005: p.77), it is essential to proceed to "more intelligent use of our millions of existing MARC 21 bibliographic, authority, and holdings records in order to improve system design and to FRBRize OPAC displays and indexes". Thus, libraries ought to develop tools which will be effective amongst heterogeneous collections and metadata schemas (Naun, 2010: p.333). The FRBR offer a contemporary perception of the bibliographic data, but, as Rajapatirana (2005) states, "*re-cataloguing is not an option*". The main challenge for libraries is, therefore, the use of existing bibliographic records in order to provide value added services. This decision leads to inventing methods which will allow the reconstruction of existing data to new formats.

FRBRization is the process of searching and synthesizing FRBR entities using records prior catalogued – encoded in other encoding schemas. Babeu (Babeu, 2008: p. 17) reports very accurately that the terms *FRBR catalog, FRBRized system, FRBR implementation* are used interchangeably in order to describe the process but without them bearing a clear meaning. Babeu herself prefers the term "FRBR Inspired catalog" to describe the FRBRization process in the context of the Perseus Project.

Starting point for any FRBRization effort is the identification of the bibliographic records which represent a *Work* and then the identification within this group of the potential *Expressions* and *Manifestations*. The identification of *Works* is the most critical step because it engages the whole database and defines all subsequent steps. Several "keys" are produced using the individual bibliographic records and comparing them in order for the clustering to be successful. Same key means same "*Work*"[1]. According to both relevant bibliography (Aalberg 2006, Freire et al 2007, LC FRBR display tool) and what the FRBR define as *Work*, there are three essential information which a key should incorporate, namely the author of the *Work*, the title of the *Work* and the type of the material (e.g. motion picture or text) in which the *Work* can be expressed.

The generation of group 1 FRBR entities is based on author-title keys. Two methods can be applied for keys' generation. In the first case the data which constitute the keys are taken directly from the bibliographic records. In the second case an Authority File is used as a mediator. A graphical representation of the process follows in Figure 3 where the dashed line designates the mediation of an Authority File for the keys' generation[2].

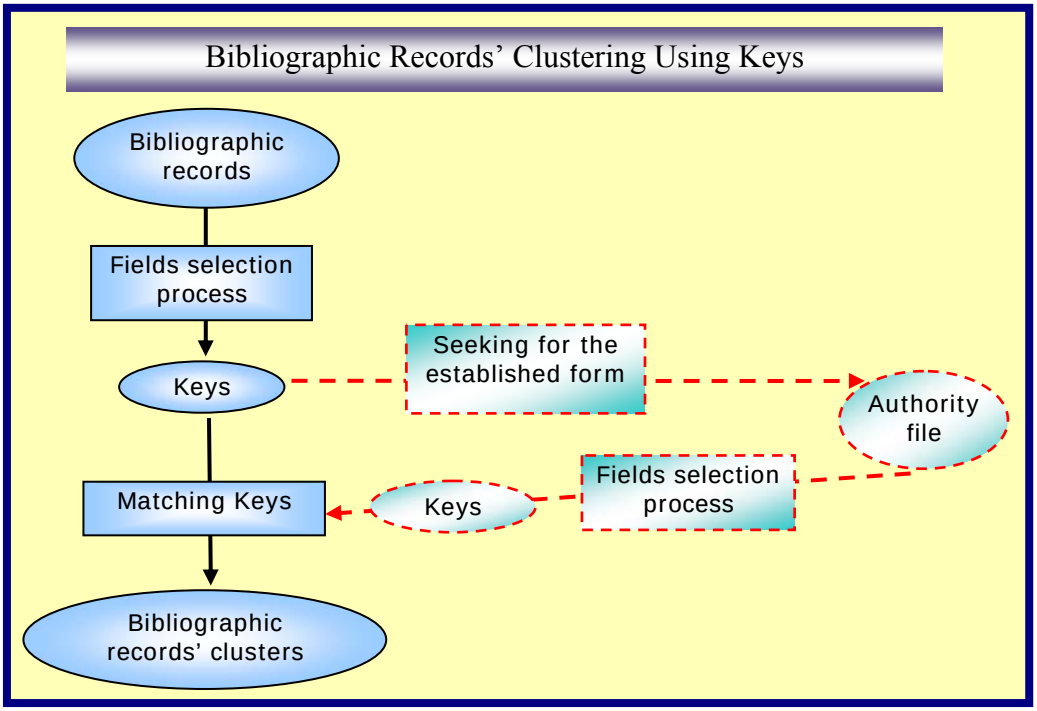


Figure 3: Process of Bibliographic Records' Clustering Using Keys (with or without using an Authority File) (Peponakis et al, 2010)

[1] This phrase is disputable considering that, in fact, different key does not necessarily mean a different *Work*. The difference between keys could be measured using a variety of similarity measures setting a limit above which the two records would be considered as belonging to the same *Work*.

