An Alternative Approach to Interoperability Testing

The Use of Special Diagnostic Records in the Context of Z39.50 and Online Library Catalogs

William E. Moen <wemoen@unt.edu>
JungWon Yoon <jungwonyoon@gmail.com>

School of Library and Information Sciences
Texas Center for Digital Knowledge
University of North Texas
Denton, TX  72603
Interoperability projects

- Funded by: U.S. Federal Institute of Museum and Library Services

- Z39.50 Interoperability Testbed, Phases 1 & 2
  - Improve Z39.50 semantic interoperability among libraries for information access and resource sharing
  - Establish and operate a testbed for interop testing of Z39.50 clients and servers with library catalogs (Phase 1: 2000-2003)
  - Explore alternative approach using Radioactive MARC Records (Phase 2: 2004-2005)
Factors affecting interoperability

- Multiple and disparate systems
  - Information retrieval systems, **search functionality**, etc.
- Multiple protocols
  - **Z39.50**, HTTP, SOAP, SRW/U, etc.
- Multiple data formats, syntax, metadata schemes
  - **MARC 21**, UNIMARC, XML, ISBD/AACR2-based, Dublin Core
- Multiple vocabularies, ontologies, disciplines
  - LCSH, MESH, AAT
- Multiple languages, multiple character sets
- **Indexing**, word normalization, and word extraction policies
Z-Interop Phase 1

- Test dataset: 400,000 MARC 21 records from OCLC
- Z39.50 reference implementations
  - Z-client, Z-server, information retrieval system
  - Configured to the profile specifications
- Test scenarios & searches
  - Searches with known result records from dataset
- Benchmarks
  - Results of test searches against reference implementations
- **Finding**: Interoperability improved dramatically using profile specs and common indexing policies
- **Issue**: Approach not suitable to interop testing for individual, local library systems
Phase 1 interop testing

Reference Z39.50 Client
Configured to Support Profile Specifications

Vendor Z39.50 Server
Configured by Vendor for Conformance to Profile

Test Dataset Loaded by Vendor or Library
Indexed by Vendor According to Vendor’s Specifications

Test Searches

Retrieval Benchmarks

Compared to

Retrieval Results
Z-Interop Phase 2

- **Radioactive MARC Records**: specially designed diagnostic records

- A set of test searches and automatic testing script that issues searches, retrieves records, and develops reports on the search and retrieval results

- A database of MARC documentation that enables the automatic identification of types of searches to issue
Radioactive MARC Records

RadMARC records loaded into database

Bibliographic Database for Online Catalog

Z39.50 Client and Interoperability Testing Management System

Uses automatic scripts to send issue queries and retrieve results

Information about RadMARC data and test searches stored in database

Analysis and Report Generation

Human mediated analysis and report writing

Automatically processes results, does preliminary analysis, and prepares for Z-Interop staff to write test results reports
Radioactive MARC records

- Specially designed diagnostic records
- Legitimate instance of MARC record structure
- Fields/subfields contain content-rich tokens
  - A token is a string of characters that has a specific structure and semantics that will serve as “words” or other data values in specific fields/subfields.
- Multiple sets of RadMARC records, distinguished by the amount of content designation populated
Structure of RadMARC tokens

- A single alpha character for left-hand padding.
  - Value = r
- A single alpha character to indicate the format of the material being described or type of record
  - Value = Selected values as defined in MARC Leader/06 – Type of Record or the Leader/07 – Bibliographic Level
- Three numbers indicating the Field Tag
  - Value = Defined in MARC 21 specifications
- A single integer to indicate number of occurrence the Field Tag
  - Value = Sequential number starting with 1
- A single alpha character to indicate the Subfield Code
  - Value = Defined in MARC 21 specifications
- A single integer indicating the offset within subfield
  - Value = Use the following scheme: 1=first token in subfield, 2=second token in subfield; 3=third token in subfield, etc.
- A single alpha character for right-hand padding
  - Value = r
Token example

- **ra2451a1r**
  - r - Left-hand padding
  - a - Type of record -- this is a Language Material type record
  - 245 - Field code
  - 1 – First occurrence of field in record
  - a - Subfield code
  - 1 - Offset within subfield, where 1 = first token in subfield
  - r - Right-hand padding

- **RadMARC example record**
Test scripts

- Automate interoperability testing and reporting
- Test searches defined by Bath Profile and US National Z39.50 Profile for Library Applications
- RadioMARC Perl module
  - Automatically generates Z39.50 queries with tokens as search terms
  - Sends searches to target servers known to contain copies of specific records
  - Generates reports dependent on whether or not the expected record(s) is present in the result set

**Sample output of testing**
MARCdocs database

- Pilot effort aimed at structuring MARC 21 documentation into a relational database
- Stores information about all content designation available in the MARC 21 Format for Bibliographic Data specifications
- Stores additional information about profile-defined searches necessary to the automatic test scripts
- Implementation uses MySQL and PhP
- Example display from MARCdocs
- Special data in RadioMARCdocs
Question space for Z-Interop2

- **Profile conformance level**: Addresses the interoperability between the Z-client and Z-server

- **Information retrieval (IR) system level**: Addresses the capability of the IR system underlying the online catalog application (e.g., types of searching)

- **Metadata record level**: Concerned with how the IR system indexes fields in the metadata record

- **Data content level**: Addresses normalization of data, hyphenated words, special characters and diacritics, etc.
So far, so good….

