Document Level Interoperability for Collection Creators

David Bainbridge
University of Waikato
Hillcrest Road
Hamilton, NZ
davidb@cs.waikato.ac.nz

Kaun-Yu (Jeffrey) Ke
DL Consulting
Innovation Park
Hamilton, NZ
jeffrey@dlconsulting.co.nz

Ian H. Witten
University of Waikato
Hillcrest Road
Hamilton, NZ
ihw@cs.waikato.ac.nz

ABSTRACT

Digital library interoperability for both documents and metadata is a critical and complex issue. Although many relevant standards have been developed, and continue to evolve, in practice things are not quite so easy as they seem. We have built a software environment called the Exchange Center that helps digital librarians manage the process of sourcing documents and metadata from various repositories, adding local content where necessary, and exporting the resulting collection into formats that are suitable for digital library repositories. This paper describes the software, which is built on Greenstone but does not require its use as the final digital library server.

Categories and Subject Descriptors: H.3.7 [Information storage and retrieval]: Digital Libraries.
General Terms: Design, Algorithms.
Keywords: Digital Library Interoperability, Software Architecture, Import and Export.

Introduction

Digital library interoperability is a critical and complex issue, with numerous facets. Standards are the key to interoperability, and the Open Archives Initiative (OAI) [2] and more recently the Metadata Encoding and Transmissions Standard (METS) [1] have made significant contributions. However, although these provide an essential foundation, it is still the case that achieving interoperability in a particular system's context often requires a programming solution that steps outside the standards and interfaces directly with a digital library system. This article describes a software environment that we call the Exchange Center, designed to help digital librarians manage the process of sourcing, combining, and exporting documents and metadata from various locations.

The Exchange Center

Greenstone, which is open source, is intended to be a flexible software architecture for digital library construction and delivery. Particularly germane to the present project is its importing component. This can stand alone—it does not have to result in a Greenstone collection—and includes a system of plugins for processing documents and metadata that is both extensive and extensible. There are plugins for literally dozens of widely-used formats. The import process combines documents and metadata into the METS format (Greenstone profile). Collections can be exported in other formats, for example DSpace [5]. The software includes an OAI server and a Z39.50 server, which are alternative ways of getting collection information out of Greenstone, as well as Greenstone's native Reader's Interface. See *www.greenstone.org* for further details of this software.

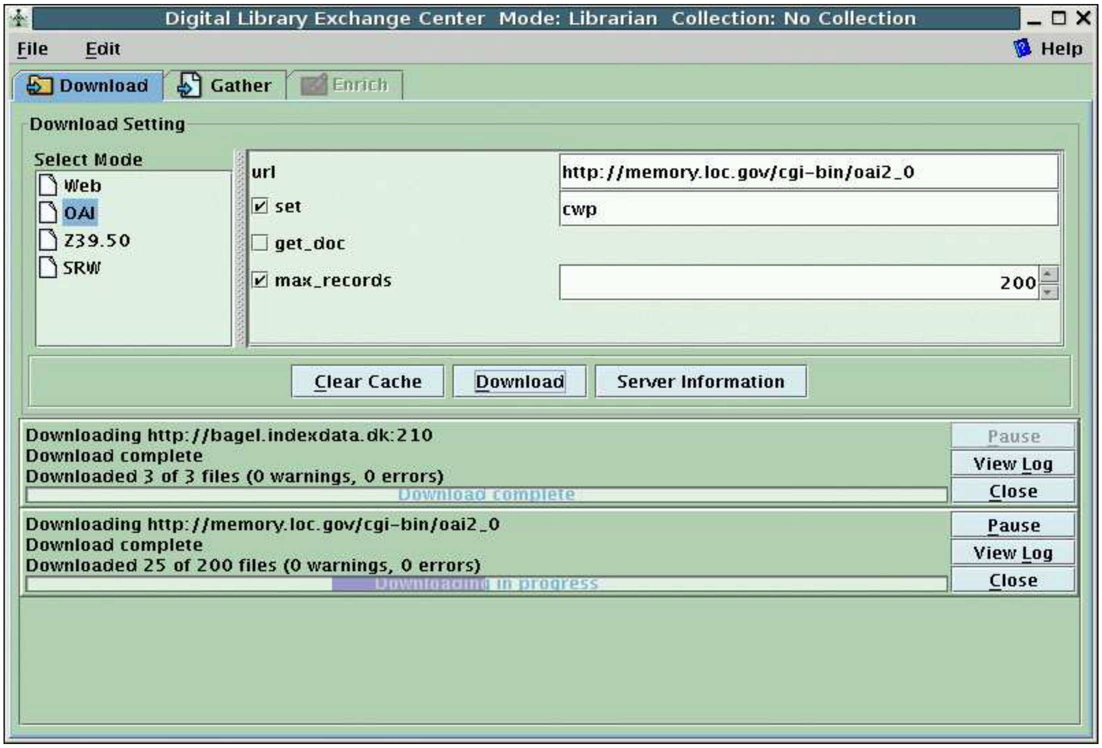
On top of this infrastructure is the Greenstone Librarian Interface (GLI) [4]. This is an interactive application that provides a graphical environment for digital librarians and others to gather, enrich, design, and create collections. GLI was originally designed specifically to create Greenstone collections from existing documents and metadata. However, this paper focuses on a recent adaptation that allows it to be used to assemble information for other digital library systems too. Since it was designed with flexibility in mind, it is not difficult to retarget it to fulfill this new role of a document and metadata exchange center.