[2] OCLC's algorithm includes the Authority File but LC tool (http://www.loc.gov/marc/marc-functional-analysis/tool.html) does not.

The benefits of the second approach are obvious because it offers the possibility of deriving extra information from the Authority File. Thus it is possible to match different linguistic representations of the same entity, as in the case of "Aristophanes", also appearing as "Aristofanis" and "Aristophanis".

Key parts: UNIMARC and MARC 21

For the construction of the keys two elements are common in all cases, i.e. Title and Author. Further specification can be achieved using the Type of Record. OCLC's algorithm does not include such information -and creates "FRBR *Work* sets", instead of *Works*- while the LC FRBR display tool does consider the Type of Record. In our approach the Record Type is taken into consideration, as well. We, therefore, construct the keys using three parts. The first part is the Author, the second part is the Title and the third is the Type of Record. For all of them there are some differences in the semantics between UNIMARC and MARC 21.

The crucial difference between MARC 21 and UNIMARC has to do with whether the Main Entry exists or not. In the context of MARC 21 the Main Entry is mandatory. On the contrary, it is optional for UNIMARC.

Author Key part
If a Main Entry Author field exists (fields 700, 710, 720), we select this field. In case the main entry is absent, we select a field with the following order (Sfakakis and Kapidakis, 2009):

- the first author name personal field, i.e. tag 701, without subfield $4 or where subfield $4 has value equal to "070" (i.e. relator code for author);
- the first author corporate body or meeting field, i.e. tag 711, without subfield $4 or where subfield $4 has value equal to "070";
- the first author family name field, i.e. tag 721, without subfield $4 or where subfield $4 has value equal to "070".

An improvement of the heuristic under testing is another rule which first explores the statement of responsibility ($f in the MARC21 245 field), and then selects the matching established form of the name from the above mentioned fields.

Title Key part
The OCLC's algorithm (Hickey and O'Neill, 2005) defines the following selection order in title fields:

- Uniform Title (Main Entry) (MARC21 130 => UNIMARC 500, Indicator 2 value 1
- Uniform Title (No Main Entry) (MARC21 240 => UNIMARC 500, Indicator 2 value 0
- Translated Title Supplied by Cataloguer (MARC21 242 => UNIMARC 541)
- Main Title, (UNIMARC 200 => MARC21 245)
- Other Variant Titles, (MARC21 246 => UNIMARC 517)
- Former Title (MARC21 247 => UNIMARC 520)

According to their definitions, the 45X UNIMARC link fields refer to records that are considered as different *Expressions* or *Manifestations* of the same *Work*, such as other editions, translations and reproductions. The previous list does not include linking fields. So, the identification and, consequently, the retrieval of the linked records are an important issue during this process.

Record Label Type Key part

As already mentioned, there is a difference in the semantics between MARC 21 and UNIMARC concerning the Record Label which defines the "Type of Record". According to UNIMARC guidelines for Electronic Resources, there is an option to catalogue digitized material (a map for example) using the Record Label value of Electronic Resource (instead of a Printed Map). Based on this, we used the value "l = electronic resources" in several groups as illustrated below. On the other hand in MARC 21 it is clearly defined that "classes of electronic resources are coded for their most significant aspect (e.g. language material, graphic, cartographic material, sound, music, moving image)". In order to bring together, meaning under the same *Work*, records with different Types of Record we suggest the following grouping. As shown in Figure 4 below, records with different values for the Record Type they might belong either to the same *Work* or to a different one (see examples with Records 4, 6 and 8 in Figure 5.)

