Realizing the Vision of
Networked Access to Library Resources

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

Ten Month Status Report to
The Institute of Museum and Library Services

May 1, 2001 through March 31, 2002

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

March 31, 2002
Introduction

This document provides a status report on the Z39.50 Interoperability Testbed Project (Z-Interop) covering the period of May 1, 2000 through March 31, 2002. Due to a family medical emergency, the Principal Investigator was not able to complete the six-month status report that was due December 1, 2001. The Principal Investigator (PI) communicated with IMLS staff about this situation, and had sent via email in November 2001 information about the progress of the project.

The Z-Interop project is at a major watershed in that we are ready to start Phase 1 interoperability testing on April 1, 2002. Much effort has been spent in the past 10 months preparing for testing. We highlight activities and accomplishments to communicate to IMLS progress on our project.

Accomplishments and Challenges

This section summarizes the key accomplishments and challenges. Subsequent sections discuss these in more detail.

Accomplishments:

- Configured Sun E220 server for loading software (June 2001)
- Installed SIRSI Unicorn software and received training (June – July 2001)
- Refined and enhanced analysis methodology for test dataset (May – December 2001)
- Established test scenarios, test searches, and benchmarks for testing (December 2001 – February 2002)
- Developed documents that describe testbed methodology, analysis procedures, testbed policies and procedures (September 2001 – March 2002)
- Redesigned Z-Interop website for Phase 1 interoperability testing (January – March 2002)
- Publicized the testbed through presentations and articles (Summer – Fall 2001)
- Presented paper on interoperability at the Joint Conference on Digital Libraries (June 2001)
- Presented poster session on interoperability testbed methods and procedures at the Annual Conference of the American Society for Information Science and Technology (November 2001)

Challenges

- Complexity of analyzing the test dataset as foundation for interoperability testing
- Hiring and retaining appropriate personnel.
Project Personnel

The Principal Investigator (PI) has committed time to the project throughout this period with a 50% effort in Summer 2001 and a 20% effort in Fall 2001 and Spring 2002.

Total staff working on the Z-Interop as of March 31, 2002:

- 2 graduate students (one Ph.D. and one Masters) working 15 hours per week
- 3 graduate students (one Ph.D. and two Masters) working 10 hours per week.

These students comprise the Z-Team.

In late Spring 2001, two of the four graduate students selected as Project Research Assistants withdrew. During Summer 2001, two Ph.D. students joined the Z-Team, one as a full Z-Team member (working 15 hours a week and receiving tuition support) and the other at 10 hours a week. These students have continued with the project in Fall 2001 and Spring 2002.

In December 2001, one of the original Z-Team members withdrew from the project. The PI recruited two new graduate students in January 2002, and they began working in February 2002. These students participate in the School of Library and Information Sciences distance education program and the project provides an opportunity to give remote as well as on-campus students research experience. The PI communicates regularly with these off-site students using email, telephone, and conferencing software (NetMeeting).

Project Management

Because of personnel changes in Spring 2001, the PI assumed overall project management responsibilities rather than delegating these to a Z-Team member. He meets with the entire Z-Team regularly and with individual team members frequently to guide them in their work. The PI is responsible for scheduling work, keeping Z-Team members on task, and moving the project forward.

Project Schedule

There has been some slippage in the project schedule due to a number of factors including changes in personnel and unanticipated complexity in the research aspects of the project. The withdrawal of team members in late Spring 2001 resulted in a slowdown on project activities while new members were recruited and trained. In addition, the PI’s personal situation with a family medical emergency also affected progress on the project.

In a research project, certain activities and the time to carry them out can only be cautiously estimated. While the original logic of the methods and analysis to prepare the test dataset and develop test scenarios has proved out, the details of carrying out the data analysis and data preparation was extremely time consuming and complex. This will be discussed below in more detail.
We are now ready, however, to begin the first phase of interoperability testing. We anticipate that the original time periods for and between phase 1 and phase 2 testing can be compressed. However, the PI will be requesting from IMLS an extension on the completion date of the project.

**Project Website**

The project website <http://www.unt.edu/zinterop/> has served as our vehicle for promoting and publicizing the project. In addition, we are using the website for online data collection of information on potential participants in the interoperability testing. Additional information about the documents on the website will be discussed below.