Example of usage

Figure 1 shows the re-purposed graphical environment. Along the top are *Download*, *Gather* and *Enrich* tabs. *Gather* refers to the act of gathering together documents and metadata into a digital library collection. It involves dragging individual files, folders, or entire file hierarchies from a file browser and dropping them into the nascent collection. *Download* invokes utilities that bring in documents or metadata over various standard protocols (see below); this is a precursor to the gathering stage for collections in which the user reaches out, if they so desire, to other digital libraries—or to the web at large—to access their content. The downloaded files are placed in a cache where they can be examined and dragged into the collection. Finally, *Enrich* is used to add metadata manually to documents in the collection. If there are existing metadata files, they can be dragged into the collection just as documents are. However, users often need to add metadata to documents interactively. GLI is agnostic towards the actual metadata set (or sets) deployed, and facilities exist to define new ones if desired. Exporting the collection in different forms is achieved through an *Export* item on the *File* menu (now shown in figure).

Within the download panel of Figure 1, the upper left is

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.
Copyright 2006 ACM 1-59593-354-9/06/0006 ...$5.00.



Digital Library Exchange Center Mode: Librarian Collection: No Collection

Download Setting

Select Mode

- Web
- OAI
- Z39.50
- SRW

url http://memory.loc.gov/cgi-bin/oai2_0

☑ set cwp

□ get_doc

☑ max_records 200

Downloading http://bagel.indexdata.dk:210
Download complete
Downloaded 3 of 3 files (0 warnings, 0 errors)
Download complete

Downloading http://memory.loc.gov/cgi-bin/oai2_0
Download complete
Downloaded 25 of 200 files (0 warnings, 0 errors)
Downloading in progress

Figure 1: Exchange Center interface – Download Panel.

the operation area. This presents a list of protocols that can be used: currently Web, OAI, Z39.50, and SRW [3]. In this case OAI is selected. At the upper right is the configuration area, in which various settings relevant to the selected protocol are available. Here the user has entered the name of an OAI server, *http://memory.loc.gov/cgi-bin/oai2_0*, decided to download a maximum of 200 metadata records from the set *cwp* (Civil War Photographs), and pressed the *Download* button. In the lower portion of the panel is the download area. Here a progress bar has appeared (the lower one), providing feedback about how the download operation is proceeding. The upper progress bar shows the result of importing metadata over the SRW protocol, which was invoked earlier and is now complete.

Flexibility

Information can be downloaded over HTTP using standard web mirroring facilities, from OAI metadata repositories (including the documents themselves where applicable), and from Z39.50/SRW library servers. Of course, documents and metadata in local files can also be included in collections. This facilitates greater use of existing standards. Collections can be exported in a form suitable for DSpace's batch input process and in various METS profiles. There is also the option of applying a user specified XSLT to the generated XML allowing for an even wider range of export formats to be supported.

The system is by no means restricted to the current set of input and output protocols and formats. Extensibility is absolutely fundamental to the design. By supporting four very different protocols, HTTP, OAI-PMH, Z39.50, and SRW we have illustrated the flexibility and extensibility of the framework. These protocols span a wide spectrum of complexity, from pure browsing with HTTP, through the hybrid OAI-PMH which includes selective filtering, all the way to full and comprehensive search with Z39.50.

Summary

In summary, we have outlined an interactive system, the Exchange Center, that is intended to help practicing digital librarians solve practical interoperability problems. The work capitalizes upon the extensive and mature infrastructure previously developed to meet the needs of the Greenstone digital library project, making it available to users of other digital library systems.

The Exchange Center's *Download* panel provides easy-to-use interactive access to the features of various protocols. This radically lowers the complexity of information gathering from non-local information sources, which previously had to be done by invoking arcane scripts. We hope it will empower practicing digital librarians to undertake practical interoperability projects.

References

[1] M. Cundiff. An introduction to the Metadata Encoding and Transmission Standard (METS). *Library Hi Tech*, 22(1):52–64, 2004.

[2] C. Lagoze and H. Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In *Joint ACM/IEEE Conference on Digital Libraries*, pages 54–62, Roanoke, Virginia, June 2001.

[3] T. Storey. Moving Z39.50 to the web. *OCLC Newsletter*, 263, 2004.

[4] I. Witten and D. Bainbridge. Creating digital library collections with Greenstone. *Library Hi Tech*, 23(4):541–560, 2005.

[5] I. Witten, D. Bainbridge, R. Tansley, C.-Y. Huang, and K. Don. StoneD: A bridge between Greenstone and DSpace. *D-Lib Magazine*, 11(9), September 2005.