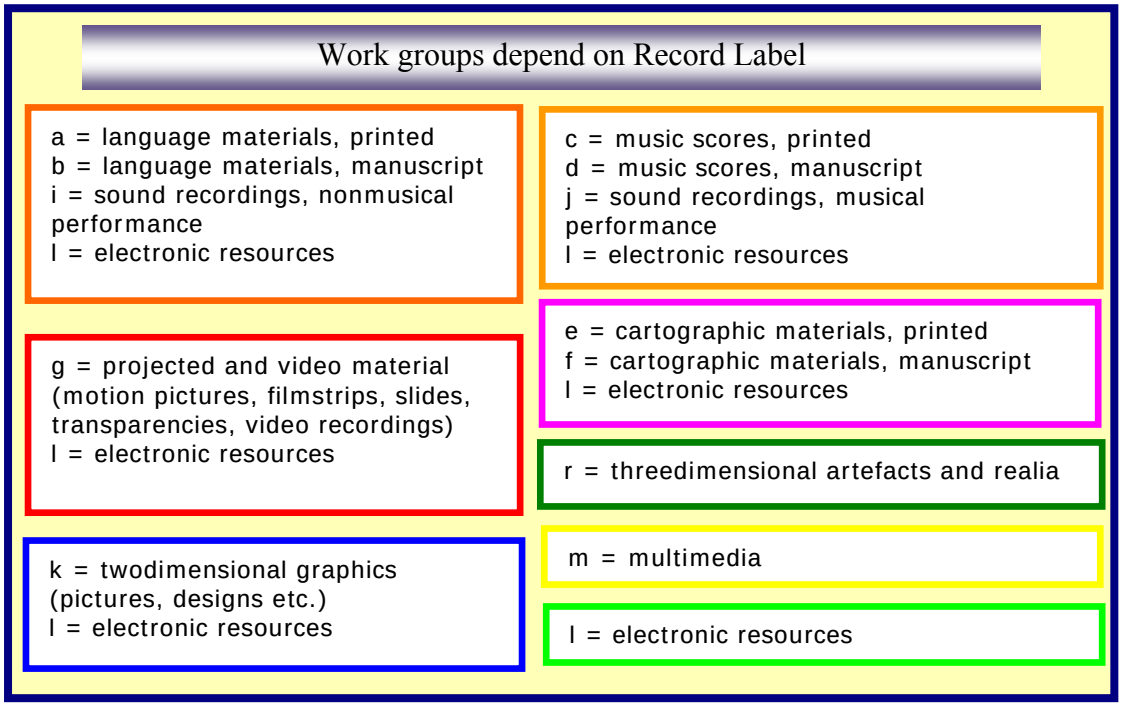


Figure 4: Grouping Suggestion based on the Record Label

An example

According to the FRBR the following three records belong to the same *Work* which consists of two *Expressions* and three *Manifestations*. Based on the above, the key that will gather together all the following records to the same *Work* will be "Author = **HOMER** – Title = **ILIAD** – RecType = **TEXT**".

Record 1 - **Book**	
Title / Author	The Iliad / Homer ; translated by E.V. Rieu
Publication	Harmondsworth : Penguin Books , 1954
Physical description	xxv, 466 p., 20 cm.
Uniform Title	**Iliad**
Author	**Homer**
Translator	Rieu, Emile Victor, 1887-1972
Language of text	English
Record 2 - **Book**	
Title / Author	The Iliad / Translated by E. V. Rieu
Publication	Baltimore : Penguin Books , [1964, c1950]
Physical description	469 p., 18 cm.
Uniform Title	**Iliad**
Author	**Homer**
Translator	Rieu, Emile Victor, 1887-1972
Language of text	English
Record 3 - **Book**	
Title / Author	Ομήρου Ιλιάδα / μετάφραση Ν. Καζαντζάκη, Ι. Θ.Κακριδή
Publication	Αθήνα : Εστία, [1997]
Physical description	401 σ., 22 εκ.
Uniform Title	**Iliad**
Author	**Homer**
Translators	Καζαντζάκης, Νίκος ; Κακριδής, Ιωάννης Θ.
Language of text	Modern Greek

Table 1: Three Records constitute one *Work*, two *Expressions* and three *Manifestations*[3]

Building on the Link Fields

Taking into account the fact that UNIMARC allows for both the existence and the absence of the Control Number of the record towards which the link is being made, we deal with each option separately. Mainly, the existence (or not) of the Control Number is related to which linking technique is implemented. Usually, in the case of embedded fields technique, the Control Number of the record exists while it is absent in the case of the standard technique.

[3] The *Work* is the text of Iliad of Homer, the first *Expression* is the English translation by Rieu (Record 1 and 2) and the second *Expression* is the Modern Greek translation by Kazantzakis and Kakridis (record 3). Every record represents a different *Manifestation*.

