Against the Grain

Volume 23 | Issue 1 Article 11

February 2011

Building and Maintaining Knowledge Bases for Open URL Link Resolvers -- Processes, Procedures, and Challenges

Christine Stohn
Ex Libris Group, Christine.Stohn@exlibrisgroup.com

Sherrard Ewing
Serials Solutions, Sherrard.Ewing@serialssolutions.com

Sheri Meares
EBSCO Information Services, sheri.meares@ebsco.com

Paul Moss
OCLC, mossp@oclc.org

Building and Maintaining Knowledge Bases for OpenURL Link Resolvers — Processes, Procedures, and Challenges

by **Christine Stohn** (SFX Product Manager, Ex Libris Group) <Christine.Stohn@exlibrisgroup.com>

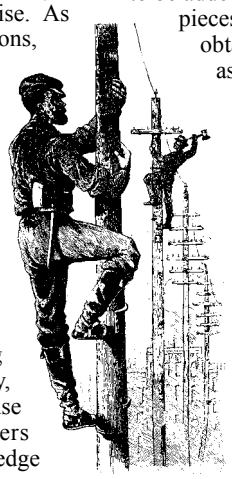
and **Sherrard Ewing** (Provider Relations Analyst, Serials Solutions) <Sherrard.Ewing@serialssolutions.com>

and **Sheri Meares** (E-Content Manger, EBSCO Information Services) <sheri.meares@ebsco.com>

and **Paul Moss** (Product Manager for the WorldCat knowledge base, OCLC) <mossp@oclc.org>

By covering as many available electronic resources as possible — both licensed and free — a global knowledge base seeks to make identifying and managing resources as easy as possible for individual institutions. Building and maintaining such a knowledge base involves a cycle of numerous processes, including building relationships with content providers; gathering data; validating, correcting, and enriching the data; converting it to the internal knowledge base format; performing quality assurance; and keeping the knowledge base up-to-date. A knowledge base with thousands of resources and millions of linked titles can receive data from several hundred providers. Such data can vary greatly in format and in the degree of accuracy, consistency, and completeness. The recommendations provided by the Knowledge Bases And Related Tools (KBART) working group are the answer to a clear need for a common format for this data supply.

A key task for any maintainer of a knowledge base is managing the relationship with content providers — agreeing on and organizing the data supply, formats, and frequency of updates. In addition, a knowledge base team works with content providers to identify and resolve any problems that might arise. As a starting point for these conversations, KBART can help facilitate an understanding of the benefits of knowledge bases, such as optimized visibility and increased usage.



Ideally, data from content providers comes in a consistent format and is updated frequently, platform or data changes are announced well in advance, and any changes required in the knowledge base can be tested before being released. However, not surprisingly, this scenario is rarely the case, because the requests that content providers receive often vary from one knowledge base vendor to another. This clearly shows the need for a consistent, agreed-on format for data delivery.

Much of the work associated with a knowledge base revolves around the correction and enrichment of data. The large amounts of data are bound to generate errors that can have many repercussions, such as an inaccurate availability status for a resource, title changes that are not recognized, and titles that are associated with the wrong package. Any problem in the data can cause a title to be unavailable to an end user at the point of need.

Various problems can occur with files supplied by content providers. For example, date coverage can be reported in many different formats, making it difficult for knowledge bases to process the data accurately. Another example is the parsing out of data incorporated in a string, such as "Vol. 2, no. 10 (Jan. 1996)-v. 5, no. 7 (Jan. 1999)"; if the provider changes this string, the parsing mechanism fails and has to be adjusted. Some providers furnish files in several formats, but the files may contain slightly different content; as a result, the content has to be compared and the correct version identified. Sometimes part of the data is missing and has to be added, either by requesting the required pieces from the content provider or by obtaining them from elsewhere such as Websites, listservs, alert services, and libraries.

Automation is of key importance for handling such large amount of data. However, the software used to automate the tasks must be able to handle many variations in data as well as errors and inaccuracies, all while delivering high-quality output. Data validation, correction, and enrichment therefore involve a combination of many automatic, semiautomatic, and manual processes. Using defined rules and routines, smart tools can automate processes such as the downloading of data, data extraction, data validation, corrections, and conversions. For example, a smart automatic tool can generate a holdings report from a content provider's Website, read and combine multiple spreadsheets, and construct the date coverage out of completely irregular publication dates. Because data can be supplied in many formats and can vary in accuracy, the number of rules and routines can easily be in the tens of thousands for a knowledge base of 2,000 to 3,000 packages.

The more complex the data validation, correction, and enrichment processes are, the greater the amount of work required for quality assurance. Tools that perform data validation and correction are usually designed to generate reports that the quality assurance team has to review manually for errors and inconsistencies in the data. By focusing on parsing a single format, as recommended by KBART, instead of a multiplicity of formats, a knowledge base provider would be able to spend significantly more time enriching content and assisting users than on fixing validation errors.

In an ideal world, the provider of a knowledge base would collect lists that contain all relevant data (metadata, date coverage, title relations, title changes, and cut-off dates for current and archival packages), are consistently formatted, and are available on a regular basis from the same location. Furthermore, all titles available from a content provider's platform would exhibit consistent linking syntax with no exceptions. Many content providers already meet at least some of these requirements, but other providers have yet to begin moving in this direction. KBART represents a significant milestone by bringing to light many of the issues faced by knowledge base providers and offering guidance to content providers to help standardize this work.

From a knowledge base provider's perspective, the recommendations developed by the KBART working group can help solve many of the issues described here. For the first time, a unified way in which content providers can supply data about their resources to all (or most) OpenURL link resolver knowledge bases has been proposed. A common format with consistent and accurate data lowers the risk of errors in knowledge bases, increases timeliness in the delivery of access to end users, reduces the effort required for correcting and comparing common data, and enables knowledge base developers to focus on enhancing and enriching the data to provide the best possible experience for users.

Rumors
from page 20

Got an email out of the blue the other day from another "true Brit" — **Liz Chapman** <e.chapman@lse.ac.uk>. **Liz's** daughter **Isabelle** is currently doing her **MFA** at **Parsons** in New York, and she is planning to travel to Charleston for Spring Break. **Liz** wanted to give **Isabelle** my phone number in case she needs a contact person. Sounds like a great motherly plan! **Liz** says she is going to NY to see **Isabelle**'s final exhibition in May and then she is heading to **The Fiesole Retreat** in St. Petersburg, May 11-13.

continued on page 28