Against the Grain

Volume 19 | Issue 4 Article 39

September 2007

Standards Column -- Standards, Scalability, and the Efficiency of Digital Libraries

Todd Carpenter
NISO, tcarpenter@niso.org

Standards Column — Standards, Scalability, and the Efficiency of Digital Libraries

by **Todd Carpenter** (Managing Director, NISO) <tcarpenter@niso.org>

Digital content has opened a world of new possibilities for users and librarians. The greater distribution digital content allows provides a range of benefits, from greater access to and uses of content to easier administration and simplified management-related data collection. But with these expanded benefits, there are also challenges that make maintenance of digital information as challenging as print — possibly more so. Whereas many of the distribution, organization, storage, and preservation issues have been long-resolved in the analog world, many problems related to digital distribution are only now being discovered and addressed. With digital content growing exponentially, the scalability of community and institutional workflows needs to be addressed.

Industries have relied on standardization to improve productivity for centuries. From the agreement on rail-gauge size in the 19th century to the modern-day light bulbs, consensus-based agreement on production methods have allowed for immeasurable advances in our capacity to create and distribute things of all types. Information is no different. From cataloging record formats and paper standards to OpenURL and RFID, we are seeing the impact of and benefits of standards grow and change in the library world.

But scalability is not just a library issue, and it's not just a publisher issue. It affects every organization in our community. Everyone from systems vendors and subscription agents, small society publishers and community college libraries, to commercial libraries and the largest commercial publishers — all of us are facing issues that are created by the scale of the long tail. The problem is that this vast information community nonetheless does not have the human time to manage this wealth of information in the same ways that it has done over the previous decade.

Initially, large university libraries, consortia, and the larger publishers were the first organizations to test electronic delivery and distribution. They were the groups with resources to invest in the technology and staff resources with the skills to build and maintain distribution systems for digital content. Of course, these organizations still are leaders in technological innovation and are constantly pushing the boundaries of information distribution. New functionality, tools, and discovery methods are constantly being added and improved.

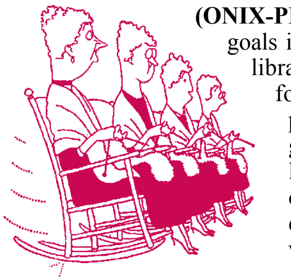
But in order for digital information distribution to be manageable on a broad scale, as print collections have been managed for decades, the acquisition, cataloging, maintenance, and preservation processes need to be standards-based so that this work can be accomplished efficiently. This is particularly true in the library environment, where staff and other resources are particularly limited and, even more so, are not growing to meet the pace of content growth.

Licensing, for example, was a reasonably manageable process when the number of digital products purchased numbered in the dozens or low hundreds. But imagine a "digital future", where most, if not all, content is acquired digitally. The average number of serials held by an **ARL** library is 40,598,[1] even presuming a generous 80% were in aggregated collections with a handful of licenses, the other 20% or more than 8,000 titles would still need to be individually managed. While most libraries are not as large as the **ARL** member libraries, the ratio of single titles would likely hold particularly in comparison to the library's staff size. There are simply not enough hours in the fiscal year to negotiate a subscription license for each product, or even a majority beyond the several hundred largest. While we certainly can't standardize business policies or purchasing activity, there are standards-based ways that we can use to try to address these problems.

> ***"Digital content has opened a world of new possibilities for users and librarians."***

Three examples of areas where **NISO** is working to help simplify and streamline the management of digital materials are found in the **Standardized Usage Statistics Harvesting Initiative (SUSHI)**, the **Simplified E-Resources Understanding (SERU) Working Group**, and the **License Expression Working Group (LEWG)**. All of these groups are developing consensus or standard-based solutions that are focused on alleviating bottlenecks in the distribution and management chain of digital content.

LEWG grew out of the work begun by the **Digital Library Federation's Electronic Resource Management Initiative (ERMI)**. The focus of the group was to determine an effective way to standardize the electronic encoding of license information into digital content management systems. Working with **EDItEUR**, **LEWG** has been developing a mapping **ERMI**'s license terms vocabulary to the widely used **ONIX** system of managing publications information and its new **Publications License (ONIX-PL)** format. Among the group's goals is to have a structure in which librarians can code their licenses for easy access and informing patrons of what rights were granted or prohibited in the license, without limiting rights or creating a machine-based enforcement system. Working with publishers, then, the work of **LEWG** is helping to create a template of rights that might be easily created and imported into an ERM system, improving the storage and distribution of information relating to license agreements.



