



U.S. Federal Institute of Museum and Library Services
National Leadership Grant

**Realizing the Vision of Networked Access to
Library Resources**

*An Applied Research and Demonstration Project to
Establish and Operate a Z39.50 Interoperability Testbed*

**Decomposing MARC 21 Records
for Analysis**

William E. Moen, Ph.D.
<wemoen@unt.edu>
Principal Investigator

School of Library and Information Sciences
Texas Center for Digital Knowledge
University of North Texas
Denton, TX 76203

October 1, 2001
Revised January 1, 2002

Table of Contents

1. Introduction
2. Overview
3. Specific Requirements and Parameters
 - 3.1 Defining a “Word”
 - 3.2 Information Recorded in a Decomposed Record
 - 3.3 Structure of the Decomposed Records for Import into Database
 - 3.4. Reconstructing Views of the MARC 21 Record from Decomposed Records
4. Summary

Appendix A: Sample of Decomposed Records

Decomposing MARC 21 Records for Analysis

1. Introduction

To prepare the test dataset of the 1% sample of MARC 21 records from the WorldCat database for use in the Z39.50 Interoperability Testbed, we need to be able to efficiently analyze the records to determine relevant records to be returned for a set of test searches. The first step in that analysis is to determine the occurrence of test search terms in specific records. This document describes the general approach for this analysis and identifies specifications for the analysis.

2. Overview

The approximately 400,000 records from OCLC's WorldCat database need to be analyzed at a "word" level to identify which records contain specific words that will be used in interoperability test searches. We not only need to identify which records contain a specific word but also identify where the word appears in the record (e.g., is a word in a title-related field or a subject-related field).

Discussions with Ed O'Neill, OCLC's Office of Research, suggested that the test dataset could be decomposed and provide a more discrete view of the records' data. He suggested that a single pass through the test dataset could produce a decomposed view of the individual records. Each MARC 21 record would be decomposed into separate subrecords that consist of data including the:

- OCLC Number for the record
- each word appearing in the record
- the location (i.e., field and subfield) in which each word appears.

As an example of this approach, a book with the author *Peter Herson* and the title *Evaluation and library decision...* would be decomposed as shown in the table.

OCLC#	Field	Subfield Code	Word
21910462	100	a	Herson
21910462	100	a	Peter
21910462	245	a	Evaluation
21910462	245	a	and
21910462	245	a	library
21910462	245	a	decision

Once this parsing has been completed, the resulting subrecords will be imported into a database application for subsequent analysis.

OCLC agreed to run the 400,000 records through a parsing program that would produce a decomposed record view. This was done, resulting in a file of approximately 33,000,000 subrecords for the 400,000 MARC 21 records. The following details provide more specific information about the requirements for this decomposition.

3. Specific Requirements and Parameters

This section identifies the details of the parsing and addresses specific concerns such as the data captured in the subrecords. The first decision was what would count as a “word” for the decomposition.

3.1 Defining a “Word”

Z-Interop testing required the identification of specific words in the 400,000 records. We defined a word as a “string of characters bounded by spaces.” Such a definition has some problems when it comes to MARC records since there are fields (e.g., 008) that have strings of characters but also blanks as part of the data in the field. Therefore, we treated the content of MARC fields 001-009 as a single word regardless of spacing since these fields are coded and not subfielded. The other complication is the occurrences of punctuation in the fields. Therefore, we qualified the definition of word as follows: “any string of characters bounded by spaces including all punctuation and other special characters.”

Because a string of characters might include leading, internal, and ending punctuation, data normalization to remove specific punctuation (e.g., leading and ending punctuation) would be necessary. This was carried out in separate procedures carried out on the decomposed records. These data normalization procedures are documented in a separate Z-Interop document, **Data Normalization Procedures on Decomposed MARC 21 Records**.

3.2 Information Recorded in a Decomposed Record

Discussions with Ed O’Neill concluded that each subrecord of the decomposed MARC 21 record would contain information necessary for analysis of words to determine not only in which MARC 21 record they appeared in but also their position in the record. The following lists the information included for each subject record (when applicable) and shown in the example table below:

- **OCLC Number (OCLC#)** – Necessary for identifying in which record a word appears
- **Field Tag (Field)**
- **First Indicator Value (1st Indicator)**
- **Second Indicator Value (2nd Indicator)**
- **Subfield Value (Subfield)**
- **Field Position in Record (Field Position)** – Necessary in cases where a field is repeated
- **Subfield Position in Record (Subfield Position)** – Necessary in cases where a subfield is repeated
- **Word Position in Field/Subfield (Word Position)** – Necessary to be able to identify “phrases” where several adjacent words are a search term
- **Specific Character String (Word)**

OCLC#	Tag	1st Indicator	2nd Indicator	Subfield	Field Position	Subfield Position	Word Position	Word
3	110	2		A	11	1	1	national
3	110	2		A	11	1	2	study
3	110	2		A	11	1	3	service
3	245	1	0	A	12	1	1	illegitimacy
3	245	1	0	A	12	1	2	and
3	245	1	0	A	12	1	3	adoption
3	245	1	0	B	12	2	1	report

For a more complete example of the decomposed records, see Appendix A.

