Application of FRBR to existing MARC data: A comparison of outcomes

Erik Mitchell
Wake Forest University
Winston Salem, NC 27106
mitcheet@wfu.edu

Carolyn McCallum
Wake Forest University
Winston Salem, NC 27106
mccallcj@wfu.edu

Patricia Strickland
Wake Forest University
Winston Salem, NC 27106
strickpk@wfu.edu

ABSTRACT
This poster examines the outcomes of the application of FRBR algorithmic approaches for conversion of existing MARC datasets. The source data in this research includes the records from a medium sized academic library and consist of 1.6 Million records. The conversion is compared with other FRBR approaches. The authors utilized the OCLC FRBR work-set algorithm to create programs in the Python platform to generate match fields. These match points were loaded into a MySQL database and compared to generate match points.

This poster presents the outcomes of the first comparisons of this data based primarily on the OCLC algorithm. It documents the process of creating the programs and discusses next steps in both formalizing the code for deployment and refinement to enable more complete matching of FRBR keys.

Keywords
Functional Requirements for Bibliographic Records, MARC, Python, PyMARC

INTRODUCTION
FRBR (Functional Requirements for Bibliographic Records) is of growing importance for library metadata. This is due in part to the emergence of RDA (Resource Description and Access) but also in part to a response to a growing ability of library systems to represent FRBR relationships in discovery systems. Although the FRBR data model is well defined and work-set algorithms have been defined to migrate traditional MARC data into FRBR modeled data, the application of these algorithms have typically been done in a ‘perfect’ world environment. This study examines the success of and issues raised when FRBR work-set algorithms are applied to a catalog of library data from a medium sized academic library.

LITERATURE
Although FRBR as a standard emerged in the late 1990s (IFLA, 2008), it has not seen widespread adoption in the library field despite the development of conversion guidelines (Hickey, 2009) and the trial application on both limited and expansive sets of metadata (Hickey, 2009; Zhang, 2011). FRBR deconstructs the bibliographic representation of MARC data into four related record structures. The primary entity (the Work) contains information about the intellectual property of the work (e.g. the title, author, and content). The expression dataset contains a specific “intellectual or artistic form” of the work (IFLA, 2008). While further refinements (manifestation and item) also exist, this research examined only the generation of Work level FRBR matches.

Research in the field has pointed to some key challenges behind the creation of these entities. First, only recently have systems been created that support faceted search and FRBRized display of data (Chew Chiat, 2010). Second, cataloging and data encoding standards have only recently been introduced that make widespread adoption of a FRBRized dataset possible. A key limiting factor beyond these two issues however is the existence of a heterogeneous and inconsistently maintained set of MARC metadata in library catalogs.

Library catalogs tend to have metadata that has not been updated in many years and is often characterized by varied local standards and approaches to metadata creation and maintenance. Because FRBR work-set rules rely in particular on normalized Author data, this work is complicated where authority data has not been regularly managed over time.

In order to better understand the real-world challenge of FRBRization of library metadata, this research project seeks to take sets of library metadata through published FRBR algorithmic processes to determine what challenges and opportunities exist in regards to the conversion of this data.
METHODOLOGY
The initial research on this project was divided into three sections. First, the FRBR Work-Set Algorithm from OCLC was interpreted and coded into a Python application. This coding process involved interpretation of rules and coding using the Python and PyMARC platform. The resulting application is capable of parsing MARC records using the PyMARC interpreter with the purpose of extracting and normalizing FRBRized keys. The created keys were colocated with unique bibliographic identifiers allowing the researchers to trace matched and unmatched records back to source titles in the integrated library system (ILS).

The second section of the research included the creation of a relational database to store these keys for later comparison. The researchers utilized a MySQL database to gather and analyze keys using SQL (Structured Query Language) queries. The final section of this initial research included a comparison of match results with previously completed comparisons. This poster presents the results of the initial phase of this research. This includes the encoding of the OCLC work-set algorithm into a program using PyMARC and the Python platform to generate matching keys, the processing of data to prepare it for comparison and the results from the analysis of these keys to create Work records.

FINDINGS
The OCLC FRBR work-set keys are based on four match points. These keys were generated through the application of a series of rules that included subfield extraction and entry normalization according to NACO rules. The initial comparison between the percentage of fields matched was found to be similar to both the OCLC (Hickey, 2009) and Kent State (Zhang, 2011) comparisons.

Keys were compared to discover how many “Works” could be created from each key. An initial comparison of selected titles indicates that title field inconsistencies and lack of authority control are potential barriers to good matches.

Initial attempts to create manifestation comparisons resulted in overall low match rates. For example, key 1 match resulted in a match rate of only 8.4%. The comparison found sizable drop-off in other match rates (e.g. key 2 only resulted in .8% matches), this indicates that a large number of titles in this dataset would not find corresponding matches. In phase 2 of this research, more detailed analysis of this gap will be explored to better understand what expected match points exist.

DISCUSSION
The initial comparison of data shows sizeable issues that are likely to be encountered by libraries seeking to FRBRize data. Lack of universal application of OCLC number, inconsistent maintenance of open fields and inconsistent sources of records are likely causes. Both Kent State (Zhang, 2011) and National Library of Australia (Ayres, 2004) took on real-world implementations of FRBR on existing data-sets and it would be interesting to analyze their approaches as alternative normalization rules and matching methods are considered.

CONCLUSION
The initial work on this project indicates that there is considerable work to be done in understanding the issues that exist with real-world library metadata. The interest in FRBRization of MARC data and the increasing need to find a solution with the coming adoption of RDA means that interest in this area of research is likely to grow in the coming years. Further work is required to explore what these issues are and how they could be dealt with in a production environment.

In phase two of this research, more detailed examination of the issues affecting match rates and potential alternative approaches for FRBR matches will be considered. This will be accomplished both by statistical comparison of matches and examination of specific titles.

REFERENCES


