
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*

**Analysis Logic and Procedures for
Creating a Test Dataset of
MARC 21 Records for the
Z39.50 Interoperability Testbed**

Phase 1 Testing

William E. Moen, Ph.D.

<wemoen@unt.edu>

Principal Investigator

&

Haley K. Holmes

<holmehk0@hotmail.com>

Z-Interop Research Assistant

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 of Procedures for Creating Descriptive Profile of Test Dataset**
 - 2.1 Goal of Procedures for Developing the Descriptive Profile
 - 2.2 Detailed Analysis Procedures
 - 2.3 Reports and Outputs
- 3. Overview of Procedures for Determining Content of Dataset**
 - 3.1 Goal of Procedures
 - 3.2 Procedure: Word Frequency Count
 - 3.2.1 Goal of Procedure
 - 3.2.2 Detailed Analysis Procedures
 - 3.2.3 Reports and Outputs
 - 3.3 Procedure: Identifying the Aggregate Record Groups
 - 3.3.1 Goal of Procedure
 - 3.3.2 Detailed Analysis Procedures
 - 3.3.3 Reports and Outputs
 - 3.4 Procedure: Identifying the Level 0 Profile Searches Candidate Record Groups
 - 3.4.1 Goal of Procedure
 - 3.4.2 Detailed Analysis Procedures
 - 3.4.2.1 Title Keyword Candidate Record Group
 - 3.4.2.2 Author Keyword Candidate Record Group
 - 3.4.2.3 Subject Keyword Candidate Record Group
 - 3.4.2.4 Any Keyword Candidate Record Group
 - 3.4.3 Reports and Outputs
- 4. Summary of Functionality Needed for Tools to Conduct Analyses**
 - 4.1 Descriptive Profile of Test Dataset
 - 4.1.1.a Analysis of Type of Material
 - 4.1.1.b Reporting
 - 4.1.2.a Analysis of Encoding Level
 - 4.1.2.b Reporting
 - 4.1.3.a Analysis of Descriptive Cataloging Form
 - 4.1.3.b Reporting
 - 4.1.4.a Analysis of Cataloging Source
 - 4.1.4.b Reporting
 - 4.2 Word Frequency Count
 - 4.2.1.a Analysis
 - 4.2.1.b Reporting
 - 4.3 Identifying the Aggregate Record Groups
 - 4.3.1.a Analysis
 - 4.3.1.b Reporting
 - 4.4 Identifying the Candidate Keyword Record Groups
 - 4.4.1.a Analysis of Title Keyword Group
 - 4.4.1.b Reporting
 - 4.4.2.a Analysis of Author Keyword Group
 - 4.4.2.b Reporting
 - 4.4.3.a Analysis of Subject Keyword Group
 - 4.4.3.b Reporting
 - 4.4.4.a Analysis of Any Keyword Group
 - 4.4.4.b Reporting

Analysis Logic and Procedures for Creating a Test Dataset of MARC 21 Records for the Z39.50 Interoperability Testbed

Phase 1 Testing

1. Introduction

This document describes the logic and procedures to create a test dataset of more than 400,000 (400K) MARC 21 records from the OCLC WorldCat database. This test dataset (hereafter referred to as the dataset) provides a controlled set of data for use in the Z39.50 Interoperability Testbed Project (hereafter referred to as Z-Interop). OCLC selected a 1% weighted sample from its WorldCat database, which contains approximately 45 million records. Weighting was based on number of holdings listed per record.

The Z-Interop researchers analyzed the entire sample set of records for the presence of certain characteristics to create a descriptive profile of the dataset. Other analysis procedures determined the content of the dataset's records, identified Aggregate Record and Candidate Record Groups for the occurrences of selected words and terms. For each procedure, we list the goal and expected outputs and reports. In some cases, the analyses were performed on the complete MARC 21 records in the dataset and in other cases the analyses were performed on the decomposed MARC 21 records (see the Z-Interop document, *Decomposing MARC 21 Records for Analysis*, for additional information).

A descriptive profile of the dataset provides an overall view of the records and others can assess how representative this dataset is against their own online catalog databases. A word frequency count analysis assists in the systematic selection of terms for use in subsequent analysis of the dataset. Aggregate Record Groups are a set of dataset records that contain a specific term anywhere in the record. A Candidate Record Group contains the specific term in selected fields and subfields related to the types of test searches used in interoperability testing (e.g., Author, Title, and Subject). During the interoperability testing, the terms based on analysis described in this document will be used as search terms in the test searches. The explanation and details about the test searches for interoperability testing is outside the scope of this document; details can be found in other Z-Interop documents.