To get ready for the launch of Phase 1 testing, Z-Interop staff redesigned the website to emphasize the readiness of the project to begin interoperability testing. The new website is being launched as part of Phase 1 testing. The URL above points to the revised website.

**Project Technology and Software**

We acquired in May 2001 the Sun E220 server as the main platform for test dataset preparation, analysis, and interoperability testing. UNT’s Academic Computing Services configured the hardware and operating system in June 2001 in preparation for loading the SIRSI Unicorn software. This was a delay in equipment acquisition and affected the project schedule.

SIRSI staff loaded and configured the Unicorn software on the system in June 2001. In July 2001, Slavko Manojlvcich from Memorial University in Newfoundland and a SIRSI trainer provided a two-day training session on the Unicorn software. This provided a foundation for Z-Team members to continue configuration and setting indexing policies on the system.

In addition to the Sun E220 server, we also installed (at no cost to the project) a Sun workstation to perform a variety of development activities related to analyzing and preparing the test dataset. Z-Interop staff worked with public domain, open source software (e.g., MySQL) and wrote custom scripts and programs for data analysis.

**Test Dataset**

Much time during Summer/Fall 2001 and into Spring 2002 was spent in setting up and carrying out analysis procedures on the approximately 400,000 MARC 21 records contributed by OCLC from its WorldCat database. Developing this controlled test dataset presented methodological challenges given the size of the dataset and the complexity of the MARC records.

The first step was to transform the 400,000 records into a form suitable for the detailed analysis needed. This transformation resulted in the derivation of approximately 33,000,000 subrecords. OCLC provided this transformation.
The next steps were to develop procedures for interrogating those millions of records to identify which records contained specific words and alongside this to develop data normalization procedures that would provide accurate results in the interrogation. These analysis procedures are documented in several project reports (see section Project Documentation below).

Working with such a large number of records has been computing resource intensive. The Sun E220 server is able to handle the majority of the analysis procedures, but some of those procedures take hours of run time.

A Z-Team member has developed batch processing and automated many of the steps involved in the analysis, but the programming to do these took additional time. The benefit of automating these procedures is that once the programming has been completed, subsequent analysis procedures will take less time.

We anticipate these analysis procedures will be of interest to the broader information science community and expect to prepare a conference paper or article that reports on our activities.

Project Documentation

Z-Interop staff worked during this period to document a variety of activities in the project and also developed documents that reflect the overall logic and analysis procedures used to prepare the testbed. Documents evolved as analysis procedures became better understood. Now as we prepare for Phase 1 testing, the documents have been revised and formatted for publication and review.

The following section describes the various documents. The order in which the documents are listed reflect the logical and in most cases, the temporal order in which they were created. All documents are available on the project website, and most documents are available in MS Word (.doc) and Portable Document Format (.pdf) versions.

- **Analysis Logic and Procedures for Creating a Test Dataset of MARC 21 Records for the Z39.50 Interoperability Testbed, Phase 1 Testing** (revised draft dated January 1, 2002)
  This document explains the overall logic by which records were selected for the test dataset from OCLC's WorldCat database and introduces key concepts used in the Z-Interop testbed methodology: Aggregate and Candidate Record Groups. In addition, there is an explanation of the procedures for determining the aggregate and candidate record groups for specific test searches. These record groups provide the foundation for Z-Interop benchmarks and for analysis of interoperability testing results.

- **Decomposing MARC 21 Records for Analysis** (revised draft dated January 1, 2002)
  The first step in creating the aggregate and candidate record groups involved the decomposition of the 400,000 MARC 21 records into subrecords based on character strings bounded by spaces (i.e., words) in fields and subfields in the MARC 21 records. This document explains the logic and procedures for decomposing the records. The result of decomposing the 400,000 records was approximately 33,000,000 subrecords. OCLC
carried out this decomposition according to guidelines prepared collaboratively by the Z-Interop team and OCLC.

- **Data Normalization Procedures on Decomposed MARC 21 Records** (revised draft dated January 1, 2002)
  
  No data normalization was done on the MARC 21 records or during the decomposition of those records. Normalization was necessary to more efficiently carry out the procedures to create the aggregate and candidate record groups. This document describes the normalization procedures carried out on the decomposed records.

