Terminology Services in a Digital Repository

The uses of controlled vocabularies in digital library applications can be expanded with ease when thesauri are made available using a standard service oriented architecture. Adopting this approach, the Indiana University Digital Library Program has been able to easily adapt existing tools to use controlled vocabularies and to better take advantage of a wide array of controlled vocabulary sources.

Service Architecture

As the culmination of an exploration with several other institutions and OCLC Research,[1] an approach of making commercial, public and locally maintained ontologies available using the standard SRU[2] protocol was decided upon. SRU, a RESTful web service (and its SOAP counterpart SRW), is a familiar and general purpose searching protocol that is flexible enough to support results in any XML format and queries of a wide variety of specificity. The selection of this protocol allows for easy use of existing server and client code without the need to create and document new standards. A talk reflecting the results of the first phase of testing this service-based approach was presented as part of a panel at the 2008 DLF Fall Forum.[3]

Application Enhancements

Almost every phase in the life of an object's metadata can be enhanced using terminology services. At metadata creation, terms can be selected or suggested for use in descriptive records. Indexes may consult vocabulary services in order to increase the amount of information available

[1] http://www.oclc.org/research/activities/termservices/default.htm

[2] Search and Retrieve via URL; http://www.loc.gov/standards/sru/

[3] Michael Durbin, Roy Tennant, and Diane Vizine-Goetz. "Using WorldCat Grid Services in Library Applications." Panel presentation, Digital Library Federation Fall Forum, Providence, RI, 2008. http://www.diglib.org/forums/fall2008/2008fallprogram.htm

for discovery applications by expanding entered terms in accordance with hierarchical or relationship data stored in vocabularies (such as the highly hierarchical data associated with geographic place names). User queries in search applications may be augmented, and browsing may be accommodated. If and when vocabularies change, this service may be consulted as part of automatic metadata migration processes to keep data in sync with current standards.

This talk will focus on the implementation and broad utility of terminology services at the Indiana University Digital Library Program. Some time will be spent on every phase of the process from the procedure for making thesauri available through the service to the integration into a wide array of applications.

Modularity was one major strength of this service based approach. Once the protocol (SRU) and result record schema had been established, work on clients could begin even before full or appropriate terminologies had been made available in the service. Furthermore tests could be completed against remotely hosted terminology services before bringing up local versions.

Discussion of the servers and clients will start from the addition of a new vocabulary into the service, then proceed through uses of the service in a chronology consistent with the lifecycle of a hypothetical object's metadata in the repository.

Methods for importing data from many forms into the Apache Lucene-based index backing the SRU service are all relatively straightforward, but necessarily varied. Discussion of techniques and tools used to parse XML, spreadsheets and custom relational databases will demonstrate the adaptability of this sort of approach.

The next portion of the talk will be an explanation of the integration of one or more vocabularies into metadata creation tools. Two main workflow tools and several methods of integration will be discussed. The approaches taken to adapt Indiana University's photograph cataloging tool PhotoCat[4] are generally applicable to any open source application and the exploration of effective and usable interfaces for leveraging data made available through the terminology service should serve as reference for similar approaches elsewhere.

In contrast to the approach of a comprehensive and custom metadata entry tool, this talk will also demonstrate and share code to quickly and easily integrate these services as a plug-in to the more general purpose commercial XML editing tool, Oxygen.

After showing how easily terminology services facilitate improved metadata entry, the presentation will shift to the access and discovery. Use of the well defined relationships between terms referenced in the metadata allow for substitutions, expansions and suggestions in user-entered queries. This presentation will cover the wide possibilities for discovery improvements based on real-time querying of the terminology service behind the scenes. A demonstration from both the user perspective as well as the technical architecture will be included.

Some improvements through the use of terminology services can be incorporated into the search index and need not require changes to the end-user application. This presentation will cover the considerations involved in implementing these sorts of improvements as well.

The presentation will conclude with a discussion of possibilities not yet implemented, an acknowledgement of lessons learned in the process and ways this approach leverages existing and widely adopted technologies.

[4] Muzaffer Ozakca and Jon W. Dunn. "PhotoCat: Implementing a Cataloging Tool for a Live Fedora Repository." Fedora Users Group, 4th International Conference on Open Repositories, Atlanta, Georgia, 2009. http://hdl.handle.net/1853/28530