This document is focused on the analysis procedures used to prepare for Phase 1 Testing in Spring 2002. A subsequent version of this document will address the additional procedures for Phase 2 Testing scheduled for Summer 2002.

2. Overview of Procedures for Creating Descriptive Profile of Test Dataset

A descriptive profile characterizes the dataset. This profile provides empirical evidence regarding the following record characteristics:

- Type of Record
- Encoding Level
- Descriptive Cataloging Form
- Source of Cataloging.

The MARC 21 Format for Bibliographic Data says the following about each characteristic:

- The **Type of Record** character position (Leader/06) contains a one-character alphabetic code that is used to differentiate MARC records created for various types of content and material. The form of the material can be found in the 006/00 position. The category of the material is indicated in 007/00.

- The **Encoding Level** character position (Leader/17) contains a one-character alphanumeric code that indicates the fullness of the bibliographic information and/or content designation in the bibliographic record.
- The **Descriptive Cataloging Form** character position (Leader/18) contains a one-character alphanumeric code that indicates characteristics of the descriptive data in the record through reference to cataloging norms. The code particularly indicates whether the descriptive part of the record exemplifies the rules of the International Standard Bibliographic Description (ISBD), either within or outside of the framework of the Anglo-American Cataloging Rules, 2nd Edition (AACR 2).
- The **Source of Cataloging** (whether National bibliographic agency, cooperative cataloging program, or other) is indicated by a one-character alphanumeric code (008/39).

The Type of Record is important in describing the types of materials within the dataset such as books and sound recordings. The Encoding level gives the fullness of the record. The Descriptive Cataloging Form explains the conventions used to catalog the resource. Similarly, the Source of Cataloging may give insight to the quality of the cataloging by showing whether the record was created by a national bibliographic agency, a participant in a cooperative cataloging program, or some other organization.

Because of the role of the OCLC number as a record identifier during the analysis, all records will be examined to ensure there is an OCLC number. The OCLC number will be found in the 001 field in each record.

The analysis for developing the descriptive profile consists of creating frequency counts of specific values in each of the identified fields for the characteristics outlined above. These analysis procedures were performed on the full MARC 21 records in the test dataset after it had been loaded on the reference implementation online catalog.

2.1 Goal of Procedures for Developing the Descriptive Profile

The goal of these procedures is to create a descriptive profile of the dataset by identifying specific characteristics of the records based on frequency of occurrences of data values in selected areas of the MARC 21 record.

2.2 Detailed Analysis Procedures

To identify the Type of Record, each record must be analyzed for specific values in the Leader/06, 006/00, and 007/00 positions. The occurrence of each value will be calculated and reported in a frequency table. The following table represents the possible values and positions for each type of record.

Type of material	Area of Record	Values	Labels
Books	Leader/06	a or t	Language material; Manuscript language material
	006/00	a or t	Language of material; Manuscript language material
	007/00	T	Manuscript language material text
Printed or manuscript music	Leader/06	c or d	Printed music; Manuscript music
	006/00	c or d	Printed music; Manuscript music
	007/00	q	Notated music

Cartographic material	Leader/06	e or f	Cartographic Material; Manuscript cart. material
	006/00	e or f	Cartographic Material; Manuscript cart. material
Visual materials	Leader/06	g or k or r	Projected medium; 2-D nonprojectable graphic; 3-D artifact or naturally occurring object
	006/00	g or k or r	Projected medium; 2-D nonprojectable graphic; 3-D artifact or naturally occurring object
	007/00	f or g or k or m	Tactile material; Projected graphic; Nonprojected graphic; Motion picture
Sound recording	Leader/06	l or j	Nonmusical sound recording; Musical sound recording
	006/00	l or j	Nonmusical sound recording; Musical sound recording
	007/00	s	Sound recording
Computer file	Leader/06	m	Computer file
	006/00	m	Computer file
	007/00	c	Computer file
Archival/Mixed materials	Leader/06	p	Mixed materials
	006/00	p	Mixed materials
Serial	006/00	s	Serial
	Leader/07	b or s	Serial component part; Serial

To identify the Encoding Level, each record must be analyzed for specific values in the Leader/17 position. The occurrence of each value will be calculated and reported in a frequency table. The following table represents the possible values and associated descriptions for encoding levels.

Value	Label
blank	Full level
1	Full level, material not examined
2	Less-than-full, material not examined
3	Abbreviated level
4	Core level
5	Partial (preliminary) level
7	Minimal level
8	Prepublication level
u	Unknown
z	Not applicable

The Leader/18 position contains the data values for Descriptive Cataloging Form. The occurrence of each value will be reported in a frequency table. The following table displays the possible values and associated descriptions of those values.

