Preparing, Creating, and Managing a Large Dataset of MARC 21 Records for Research and Analysis

Serhiy Polyakov (sp0065@unt.edu)
Interdisciplinary Ph.D. Program in Information Science, University of North Texas, Denton, TX
Advisors: Dr. William E. Moen, Dr. Shawne Miksa, Texas Center for Digital Knowledge

Introduction

The MARC Content Designation Utilization (MCDU) Project (<http://www.mcdm.unt.edu>) is examining catalogers’ utilization of available MARC 21 content designation (e.g., fields/subfields). A critical component of this project is the development of methods, procedures, and software tools for reliable and valid analyses of the complete WorldCat database of more than 56 million MARC 21 records.

Analyses of MARC 21 content designation conducted on MARC records stored in the original format would require writing a computer program for every single question of the analysis, and that would not be efficient. Decomposing and storing MARC 21 records in a relational database allows answering various questions about content designation utilization using queries written in SQL.

The following describes the research related to development of the decomposition specification, designing parsing software, and creating a database structure to store these MARC 21 records for subsequent analysis.

Methodology

Building a very large database for storing and analyzing decomposed MARC records presented major challenges in areas of efficiency and performance of the application. The approach used was to systematically estimate all processing requirements using sample sets of the records. For example, we were able to estimate total storage requirements; calculations showed that after converting the set of 56 million records stored in raw MARC format into the relational database format suitable for the analysis storage requirement would be about 125GB (prior to indexing) and total number of rows in all tables would be about 3.8 billion.

The following is a sample MARC record in a raw format. For presenting this sample here nonprinting characters for Record Terminator, Field Terminator, Delimiter, Blank, and space have been substituted with \', \', \', \', and \', respectively.

| 00800 | 00900 | 01000 | 01900 | 02000 | 02100 | 02200 | 02300 | 02400 | 02500 | 02600 | 02700 | 02800 | 02900 | 03000 | 03100 | 03200 | 03300 | 03400 | 03500 | 03600 | 03700 | 03800 | 03900 | 04000 | 04100 | 04200 | 04300 | 04400 | 04500 | 04600 | 04700 | 04800 | 04900 | 05000 | 05100 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| \', 0800 | \', 00900 | \', 01000 | \', 01900 | \', 02000 | \', 02100 | \', 02200 | \', 02300 | \', 02400 | \', 02500 | \', 02600 | \', 02700 | \', 02800 | \', 02900 | \', 03000 | \', 03100 | \', 03200 | \', 03300 | \', 03400 | \', 03500 | \', 03600 | \', 03700 | \', 03800 | \', 03900 | \', 04000 | \', 04100 | \', 04200 | \', 04300 | \', 04400 | \', 04500 | \', 04600 | \', 04700 | \', 04800 | \', 04900 | \', 05000 | \', 05100 |

The decomposed records are stored in seven tables that represent the primary components of a MARC 21 record: Leader, Directory, Control_001_005, Control_006, Control_007, Control_008, and Data_010_999. Figure 3 gives an example of one of these tables.

Results

The decomposed records are stored in seven tables that represent the primary components of a MARC 21 record: Leader, Directory, Control_001_005, Control_006, Control_007, Control_008, and Data_010_999. Figure 3 gives an example of one of these tables.

Conclusion

Creating and managing a very large dataset for research purposes indicates the need for planning, testing, quality assurance, and a clear understanding of the types of analyses and questions the research is addressing.

Figure 1: Parser and Database Loader Software Structure

Figure 2: Parser and Database Loader Program User Interface and resulting statistics after parsing and loading of 56 million records

Figure 3: Table storing decomposed data fields (Data_010_999)