Status Report
Period: Fall Semester 2006
(updated February 15)

Serhiy Polyakov

Mark Phillips

February 15, 2006

Table of Contents

1. Introduction

This report details the progress accomplished by Serhiy Polyakov in the IMLS Grant Partner Uplift Project during the fall semester 2006. The report also briefly covers encountered problems and activities related to the project.

The work has been primarily concentrated on the subset of the tasks identified by the scope of the project as:

- Research and develop workflow for installing Z39.50 and/or SRU/SRW gateways to both closed and open databases.
- Install 3 gateways onto MySQL databases hosted by the Texas State Library and Archive Commission.
- Install a minimum 1 and maximum of 3 gateways onto DAMS hosted by the Dallas Public Library, the University of Texas at Arlington, and Stephen F. Austin State University
- Install Z39.50 or SRU/SRW gateways onto DAMS hosted by THDU IMLS grant partners to enable interoperability with the THDI search interface.

2. SimpleServer

Based on the *Review of the Tools and Software to Support Interoperability* <http://spmachine.lis.unt.edu/thdi/aris/ToolsReview_sp_15July2006.doc> compiled in summer 2006 SimpleServer has been selected as one of the possible tools to install SRU/SRW gateways onto variety of DAMS. SimpleServer is a Perl module (API) to implement Z39.50, SRU and SRW servers over the databases.

First task was to identify programming requirements for SimpleServer. In fact these requirements are formally stated in documentation but those statements do not give much information about real complexity of the task. The test installation could give more understanding about this. SimpleServer and YAZ toolkit have been installed on the test server. Also, simple test MySQL database was generated with the goal of making it SRU/SRW accessible.

The procedure required to write a set of callback functions in Perl to convert SRU queries into SQL, store the results, translate the results into XML, etc. As it turned out this is not a trivial task because it requires thorough knowledge of Object Oriented Perl (augmented classes paradigm). Also knowledge of Z39.50 query structure is desirable. Second half of the task was to write a client for specific tasks.

Discussion with Mike Taylor (developer of SimpleServer) has helped to clarify a workflow for these tasks. However, test implementation failed to work and decision has been made to move the emphasis to the Lucene/SRU/SRW.

Outcomes:
Learned: basic Perl
Developed: SimpleServer and YAZ installation procedure
Problems: Perl functions could not be implemented because of complexity

3. From MySQL to SRW

To improve search functionality the possibilities of using Lucene information retrieval API have been explored. Regardless of the chosen approach, simplified data workflow may be described as DAMS -> Lucene -> SRU/SRW -> Texas Heritage Online.

There are several ways to ingest data from DAMS into Lucene index. One is developing Lucene connectors for DAMS that use JDBC. Another is exporting data from DAMS into XML files and then ingesting them from XML files using XML to Lucene parsers.

3.1. MySQL to XML export

TSLAC ARIS databases have been used as targets. The document *Parsing Records from TSLAC ARIS Databases into XML: Notes* http://spmachine.lis.unt.edu/thdi/aris/ARIStoXML_sp_15Dec2006.doc describes the details of the procedures and includes code used.

These procedures may be briefly summarized as follows.

1. Obtained access to TSLAC ARIS databases through phpmyadmin interface
2. Dumped content of the databases into SQL files
3. Uploaded data to test server at UNT
4. Developed XML schemas for XML files containing records in native format
5. Mapped native format to unqualified Dublin Core
6. Developed and run php scripts that flatten, if necessary, and imports data into XML files according to native and DC schemas

One of the purposes of developing XML schemas was utilizing them in XML to Lucene parser at the later phases.

Generated XML documents (limited to 1000 for each database) are available at:
http://spmachine.lis.unt.edu/thdi/aris/

Mapping to DC may have been done differently. For example, each Republic Claims object is stored in multiple images and it is unclear how to preserve order of the images in DC schema without using attributes (OAI-DC schema does not prescribe usage of attributes):
<identifier>image_frame1</identifier>
<identifier>image_frame2</identifier>

In native XML dump attributes have been used:
<image_URL frame="1">image_frame1</image_URL>
<image_URL frame="2">image_frame2</image_URL>

So, entities/objects can be composed of multiple images (they are also identifiers) and in some cases order matters.

