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 <www.mcdu.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.

```
00700cem##2200253###45•0001001300000003000600013005001700019006001900000
36007000900055008004100064010001700105040002500122043001200147050000250
0159082001300184110003300197245002800230260002000258300004100278440005
300319500001700372651003800389994001900427^ocm00008028#^OCoLC^20041106
021327.0^ab##########000#0#^aj#canzn^690501s1966####txu#######a#####0#
##eng##^##$a•••74208040•^##$aDLC$cDLC$dOCL$dOCLCQ^##$an-us-tx^0#$aHC10
7.T4$bA325•no.•3^00$a330.9764^2#$aXxxxx•Xxxxxxxxxx•Xxxxxxxxxx.^10$aXxx
xxxxx•xxxx•xx•Xxxxx.^##$aAustin$c[1966?]^##$a[1]•l.,$b13•fold.•col.•ma
ps.$c27•cm.^#0$aIndustrial•economic•opportunities•series,$vno.•3^##$aC
over•title.^#0$aXxxxx$xEconomic•conditions$vMaps.^##$a11$bOCL$i00000^\
```

The parsing software developed reads MARC records from the ASCII text file, parses the records in memory and inserts the records in to MySQL database. Figure 1 illustrates the functions of the parser and the database loading programs. PHP scripting language has been used to implement the parser and 64-bit version of MySQL 5 database server is used to store and analyze the parsed records.