Value	Label
blank	Non-ISBD
A	AACR 2
I	ISBD
U	unknown

Each record must be analyzed for specific values in the 008/39 position to determine the Source of Cataloging. The occurrence of each will be reported in a frequency table. The following table represents the possible values and associated descriptions for cataloging source.

Value	Label
blank	National bibliographic agency
c	Cooperative cataloging program
d	Other
u	Unknown
	No attempt to code

2.3 Reports and Outputs

As a result of the above procedures, frequency tables will be generated for each of the characteristics. Each frequency table summarizes the occurrences of each value for the various characteristics.

3. Overview of Procedures for Determining Content of Dataset

The testbed requires a controlled dataset that has been analyzed to determine what records should be or might be retrieved for the test searches that will be used in the interoperability testing. The analysis procedures outlined in this section result in a solid understanding of the contents of the test dataset. These procedures were performed on the decomposed MARC 21 records.

The analysis begins by selecting a term to find its occurrence in the dataset. To select a term we first performed a word frequency analysis on the entire dataset. The word frequency count resulted from analysis on the decomposed records after the data in those records were normalized (for description of the data normalization procedures, see the Z-Interop document **Data Normalization Procedures on Decomposed MARC 21 Records**). We then used terms selected from that analysis (e.g., terms that occurred in less than 300 of the records) to generate Aggregate Record Groups (i.e., a list of all records in which a term occurs). The decomposed records indicate the fields and subfields in which a word occurs. Further analysis on each Aggregate Record Group identifies the specific records where the word occurs in the specific fields and subfields. This analysis results in Candidate Record Groups associated with particular types of searches (e.g., Author, Title, Subject, etc.) for a particular search term.

The following hypothetical example illustrates the procedures and results. From the word frequency analysis, we select the term "rivers," a term that appears 300 times. We interrogate the decomposed MARC record dataset and determine which records contains the term "rivers." The records containing that term constitute the Aggregate Record Group for the term "rivers." Based on further analysis, we identify the term appears in the title fields of 200 records. These 200 records become the Candidate Record Group for a Title Keyword test search in the interoperability testing.

3.1 General Goal of Procedures

The general goal of the procedures for determining the content of the dataset is to define a systematic means of selecting terms to be used in subsequent analyses, to identify all of the records where those terms appear, and to create subsets of records for use in the interoperability testing.

3.2 Procedure: Word Frequency Count

To systematically choose terms that occur in records for use in subsequent analyses, we conducted a complete word frequency analysis on the dataset. One word may appear more than once in a record. The word frequency count should give every occurrence of all words, even repetition within a record.

For the purposes of this procedure, a word is defined as: *a string of characters bounded by spaces*. Hyphenated words and contractions will be treated as a single word.

3.2.1 Goal of Procedure

The goal of this procedure is to identify terms that can be used in test searches during interoperability testing. Since some manual examination of records returned during interoperability testing may be necessary, it is practical to use search terms that will result in a relatively manageable number of records. Less frequently occurring words in the dataset will be chosen for the subsequent analyses listed below.

3.2.2 Detailed Analysis Procedures

The dataset was analyzed for the occurrences of all words. Each word was counted each time it occurred in a record. The result of these procedures is a list of all words and their occurrences.

3.2.3 Reports and Outputs

Two reports were generated based on these procedures:

1. A list of words ordered by their frequency of occurrence in the dataset
2. The list of words ordered alphabetically with the associated frequency count.

3.3 Procedure: Identifying the Aggregate Record Groups

Aggregate Record Groups consist of those records that contain a selected word somewhere in the record. The Aggregate Record Group serves as the upper boundary for the number of records retrieved for any specific search during interoperability testing since no other records contain that particular search term. From this Aggregate Record Group, subsets of Candidate Record Groups can be identified that provide yet another measure of what records might be retrieved for a particular kind of search using a particular search term.

3.3.1 Goal of Procedure

The goal of this procedure is to create a set of Aggregate Record Groups that contain selected terms based on the word frequency analysis.

3.3.2 Detailed Analysis Procedures

A list of words (e.g., 10) were chosen from the word frequency count analysis. The researchers selected the words based on their occurrence in the dataset. An initial threshold was that a candidate word cannot occur more than 300 times in the word frequency count. This threshold may be too low or too high, and practical experience with the dataset and the terms guided the final decision on the threshold.

For each selected word, the content of the decomposed records was examined to identify the specific records in which the term appears. The analysis reports the OCLC Number and other information from the record in which the word appears. These records constitute the Aggregate Record Group for the word.