Outcomes:
Learned: ARIS database semantics, XML Schema language
Developed: procedures for writing XML schemas and importing data from RDBMS into XML

3.2. XML to Lucene parser

Lucene API was installed and process was documented. Development and diagnostic tool Luke was installed for the purposes of debugging. It allows accesses to already existing Lucene indexes.

The parsers for converting/indexing data from XML documents into Lucene index have been developed. The parser comprise java program that reads XML documents from directory, indexes each file, and updates Lucene index. Developed parsers are specific for each XML schema. However, it is possible to develop universal parser that would utilize XML schemas for the input documents.

200 XML documents (derivatives of ARIS Service Records database) have been indexed into Lucene.

The document *XML to Lucene to SRW*
<http://spmachine.lis.unt.edu/thdi/aris/XMLtoLuceneToSRW_sp_15Jan2007.doc> describes the details of the procedures and includes code used.

Outcomes:
Learned: Lucene API
Developed: XML to Lucene parsers for each database
Things to watch: Some examples of the code provided in the literature and on the web work with Lucene 1.9 but not with Lucene 2.0 (latest version).

3.3. SWRLucene

Ralph Levan at OCLC has developed SRWLucen service that provides SRW interface to Lucene index. The service runs under Tomcat, a Servlet and Java Server Pages container.

Tomcat service has been installed on test server on port 8080. SRWLucene service has been deployed and configured for the test Lucene index containing 200 records from ARIS database.

The service can be accessed trough web interface:
http://spmachine.lis.unt.edu:8080/SRWLucene/search/THDI_ARIS_service_records
This will show Explain record for the database. Search and Index Browse functionality are available. To use the service with any kind of SOAP client the following pointer to Web Service Definition can be used:
http://spmachine.lis.unt.edu:8080/SRWLucene/search/THDI_ARIS_service_records?wsdl

The document *XML to Lucene to SRW* <http://spmachine.lis.unt.edu/thdi/aris/XMLtoLuceneToSRW_sp_15Jan2007.doc> describes the details of the procedures and includes code used.

Outcomes:
Learned: Tomcat, SRWLucene service
Developed: Procedure to install Tomcat and SRWLucene service
Things to watch: SRWLucene requires Java 1.5+, Fedora built-in Tomcat hard wired to use open source implementation of java (1.4.2) which comes with Fedora. Standalone JDK 1.5+ and Tomcat should be installed.

4. Other Activities

4.1. TH Online review

Reviewed Texas Heritage Online search interface www.thdi.org

4.2. LITA Annual Forum

Plumer, D. C., Polyakov, S., Phillips, M. (October, 2006) One Size Does Not Fit All: Multi-Component Federated Search. Concurrent session presentation at the Library & Information Technology Association 2006 National Forum, Nashville, TX.

4.3. SRU Workshop

Attended TSLAC workshop on SRU

4.4. Meetings at UNT

August 24, 2006
Dr. Bill Moen, Serhiy Polyakov.

Discussed three models of interoperability for the purposes of THDI Uplift Project:
A) Z39.50 installed on DB or DAMS. Use Simple Server, ZContent.
B) Dumping data and metadata from DAMS into IR system (i.e. Zebra) and provide Z39.50/SRU/SRW to the IR system
C) Use Lucene IR library and SRWLucene plugin to provide real time access to DAMS

October 16, 2006
Mark Philips, Serhiy Polyakov.
Discussed Keystone, Zebra, pyLucene.
Discussed methodology of dumping data from DAMS into XML and parsing XML records into Lucene index.

November 14, 2006
Dr. Bill Moen, Serhiy Polyakov.
Discussed XML schemas for exporting data from DAMS, problem of pointers to actual objects and availability of the objects.

November 17, 2006
Mark Philips, Dr. Bill Moen, Serhiy Polyakov.
DAMS to XML dumping. Data dictionaries. Problems of installing Lucene on remote servers.
Models of aggregation and federation with SRWLucene.