UNIMARC records with link fields that embed 001 field

In the case of an existing Control Number of a linked record, all the records that are linked with 45X fields are considered belonging to the same *Work* if the Record Label allows it, regardless of the outcome of the key implementation. In the case of different record label groups, they constitute different but, still, related *Works*. For example, in Figure 5, Record 4 is linked with Record 6 and Record 8; but only Record 4 and 6 belong to the same *Work*. Record 8 does not because it has a different Record Type (for the Record Type groups see figure 4).

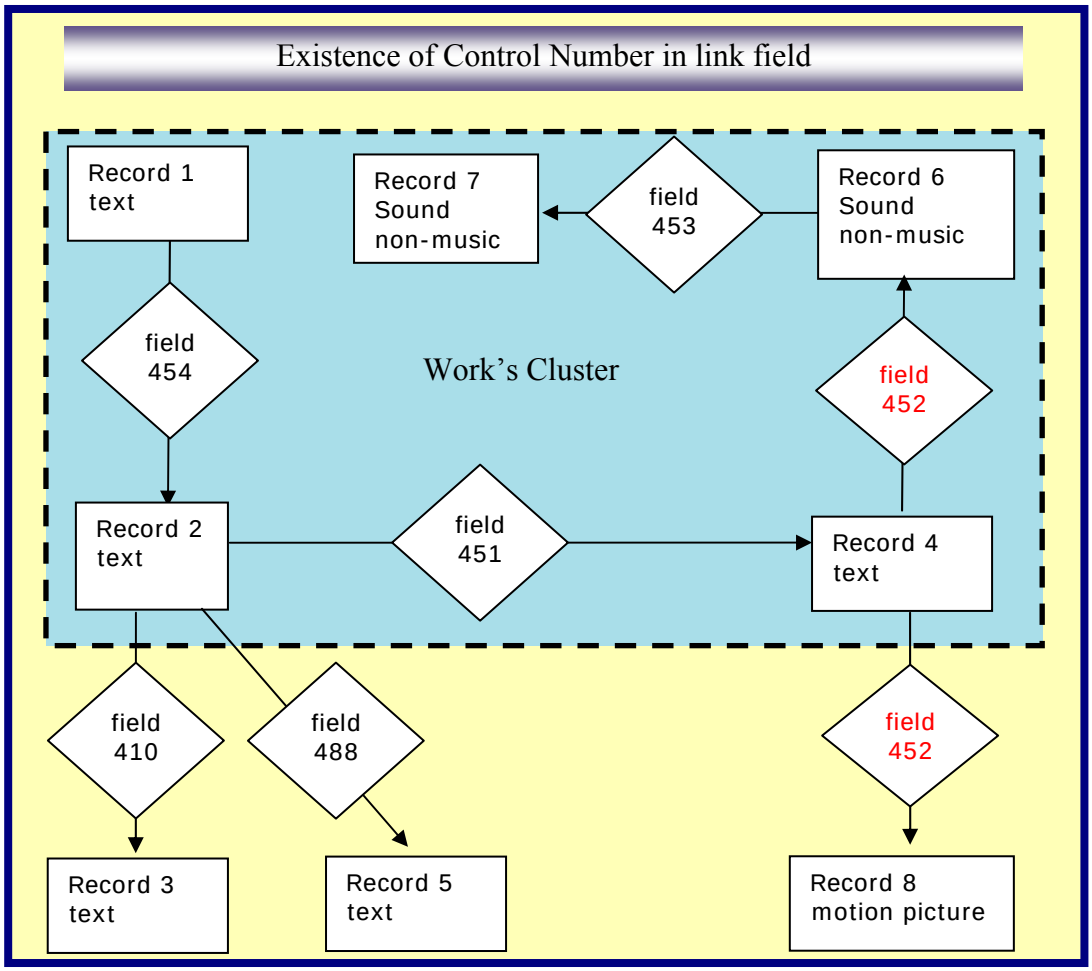


Figure 5: Embedded 001 fields. The light blue background colour (included within the dashed line) illustrates the *Work*

UNIMARC records with link fields that do not embed 001 field

In this case, data from the link fields can be used for key generation. We observed that in the Standard Technique the information of 45X field is more formal than the 200$a. Actually, there is no description of a specific *Manifestation* but rather a more formal (closer to Uniform) title. So, even in the case of the fields 451,452 455 456, it is more effective to use these fields instead of 200$a.

To define the link fields selection order (especially in the case of "453 Translated As" and "454 Translation Of") we considered the 101 Language Field. If the indicator was "1=Item is a translation of the original work or an intermediate work", the field "454 Translation Of" was set right below the Uniform title. If the indicator 1 was "0= Item is in the original language(s) of the work", we did not use field 453.