3.3.3 Reports and Outputs

For each word analyzed, two reports were generated. One report contains only the list of OCLC numbers for records that contain the word. The second report contains the following information:

- OCLC Number
- Field Tag(s)
- Subfield code (s).

This second report is also represented as a data file upon which subsequent analysis can be performed.

3.4 Procedure: Identifying the Level 0 Profile Search Candidate Record Groups

Once the Aggregate Record Group is established for a term, the group is further analyzed to identify the Candidate Record Group for a particular kind of search. This analysis is associated with the occurrence of the term in a set of appropriate fields/subfields containing data related to:

- Author
- Title
- Subject

For Phase 1 testing, the set of Candidate Record Groups produced is associated with the following Bath and US National Profiles' Functional Area A, Level 0 searches:

- Author Search – Keyword (Bath and US National Profile)
- Title Search – Keyword (Bath and US National Profiles)
- Subject Search – Keyword (Bath and US National Profiles)
- Any Search – Keyword (Bath and US National Profiles)

This procedure examines the decomposed records to find records where the word appears in specified fields/subfields appropriate for these searches. The analysis results in a list of OCLC numbers along with the field(s)/subfield(s) in which the word appears.

Repeating the example from above, the word “rivers” appears in 300 records (Aggregate Record Group). Further analysis identifies that the word occurs in one or more title-related fields/subfields in 200 records. These records constitute the Candidate Record Group for this word for a Title Keyword Search.

The Candidate Record Group report contains a list of all records that meet the field/subfield criteria.

3.4.1 Goal of Procedure

The goal of this procedure is to identify the Candidate Record Groups in which a specified word appears in title, subject, and author-related fields/subfields. Candidate Record Groups are used as a basis for establishing a benchmark in the Z-Interop Reference Implementations and ultimately in assessing the results of test searches of participant implementations during interoperability testing.

3.4.2 Detailed Analysis Procedures

The analysis procedures for determining Candidate Record Groups for keyword searches follow a similar pattern. The following sections provide the details on the analysis procedures.

3.4.2.1 Title Keyword Candidate Record Group

The analysis procedure examined the decomposed records the occurrence of a selected word in title-related fields and subfields. The Z-Interop document, **Indexing Guidelines to Support Z39.50 Profile Searches**, provides the basis for the title-related fields/subfields interrogated. Every record that contains the term in any one of specified fields/subfields is included in the Title Keyword Candidate Record Group. These records and only these records should be retrieved in a Title Keyword search during benchmarking and interoperability testing.

3.4.2.2 Author Keyword Candidate Record Group

The analysis procedure examined the decomposed records the occurrence of a selected word in author-related fields and subfields. The Z-Interop document, **Indexing Guidelines to Support Z39.50 Profile Searches**, provides the basis for the author-related fields/subfields interrogated. Every record that contains the term in any one of specified fields/subfields is included in the Author Keyword Candidate Record Group. These records and only these records should be retrieved in an Author Keyword search during benchmarking and interoperability testing.

3.4.2.3 Subject Keyword Candidate Record Group

The analysis procedure examined the decomposed records the occurrence of a selected word in subject-related fields and subfields. The Z-Interop document, **Indexing Guidelines to Support Z39.50 Profile Searches**, provides the basis for the subject-related fields/subfields interrogated. Every record that contains the term in any one of specified fields/subfields is included in the Subject Keyword Candidate Record Group. These records and only these records should be retrieved in a Subject Keyword search during benchmarking and interoperability testing.

3.4.2.4 Any Keyword Candidate Record Group

In the case of Any Keyword searches, the Aggregate Record Group for a word could be considered the Candidate Record Group for that word. The definition of an Any Keyword search is problematic. At the very least, the word should at least appear in one or more of the author, title, and subject fields/subfields. A conservative Candidate Record Group contains only those records in which the word appears in more author, title, and subject fields/subfields. This Candidate Record Group is logically a union set of the Title, Author, and Subject Candidate Record Groups for a particular term.

3.4.3 Reports and Outputs

For each of the procedures resulting in a Candidate Record Group, a report will be generated that contains the following:

- OCLC Number
- List of fields and subfields (and their contents) in which the term appears.

Based on these procedures, we now have Candidate Record Groups for each type of keyword search and each type of keyword truncation search.

4. Summary of Functionality Needed for Tools to Conduct Analyses

The procedures listed previously assume that software can conduct the necessary analyses and produce the required reports. This section summarizes briefly the functionality needed in the software for each of the analyses.