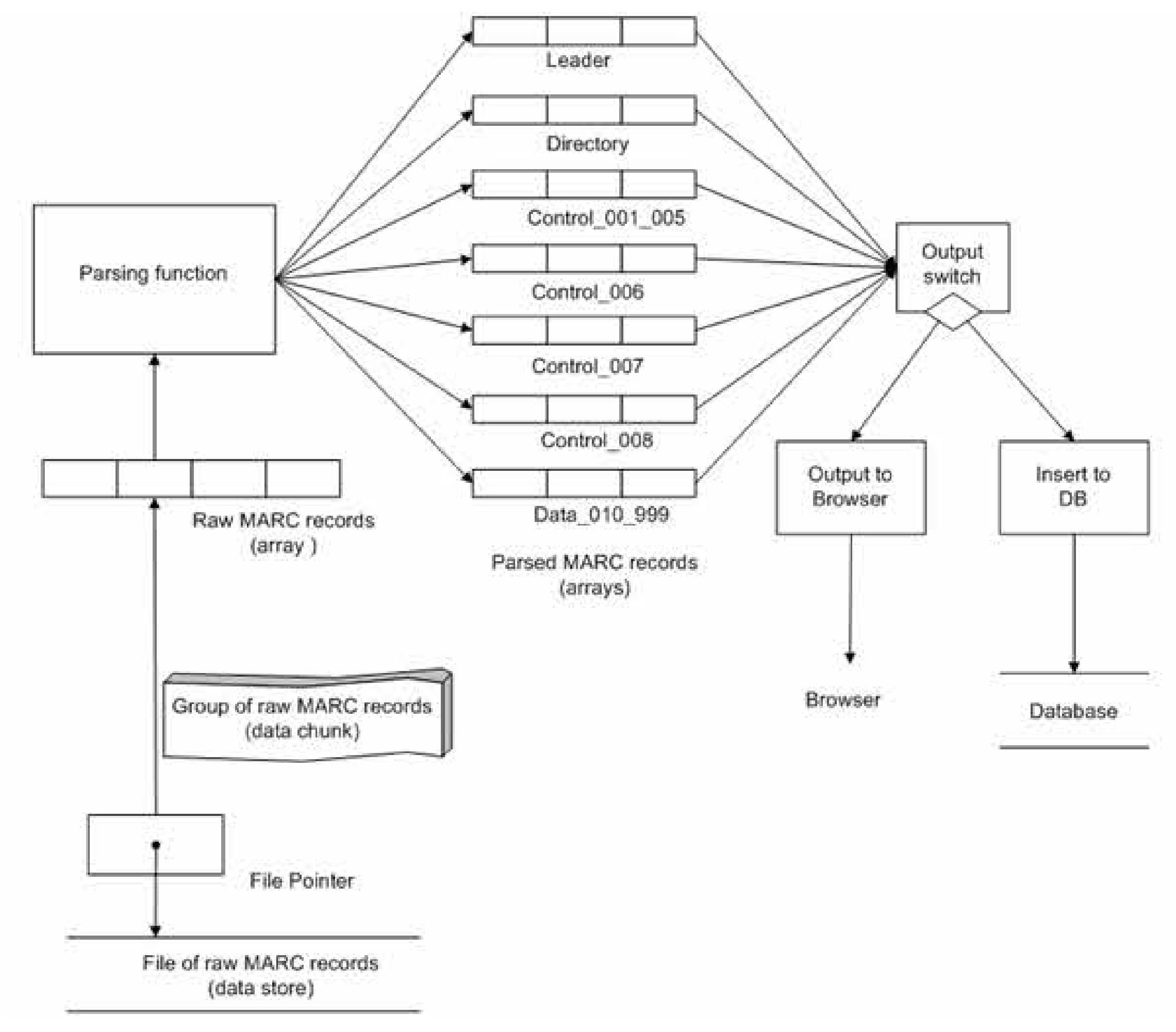
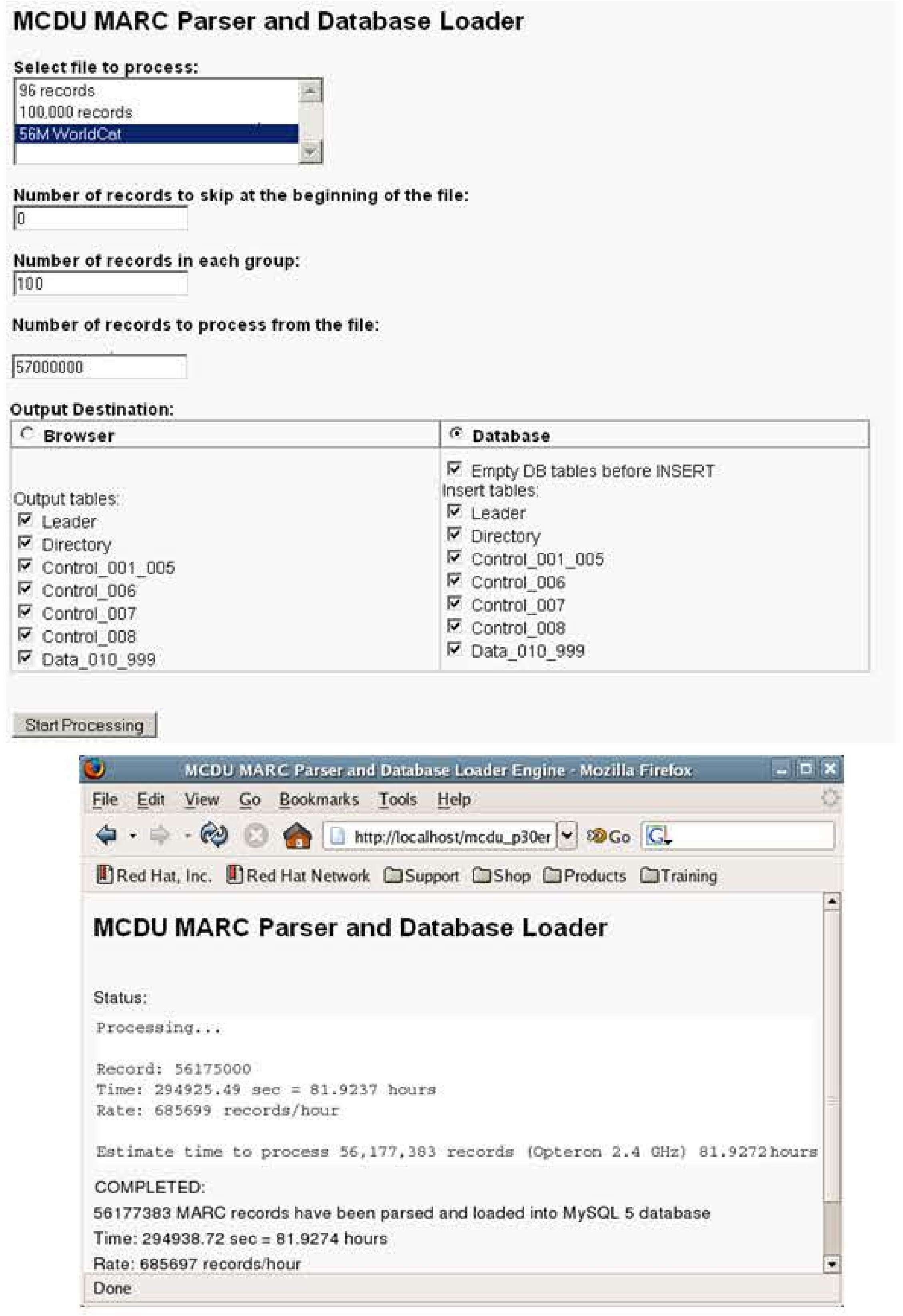