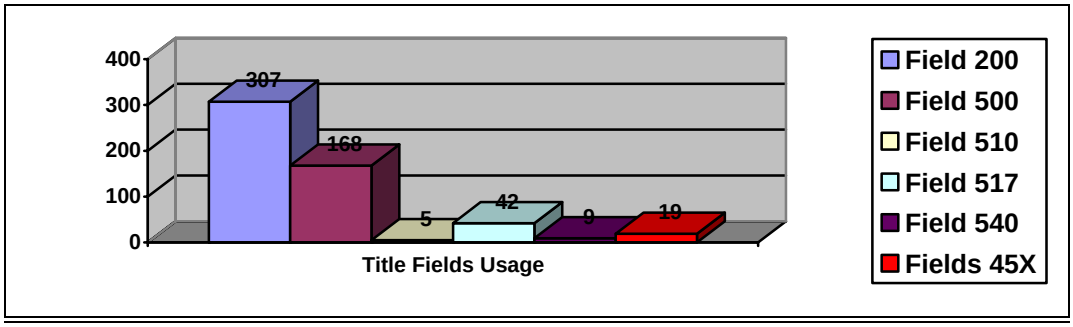
Evaluation of adding link fields to key generation

In order to strengthen the hypothesis that linking fields can be used to increase the effectiveness of the recall, we set up an experiment. For test set we used sample records from the *Union Catalogue of Hellenic Academic Libraries*. This is a large UNIMARC database with more than 3,500,000 records from 54 libraries. Some of the main characteristics of this database are the multilingual data, the absence of common Authority File and the different cataloguing policies implemented from the partners.

We selected *Works* of ancient Greek authors because the *Works* of classical writers have both many *Expressions* and *Manifestations* and constitute an ideal "area" for testing the effectiveness of the FRBRization algorithms. In order to avoid controversial results we manually excluded from our sample all the records which represented fragments of *Works* or *Works* bound together in a single volume.

Due to the fact that the link field policy of the *Union Catalogue of Hellenic Academic Libraries* uses no 001 field in the link fields, we applied only the method where linking fields are used to construct keys. First we used a slight modification[4] of the OCLC algorithm to verify the effectiveness of the FRBRization procedure with our set of data. The main issue was the low recall rate; the algorithm was gathering together only a limited number of records. As far as precision was considered it seemed to work adequately.

The sample consisted of 307 records belonging to 12 *Works*. Therefore, the total success would be a result of 12 generated keys. The most significant title for the *Work* identification is the field of Uniform Title. Unfortunately, as graph number 1 shows, only about half of the records had such a field. Specifically, all 307 records (100%) had a 200 field; 168 records (54,7%) had a 500 field; 5 records (1,6%) had a 510 field; 42 records (13,6%) had a 517 field; 9 records (2,9%) had a 540 field and 19 records (6,1%) had 45X fields. Only 3 records (0,97%) had both Uniform Title and 45X fields.



Graph 1: Title fields distribution

In order to evaluate the effectiveness of our algorithm we applied single link clustering on two *Work key* sets produced from our sample records. The first *Work key* set consisted of

[4] We did not use an Authority File and our metadata was in UNIMARC instead of MARC 21.

keys generated without using link fields (OCLC based), while the second set used link fields as described in the previous section. Applying clustering on the sets, 85 clusters where produced from the first key and 78 clusters where produced from the second one.

The use of link fields improves the effectiveness of the *Work* synthesis about 9%. Even though the comparison of the numbers of the resulting clusters does not provide alone an accurate indication on the effectiveness of the process in general, in our case where the content of the clusters were checked and any cluster contained only similar records, we see that the proportion of the additional cluster between the two approaches is 0.9. Moreover, the improvement was also confirmed from the clustering evaluation measures such as the *corrected RAND* index and the average *silhouette width information*. *RAND* index measures the percentage of decisions that are correct (correct key matches), while the *corrected RAND* increases the sensitivity of the measure. *Silhouette width* evaluates how successfully a key is being clustered, i.e. placed in the correct cluster. More specifically, the values for the *corrected RAND* index and the average *silhouette width information* were equal to 0.56 and 0.81 respectively, while the values for the clustering on the second *Work key* set using the link fields were improved to 0.61 and 0.83 respectively. *RAND* index is closer to our estimation, while the existence of many singleton clusters affects the high improvement of the *silhouette width information*.