4.1. Descriptive Profile of Test Dataset

4.1.1.a Analysis of Type of Material

Examine the Leader/06, 006/00, and 007/00 for occurrences of the codes:

a, b, c, d, e, f, g, i, j, k, m, p, q, or t

to identify the type of the records. Create a frequency table for each type of material and the number of times each value occurs.

4.1.1.b. Reporting

Report results as a frequency count by the following categories of type:

Books	#of records
Printed or manuscript music	#of records
Cartographic material	#of records
Visual materials	#of records
Sound recording	#of records
Computer file	#of records
Archival/Mixed materials	#of records
Serial	#of records

4.1.2.a. Analysis of Encoding Level

Examine the Leader/17 for occurrences of the codes:

blank, 1, 2, 3, 4, 5, 7, 8, u, or z

to identify the encoding level of the record. Create a frequency count of each value.

4.1.2.b. Reporting

Report results as a frequency count by the following encoding levels:

blank	#of records
1	#of records
2	#of records
3	#of records
4	#of records
5	#of records
7	#of records
8	#of records
u	#of records
z	#of records

4.1.3.a. Analysis of Descriptive Cataloging Form

Examine the Leader/18 for occurrences of the codes:

blank, a, i, or u

to identify the descriptive cataloging form. Create a frequency count of each value.

4.1.3.b. Reporting

Report results as a frequency count by the following codes for the descriptive cataloging form:

blank	#of records
a	#of records
i	#of records
u	#of records

4.1.4.a. Analysis of Cataloging Source

Examine the 008/39 for occurrences of the codes

blank, c, d, u, or |

to identify the source of cataloging. Create a frequency count of each value.

4.1.4.b. Reporting

Report results as a frequency count by the following codes for the source of cataloging:

blank	#of records
c	#of records
d	#of records
u	#of records
	#of records

4.2. Word Frequency Count

4.2.1.a. Analysis

Examine every field and subfield in every record for the occurrence of a word. Count every occurrence of a word and how many times it appears in a record. A word is defined as: *a string of characters bounded by spaces*. Hyphenated words and contractions and plurals/possessives will be treated as a single word. A standard list of stop words will be used to remove some of the most commonly occurring words from the analysis.

4.2.1.b. Reporting

Two reports will be generated:

- 1) a list of words ordered by their frequency of occurrence in the dataset
- 2) the list of words ordered alphabetically with the associated frequency count.

4.3. Identifying the Aggregate Record Groups

4.3.1.a. Analysis

Examine dataset for every field/subfield where a chosen term appears. Identify the records in which the term occurs and list the fields and subfields that contain the term.

4.3.1.b. Reporting

A report for each term that lists a brief record containing:

- OCLC Number
- List of fields and subfields (and their contents) in which the term appears.

4.4. Identifying the Keyword Candidate Record Groups

4.4.1.a. Analysis of Title Keyword Group

The tool should take a term, examine the Aggregate Record Group for that term, and find every instance where it appears in a field or subfield identified as containing title information. See **Introduction to the Indexing Guidelines for MARC21 Records to Support Z39.50 Profile Searches** for the list of Title-related fields and subfields to examine.

4.4.1.b. Reporting

A report for each term that lists a brief record containing:

- OCLC Number
- List of fields and subfields (and their contents) in which the term appears.

4.4.2.a. Analysis of Author Keyword Group

The tool should take a term, examine the Aggregate Record Group for that term, and find every instance where it appears in a field or subfield identified as containing author information. See **Introduction to the Indexing Guidelines for MARC21 Records to Support Z39.50 Profile Searches** for the list of Author-related fields and subfields to examine.

4.4.2.b. Reporting

A report for each term that lists a brief record containing:

- OCLC Number
- List of fields and subfields (and their contents) in which the term appears.

4.4.3.a. Analysis of Subject Keyword Group

The tool should take a term, examine the Aggregate Record Group for that term, and find every instance where it appears in a field or subfield identified as containing subject information. See **Introduction to the Indexing Guidelines for MARC21 Records to Support Z39.50 Profile Searches** for the list of Subject-related fields and subfields to examine.

4.4.3.b. Reporting

A report for each term that lists a brief record containing:

- OCLC Number
- List of fields and subfields (and their contents) in which the term appears.

4.4.4.a. Analysis of Any Keyword Group

The tool should take a term, examine the Aggregate Record Group for that term, and find every instance where it appears. Minimally, title-related, author-related, and subject-related fields and subfields (see Appendices A, B, C) should be examined.

4.4.4.b. Reporting

A report for each term that lists a brief record containing:

- OCLC Number
- List of fields and subfields (and their contents) in which the term appears.