- Verified procedures and test scripts with the Z-Interop reference implementation server
- Completed testing with local library
  - Loaded RadMARC records successfully
  - Used the test script and procedures to issue searches
- Created two sets of RadMARC records
RadMARC record sets

- What content designation should be populated in RadMARC records to support interoperability testing?
- MARC 21 defines approximately 2,000 structures for holding data
- Z-Interop2 approach
  - Develop multiple RadMARC record sets
  - Increasing amount of content designation populated
- Informed by MARC content designation analysis
- More on this analysis in Metadata Quality and Evaluation Panel, Tuesday, 1:30pm
### Fields used in Z-Interop dataset

<table>
<thead>
<tr>
<th>MARC 21 Field Groups</th>
<th>Currently Defined</th>
<th>Obsolete</th>
<th>Unlikely Used</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>00x</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>0xx</td>
<td>96</td>
<td>1</td>
<td>33</td>
<td>130</td>
</tr>
<tr>
<td>1xx</td>
<td>49</td>
<td>0</td>
<td>2</td>
<td>51</td>
</tr>
<tr>
<td>2xx</td>
<td>81</td>
<td>0</td>
<td>19</td>
<td>100</td>
</tr>
<tr>
<td>3xx</td>
<td>23</td>
<td>6</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>4xx</td>
<td>10</td>
<td>0</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>5xx</td>
<td>128</td>
<td>1</td>
<td>3</td>
<td>132</td>
</tr>
<tr>
<td>6xx</td>
<td>104</td>
<td>1</td>
<td>7</td>
<td>112</td>
</tr>
<tr>
<td>7xx</td>
<td>205</td>
<td>0</td>
<td>5</td>
<td>210</td>
</tr>
<tr>
<td>8xx</td>
<td>105</td>
<td>3</td>
<td>8</td>
<td>116</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>807</strong></td>
<td><strong>12</strong></td>
<td><strong>107</strong></td>
<td><strong>926</strong></td>
</tr>
</tbody>
</table>
Occurrence summary

Total number of fields/subfields instances in dataset = 13,849,499

<table>
<thead>
<tr>
<th>Frequency</th>
<th># of Fields/Subfields</th>
<th>% of All Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 600,000</td>
<td>1</td>
<td>4.4%</td>
</tr>
<tr>
<td>500,000 &gt; 599,999</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>400,000 &gt; 499,999</td>
<td>13</td>
<td>39.9%</td>
</tr>
<tr>
<td>300,000 &gt; 399,999</td>
<td>6</td>
<td>14.3%</td>
</tr>
<tr>
<td>200,000 &gt; 299,999</td>
<td>6</td>
<td>10.6%</td>
</tr>
<tr>
<td>100,000 &gt; 199,999</td>
<td>10</td>
<td>10.3%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>36</strong></td>
<td><strong>79.5%</strong></td>
</tr>
</tbody>
</table>

Only 4% of all fields/subfields account for 80% of all occurrences or 96% of all fields/subfields account for 20% of all occurrences
Indexing & MARC

- *Indexing Guidelines to Support Z39.50 Profile Searches* (available on Z-Interop website)

- Identified **all MARC 21** fields/subfields that can contain author, title, or subject data
  - Author-related fields/subfields: 119
  - AuthorTitle-related fields/subfields: 21
  - Title-related fields/subfields: 253
  - Subject-related fields/subfields: 144
Occurrences in test dataset

- 537 fields/subfields can contain author, title, subject data
- 381 of these **actually occur** in Z-Interop dataset
- Total occurrences of the 381 = 4,397,712
- 19 of the 381 (5%) account for 80% of all occurrences
  - 9 of 19 are subject-related
  - 5 of 19 are author-related
  - 5 of 19 are title-related
- Preliminary testing using only 19 indexed fields:
  - 95% - 100% of correct records retrieved!
- The 19 fields/subfields
Initial RadMARC sets

- **Set 1**
  - 10 records
  - Populate 19 most frequently occurring Author, Title, Subject fields
  - Distinguished by types of materials cataloged

- **Set 2**
  - 4 records (100, 110, 111, 130 main entry fields)
  - Populate the Author, Title, Subject fields occurring 1000 or more times (approximately 50 fields/subfields populated)
  - [Sample Set 2 RadMARC Record](#)
Extensibility of RadMARC

- Records can be as simple or as complex as needed
- Custom records to interrogate system behavior for a library that wants specific assessment of indexing or other policies
- Assess normalization of characters
- Testing transformation from one metadata scheme to another
  - MARC Record
  - MARCXML Transformation
  - MODS Transformation
  - DC Transformation
- Other metadata environments?
Concluding thoughts

- Exploring an innovative conceptual and technical approach for interoperability testing.
- Conducting a proof-of-concept for a radioactive metadata record approach for diagnosing interoperability factors in an identified question space.
- Extensible in terms of the current focus.
- Extensible to other application environments, metadata schemes, and protocols.
References

- Z39.50 Interoperability Testbed
  - [http://www.unt.edu/zinterop/](http://www.unt.edu/zinterop/)

- MARC Content Designation Utilization Project
  - [http://www.mcdu.unt.edu/](http://www.mcdu.unt.edu/)

- Indexing Guidelines to Support Z39.50 Profile Searches

- RadioMARC Perl module
  - [http://search.cpan.org/~mirk/Net-Z3950-RadioMARC-0.06/](http://search.cpan.org/~mirk/Net-Z3950-RadioMARC-0.06/)

- MARCdocs Database (public interface)
  - [http://meta.lis.unt.edu/MARCdocs2](http://meta.lis.unt.edu/MARCdocs2)