Conclusions and further work

First and foremost it has to be made clear that sometimes we do not come across *Works* but *Work* Sets. They resemble to OCLC's *Work Sets* but in our case they bear the significant difference that they are more explicitly distinguished as far as the type of record is concerned. Furthermore, as stated in the FRBR "The concept of what constitutes a *work* and where the line of demarcation lies between one *work* and another may in fact be viewed differently from one culture to another. Consequently the bibliographic conventions established by various cultures or national groups may differ in terms of the criteria they use for determining the boundaries between one *work* and another" FRBR (p. 16).

The results reveal a low recall rate even with the addition of the link fields for the key generation. The main reason of this poor performance is the absence of Uniform Titles (field 500) in combination with the great diversity of existing main titles (field 200). In 307 records there were 141 unique main titles (field 200) while, as graph number 1 shows, there were 550 title fields in total. Using only one field from every record seems to ignore the significance of 243 titles, which constitutes an amount almost equal to the data actually used. Targeting to a meaningful increase of the recall rate we plan to also use this data, i.e. the field titles previously ignored, in order to identify the *Works*. Instead of selecting only one title field, we will compare all fields with each other.

References

Aalberg, T. (2006). A Tool for Converting from MARC to FRBR. In**:** *ECDL 2006, Alicante, Spain, 17-22 September 2006*. Gonzalo, J. et al. (eds.) Berlin, Heidelberg: Springer, pp. 453–456. Available at http://www.springerlink.com/content/5356711834963732/fulltext.pdf. [Last accessed 29/05/2011].

Babeu, A. (2008). Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience. [Internet] Perseus Digital Library. Available at http://www.perseus.tufts.edu/~ababeu/PerseusFRBRExperiment.pdf. [Last accessed 29/05/2011].

Freire, N., Borbinha, J. and Calado, P. (2007). Identification of FRBR Works Within Bibliographic Databases: An Experiment with UNIMARC and Duplicate Detection Techniques. In: *ICADL 2007, Hanoi, Vietnam, 10-13 December 2007*. Berlin, Heidelberg: Springer, pp. 267–276. Available at http://www.springerlink.com/content/d06r28v440n1x420/.

Hickey, T.B. and O'Neill, E.T. (2005). FRBRizing OCLC's WorldCat. Cataloging & Classification Quarterly. 39 (3/4), pp. 239-251.

IFLA (1998). Functional Requirements for Bibliographic Records. Available at http://www.ifla.org/VII/s13/frbr/frbr.pdf. [Last accessed 29/05/2011].

LC FRBR Display Tool (The Library of Congress' Network Development and MARC Standards Office) http://www.loc.gov/marc/marc-functional-analysis/tool.html.

Manguinhas, H., N. Freire, and J. Borbinha. "FRBRization of MARC records in multiple catalogs." In *Proceedings of the ACM International Conference on Digital Libraries*, 225-234, 2010

Naun, C.C. (2010) "Next generation OPACs: A cataloging viewpoint." *Cataloging and Classification Quarterly* 48 (4), pp. 330-342.

Peponakis, M.; Sfakakis, M.; Kapidakis, S. (2010) "FRBRization: Seeking for the "key" to Works' Identification" (written in Greek). In *Proceedings of the 19th Hellenic Conference of Academic Libraries*. Available at http://library.panteion.gr/19libconf/conference_en.php [Last accessed 29/05/2011]

Rajapatirana, B. and Missingham, R. (2005). The Australian National Bibliographic Database and the Functional Requirements for the Bibliographic Database (FRBR). *The Australian Library Journal*. 54 (1), pp. 31-42. Available at http://www.alia.org.au/publishing/alj/54.1/full.text/rajapatirana.missingham.html. [Last accessed 29/05/2011]

Sfakakis, M. and Kapidakis, S. (2009). Eliminating query failures in a work-centric library meta-search environment. *Library Hi Tech*. 27 (2), pp. 286-307

Tillett, B. (2004). What is FRBR? A conceptual model for the bibliographic universe. [Internet]. Available at http://alia.org.au/publishing/alj/54.1/full.text/tillett.html . [Last accessed 29/05/2011].

Yee, M.M. (2005). FRBRization: a Method for Turning Online Public Finding Lists into Online Public Catalogs. *Information Technology and Libraries*. 24 (3), pp. 77-95. Post print available at http://repositories.cdlib.org/postprints/715/. [Last accessed 29/05/2011]