- **SQL Data Analysis Procedures to Create Aggregate and Candidate Record Groups on a Sample of Decomposed MARC Records, Phase 1 Testing** (revised draft dated January 1, 2002)
  
  Once the decomposed records had undergone data normalization, they were now ready for the procedures to create aggregate and candidate record groups. For Phase 1 testing, four US National and Bath Profile searches (Functional Area A, Level 0) searches are being tested. The procedures documented here describe how aggregate and candidate record groups were created for Author Keyword, Title Keyword, Subject Keyword, and Any Keyword searches for specific search terms. The procedures described were informed by the indexing guidelines developed as part of the Z-Interop project, the Texas Z39.50 Profile, and the Bath and US National Profiles.

- **Indexing Guidelines to Support Z39.50 Profile Searches** (revised draft dated February 1, 2002)
  
  To assure rigor in the testbed, we have developed a set of guidelines for indexing the MARC 21 records to support Z39.50 profile searches. Z-Interop staff used these guidelines to index the 400,000 MARC 21 records that comprise the Z-Interop reference implementation of the Z39.50 server and online catalog. The guidelines reflect input and revisions based on public review. At this point, indexing guidelines are available for author, title, and subject, and any keyword searches. The guidelines can be used by interoperability testbed participants.

  
  This report identified the MARC fields and subfields that are indexed in the Unicorn system to support various Z39.50 Bib-1 Use Attributes. This serves as a confirmation of the actual indexing policies set up on the reference implementation.

- **Z-Interop Interoperability Testing Policies and Procedures, Phase 1 Testing** (revised draft dated February 1, 2002)
  
  This document provides an overview and the details of the policies and procedures of the Z39.50 Interoperability Testbed Project. Specifically, the document lays out the responsibilities and obligations of the Z-Interop Testbed and the organizations that participate in interoperability testing. It includes an agreement for the appropriate use of the test dataset.

- **Agreement for the Appropriate Use of Test Dataset by Participants in the Z39.50 Interoperability Testbed (Third Party Vendor Agreement)**
  
  This document includes a Third Party Vendor Agreement that must be signed by Z-Interop participants for the use of the test dataset of 400,000 MARC 21 records from the OCLC WorldCat database.
Project staff will continue to develop documentation that is a record of our activities and the procedures used in operating the testbed during Phase 1 testing.

**Project Information Dissemination**

The PI has made numerous presentations since Sprint 2001 in which he included discussions about the project (and acknowledged IMLS funding in all cases). These have occurred in national and international meetings including:

- Joint Conference on Digital Libraries, Roanoke, VA, (June 2001)
- IFLA Council and General Conference, Boston, MA (August 2001)
- Access 2001, Winnipeg, Canada (September 2001)
- International Z39.50 Implementors Group Meeting, Boston Spa, UK (October 2001)

Of particular note was a paper by the PI, “Mapping the Interoperability Landscape for Information Retrieval,” that was accepted for presentation at the Joint Conference on Digital Libraries held in June 2001.

Z-Team members presented at a poster session for the ASIST Conference in November 2001 in Washington, DC. The poster session focused on the project’s activities related to interoperability assessment methodologies.

Throughout this time, the PI has been in personal contact with numerous Z39.50 vendors and implementers to raise awareness of the project and to solicit their involvement in the interoperability testbed.

**Summary and Next Steps**

The past ten months have been an exciting and challenging time for the project. We are now ready, however, for Phase 1 interoperability testing:

- The testbed methodology has been documented <http://www.unt.edu/zinterop/Z39.50InteroperabilityTestingPhaseOne02_2002.htm>
- The preparation of the test dataset for this phase of testing has been completed
- Test searches and benchmarks have been established
- Online data collection forms for interoperability testbed participants are available <http://www.unt.edu/zinterop/zParticipants/>
- The Call for Participation has been issued <http://www.unt.edu/zinterop/ZinteropNew/Documents/CallForParticipationMarch2002.pdf>.

The Z-Team is now working on test searches for Phase 2 interoperability testing that will be ready in Summer 2002.