Figure 1: Parser and Database Loader Software Structure

Figure 2 presents the parsing software user interface. A user selects various options and output destination, which can be either Browser or Database. Browser output is used to perform validation confirming correct operation of the parser. If Database is chosen for the output, the system launches the database loader and loads the parsed data into the MySQL database.



MCDU MARC Parser and Database Loader

Select file to process:
96 records
100,000 records
56M WorldCat

Number of records to skip at the beginning of the file: 0

Number of records in each group: 100

Number of records to process from the file: 57000000

Output Destination:

○ Browser	● Database
Output tables: ☑ Leader ☑ Directory ☑ Control_001_005 ☑ Control_006 ☑ Control_007 ☑ Control_008 ☑ Data_010_999	☑ Empty DB tables before INSERT Insert tables: ☑ Leader ☑ Directory ☑ Control_001_005 ☑ Control_006 ☑ Control_007 ☑ Control_008 ☑ Data_010_999

Start Processing
MCDU MARC Parser and Database Loader Engine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost/mcdu_p30er
Go
Red Hat, Inc. Red Hat Network Support Shop Products Training

MCDU MARC Parser and Database Loader

Status:

```
Processing...

Record: 56175000
Time: 294925.49 sec = 81.9237 hours
Rate: 685699 records/hour

Estimate time to process 56,177,383 records (Opteron 2.4 GHz) 81.9272 hours
```

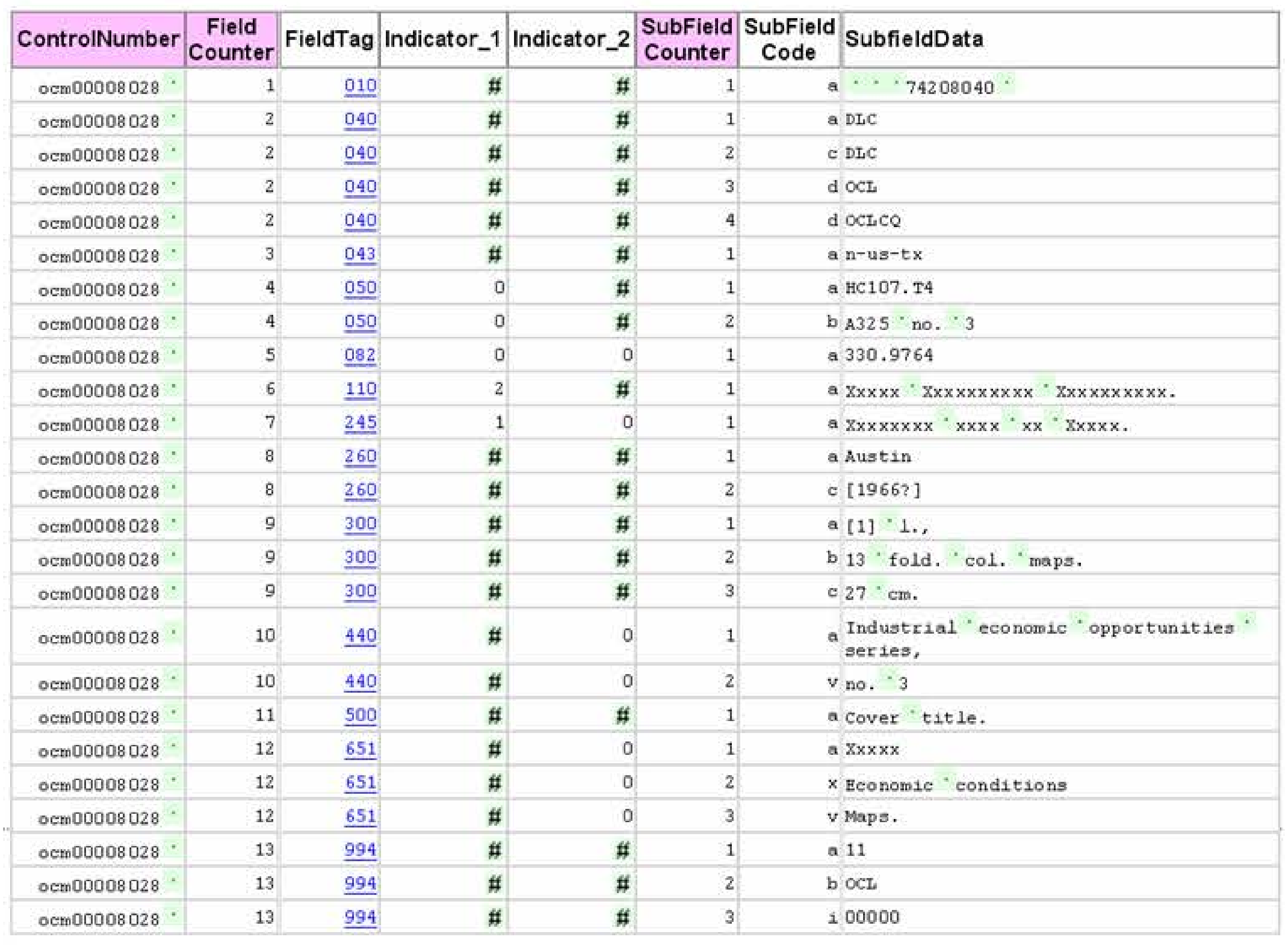
COMPLETED:
56177383 MARC records have been parsed and loaded into MySQL 5 database
Time: 294938.72 sec = 81.9274 hours
Rate: 685697 records/hour

