DLF Aquifer Functional Requirements
Phase 1 Metadata Harvesting Service

Document prepared by Jon Dunn (Chair, Aquifer Technology/Architecture Working Group) and Martin Halbert (Chair, Services Working Group) for the Aquifer Implementation Group (AIG). Document was last revised 2005-10-12.

Introduction

The document is intended to provide a brief statement of functional requirements for the creation of metadata harvesting services as part of the first phase of the DLF Aquifer project (hereafter referred to as "Aquifer phase 1 harvesters" or simply "AP1 harvesters").

Background

This document references other DLF Aquifer planning documents (listed in the "references cited" section below) concerning services, technical architecture, metadata, and collections. On June 28, 2005 the DLF Aquifer Services Working Group (SWG) presented the AIG with an initial planning report proposing twelve core Aquifer services in the form of use case scenarios. Service #8 was entitled Metadata Harvesting. The basic aim of such a service would be to quickly assemble a union database of records contributed from many DLF Aquifer collections by means of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Value of this service

The benefit to both end-users and DL intermediaries of such a union database is generally understood to be searchable access to records describing the content of DLF Aquifer collections, and re-exposure of these records through a centralized repository of metadata. The service is potentially relevant to all of the Aquifer user personas, but the primary user is the DL Developer persona, who will use this service to implement local services using Aquifer content.

Audiences for this document

Implementers seeking to create a basic metadata harvesting system for the Aquifer project are the primary audience for this document. Secondary audiences include metadata providers who have installed and maintain OAI data providers for content relevant to Aquifer efforts. These DL staff will frequently be at different institutions than the staff operating the harvesting systems, but may wish to reference these specifications to understand the ways that their metadata will be harvested.

Relevant Aquifer Phase

This document only identifies basic requirements for the implementation of metadata harvesting services by DLF Aquifer institutions *during the first phase of the Aquifer project*. Different requirements may apply to metadata harvesting services created in subsequent phases of the DLF Aquifer project.

Implementation Target

AP1 harvesters are intended to be deployed during the first phase of the Aquifer project, before April 6, 2006.

Functional Requirements

All Aquifer phase 1 harvesters must be capable of meeting all of the following requirements.

1. *Must harvest using OAI-PMH:* AP1harvesters must be capable of harvesting records by means of OAI-PMH version 2.0 requests, and incrementally harvesting new/changed records on a regular basis.
2. *Must be able to harvest MODS records:* While most metadata harvesting systems are only capable of harvesting unqualified Dublin Core (DC) metadata records, AP1harvesters must be capable of harvesting MODS metadata records that are conformant to the DLF Aquifer MODS profile.
3. *Must harvest Dublin Core records:* While unqualified Dublin Core is a relatively unsophisticated metadata format, many metadata exchange systems are still based on it. Since AP1 harvesters are also intended to serve as data providers for ad hoc re-harvesting purposes, they must also be capable of harvesting and storing DC records.
4. *Metadata Normalization:* AP1harvesters must carry out some minimal normalization or enhancement of records in order to support searching, browsing, and re-exposure requirements. The specific elements that must be normalized are concerned with dates and formats.
    a. If available in harvested records, the DC *date* and the following MODS sub-elements must be normalized when possible by being encoded using the W3C Date Time Format profile of ISO 8601: *temporal, dateIssued, dateCreated, dateCaptured, dateValid, dateModified, copyrightDate,* and *dateOther*.
    b. If available in harvested records, the DC *format* and the MODS sub-element *internetMediaType* must be normalized by being encoded using MIME media types.
5. *Simple Search Interface:* AP1 harvesters must provide access to the union database via a simple Web searching and browsing interface. This interface should provide keyword and Boolean search functions.
6. *Re-Exposure of Harvested Records with Provenance Information:* AP1 harvesters must maintain an associated OAI-PMH version 2.0 compliant data provider that re-exposes all harvested records in both DC (per OAI-PMH requirement) and MODS formats. The re-exposed records must include a statement concerning provenance, clarifying where the records originated and a summary of the changes applied to them. Also, all records harvested from a particular OAI data provider source must be available for selective harvesting by being identified in an OAI set.
7. *Re-Exposure of Harvested Records through SRW/SRU:* AP1 harvesters must provide an SRW/SRU version 1.1 compliant server for use by other DLF Aquifer services.
    a. SRU (REST-style) interface is required at a minimum; SRW (SOAP-based) is optional.
    b. The server should support the Bath profile to the extent possible, including support for appropriate portions of following CQL context sets: cql, dc, bath, rec.
    c. The server must provide a means of restricting a search to a specific contributing institution and/or collection through the database name or an added search clause.
8. *Hosting Commitment:* AP1 harvesters must agree to keep the service in operation at least until the officially designated conclusion of Phase 2 of the DLF Aquifer project, with possible option to renew.
9. *Software Licensing Commitment:* Groups creating AP1 harvesters with funding from DLF Aquifer must agree to release all code developed to support this service as open source or via a non-exclusive license to DLF and DLF member institutions.

References Cited

Bath Profile Z39.50 Specification, Release 2. http://www.collectionscanada.ca/bath/tp-bath2-e.htm

CQL context sets (SRW/SRU Version 1.1). http://www.loc.gov/z3950/agency/zing/cql/context-sets.html

DLF Aquifer Services Initial Planning Report: Aquifer Services Working Group Initial Report, Including Information Gathered to Date, Survey Plan, Use Cases, and Next Steps. Presented June 28, 2005 to the Aquifer Implementation Group. http://www.diglib.org/aquifer/

DLF Aquifer Technology/Architecture Working Group Proposed Activities. Presented June 28, 2005 to the Aquifer Implementation Group. http://www.diglib.org/aquifer/

MIME Specification (Internet Media Types). http://www.iana.org/assignments/media-types/

SRW/SRU Version 1.1 Specification. http://www.loc.gov/z3950/agency/zing/srw/

W3CDTF Specification for Date and Time Formats, W3C Note. http://www.w3.org/TR/NOTE-datetime