3.3 Structure of the Decomposed Records for Import into Database

We had a requirement that the decomposed records could be easily imported into a database for subsequent processing. The OCLC parsing process on the 400,000 MARC 21 records resulted in a tab delimited file of approximately 33,000,000 subrecords. These were then imported into a MySQL database for processing.

3.4 Reconstructing Views of the MARC 21 Record Data from Decomposed Records

A requirement of the parsing procedure was to produce subrecords that could be further analyzed. Those analysis procedures required the ability to generate a variety of reports including frequency counts of word occurrences, lists of record numbers in which a word appears, etc. (see Z-Interop document **SQL Data Analysis Procedures to Create Aggregate and Candidate Record Groups on a Sample of Decomposed MARC Records, Phase 1 Testing**). The structure of the decomposed records provided sufficient information for such reports. For example, we can generate reports for the occurrence of words in particular records, in specific fields, and specific subfields.

4. Summary

This document has described the requirements and specifications for preparing the MARC 21 records for subsequent processing and analysis. The OCLC parsing procedures successfully produced tab delimited subrecords containing the appropriate information for analysis to prepare test searches for the testbed.

Appendix A

Sample of Decomposed Records

To indicate the decomposition of MARC 21 records, the following is an example of a complete MARC 21 record and the output of the decomposition.

Example of Complete MARC 21 Record

```
LDR01019cam 2200265 4500^
001ocm00000003^
003OCoLC^
00520010925133908.0^
008690414s1963 nyu b 000 0 eng ^
010 _a 63064323^
019 _a7124033 _a10654585 _a14218190^
040 _aDLC _cDLC^
049 _aOCLC^
0500 _aHV700.5 _b.N37^
0820 _a362.7/3^
1102 _aNational Study Service.^
24510 _alllegitimacy and adoption in Maine : _breport of a study made for the Maine Committee on Children and Youth.^
260 _a[New York], _c1963.^
300 _a24 p. ; _c28 cm.^
500 _aCover title.^
504 _aBibliographical footnotes.^
650 0 _alllegitimacy _zMaine.^
650 0 _aAdoption _zMaine.^
7101 _aMaine. _bCommittee on Children and Youth.^
994 _a00 _bOCLC^
```

Example of MARC 21 Record Decomposed

OCLC#	Tag	1st Ind	2nd Ind	Subfield	Field Position	Subfield Position	Word Position	Word
3	1				1	1	1	ocm00000003
3	3				2	1	1	OCoLC
3	5				3	1	1	20010215000003.0
3	8				4	1	1	690414s1963 n yu b 000 0 eng
3	10			a	5	1	1	63064323
3	40			a	6	1	1	DLC
3	40			c	6	2	1	DLC
3	19			a	7	1	1	7124033
3	19			a	7	2	1	10654585
3	19			a	7	3	1	14218190
3	50	0		a	8	1	1	HV700.5
3	50	0		b	8	2	1	.N37
3	82			a	9	1	1	362.7/3
3	49			a	10	1	1	OCLC
3	110	2		a	11	1	1	National
3	110	2		a	11	1	2	Study
3	110	2		a	11	1	3	Service.
3	245	1	0	a	12	1	1	Illegitimacy
3	245	1	0	a	12	1	2	and
3	245	1	0	a	12	1	3	adoption
3	245	1	0	a	12	1	4	in
3	245	1	0	a	12	1	5	Maine
3	245	1	0	b	12	2	1	report
3	245	1	0	b	12	2	2	of
3	245	1	0	b	12	2	3	a
3	245	1	0	b	12	2	4	study
3	245	1	0	b	12	2	5	made
3	245	1	0	b	12	2	6	for
3	245	1	0	b	12	2	7	the
3	245	1	0	b	12	2	8	Maine
3	245	1	0	b	12	2	9	Committee
3	245	1	0	b	12	2	10	on
3	245	1	0	b	12	2	11	Children
3	245	1	0	b	12	2	12	and
3	245	1	0	b	12	2	13	Youth.

OCLC#	Tag	1st Ind	2nd Ind	Subfield	Field Position	Subfield Position	Word Position	Word
3	260			a	13	1	1	[New
3	260			a	13	1	2	York]
3	260			c	13	2	1	1963.
3	300			a	14	1	1	24
3	300			a	14	1	2	p.
3	300			c	14	2	1	28
3	300			c	14	2	2	cm.
3	500			a	15	1	1	Cover
3	500			a	15	1	2	title.
3	504			a	16	1	1	Bibliographical
3	504			a	16	1	2	footnotes.
3	650		0	a	17	1	1	Illegitimacy
3	650		0	z	17	2	1	Maine.
3	650		0	a	18	1	1	Adoption
3	650		0	z	18	2	1	Maine.
3	710	1		a	19	1	1	Maine.
3	710	1		b	19	2	1	Committee
3	710	1		b	19	2	2	on
3	710	1		b	19	2	3	Children
3	710	1		b	19	2	4	and
3	710	1		b	19	2	5	Youth.