Done

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

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.



ControlNumber	Field Counter	FieldTag	Indicator_1	Indicator_2	SubField Counter	SubField Code	SubfieldData
ocm00008028 '	1	010	#	#	1	a	' ' '74208040 '
ocm00008028 '	2	040	#	#	1	a	DLC
ocm00008028 '	2	040	#	#	2	c	DLC
ocm00008028 '	2	040	#	#	3	d	OCL
ocm00008028 '	2	040	#	#	4	d	OCLCQ
ocm00008028 '	3	043	#	#	1	a	n-us-tx
ocm00008028 '	4	050	0	#	1	a	HC107.T4
ocm00008028 '	4	050	0	#	2	b	A325 'no. '3
ocm00008028 '	5	082	0	0	1	a	330.9764
ocm00008028 '	6	110	2	#	1	a	Xxxxx 'Xxxxxxxxxx 'Xxxxxxxxxx.
ocm00008028 '	7	245	1	0	1	a	Xxxxxxxx 'xxxx 'xx 'Xxxxx.
ocm00008028 '	8	260	#	#	1	a	Austin
ocm00008028 '	8	260	#	#	2	c	[1966?]
ocm00008028 '	9	300	#	#	1	a	[1] 'l.,
ocm00008028 '	9	300	#	#	2	b	13 'fold. 'col. 'maps.
ocm00008028 '	9	300	#	#	3	c	27 'cm.
ocm00008028 '	10	440	#	0	1	a	Industrial 'economic 'opportunities 'series,
ocm00008028 '	10	440	#	0	2	v	no. '3
ocm00008028 '	11	500	#	#	1	a	Cover 'title.
ocm00008028 '	12	651	#	0	1	a	Xxxxx
ocm00008028 '	12	651	#	0	2	x	Economic 'conditions
ocm00008028 '	12	651	#	0	3	v	Maps.
ocm00008028 '	13	994	#	#	1	a	11
ocm00008028 '	13	994	#	#	2	b	OCL
ocm00008028 '	13	994	#	#	3	i	00000

Figure 3: Table storing decomposed data fields (Data_010_999)

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.

2006 ALISE Annual Conference, From Research to Practice: The Scholarship of Teaching and Learning in LIS Education, January 16-19, San Antonio, Texas