The second licensing-related project currently underway at **NISO** involves the scalability of license negotiation. Since many of the core issues in license negotiation are the same from one license to another, the **SERU** project aims to capture the majority of the terms that are commonly agreed to in licenses into a community-based and publicly held set of understandings under which the sale of electronic products could be advanced without the use of a formal, negotiated, and signed license. **SERU**'s goal, then, is to create an agreed upon framework for the sale of digital products within a situation where best practices, rights, and responsibilities are commonly understood. While not meant for every situation or every publisher or library, the **SERU** process was designed with the "long tail" of publishing in mind, where negotiation of individual licenses may be impractical or unwieldy. Although **SERU** is not a standard, per se, it is a model based on community consensus is an example of new methods by which **NISO** can help facilitate the exchange of information.

Finally, **NISO**'s most recent standard, **SUSHI**, is just now wrapping up balloting. **SUSHI** facilitates the gathering and compiling of **COUNTER** usage reports through a client-server structure built into publisher and library systems. This Web service protocol allows subscriber-based ERM systems with **SUSHI** clients installed to automatically call to the numerous publisher systems requesting their specified **COUNTER** reports. The **SUSHI** server on the publisher's system receives the request, processes the reports, and packages and returns the reports via Web transfer protocols. **SUSHI** will help to alleviate one of the most challenging bottlenecks in managing and analyzing the use of digital materials. By utilizing an ERM system with **SUSHI** installed, librarians will be able to more effectively scale their oversight of online materials.

Similarly, other issues of usage measurement, cataloging, authentication, and preservation are unlikely to be manageable if they are not standardized. Many of **NISO**'s future activities will be aimed at further identifying and then working with the information community as a whole to come to agreement on the standards necessary to cope with the scale of digital information distribution. One of these projects, a series of Thought Leader

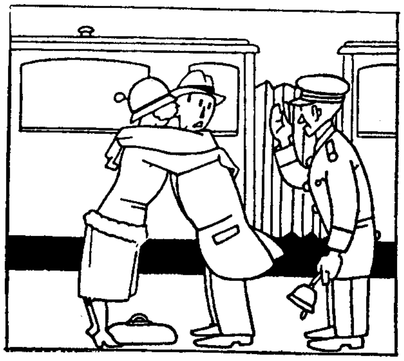
continued on page 78

I Hear the Train A Comin' — Penn Tags

Column Editor: **Greg Tananbaum** (Consulting Services at the Intersection of Technology, Content, and Academia) <gtananbaum@gmail.com>

As an elementary school-aged boy in the 1970s, I had very straightforward criterion for prospective friends. You had to drink Orange but not Purple Hi-C. This issue was *important*. It provided a sort of shorthand for me to determine compatibility. If you were a Purple Hi-C kid, I knew immediately that our broader interests were likely divergent. If you liked Orange Hi-C, I could trust your judgment on other key matters (like Star Wars action figures and Saturday morning cartoons). I broach the example of my younger self because so much of what we encounter within the Next Big Web Thing discussion today relies on sophisticated Hi-C litmus tests. **Facebook** and **MySpace** allow users to discover what is new and what is important among their peers by revealing commonalities within what people are reading, listening to, watching, and so forth. Twitter takes this to a new extreme. It connects people by revealing the connections within Joycean streams of consciousness posted by its users. Literally thousands of sites are devoted to a variation of "I like X," or "I read Y," or "I use Z." Why? First and foremost, because I want to meet people like me who value Orange Hi-C and disdain its purple counterpart. These people are potential friends. Beyond companionship, these like-minded souls can provide a valuable service. The information age breeds clutter, so much clutter that I need not just myself, but Proxy Me's, to cut through the tangle and help me uncover the music that I will love or the video that will make me laugh or the paper that will help my research. I need an army of Orange Hi-C drinkers at my disposal.

My column this issue focuses on one specific Hi-C tool, **PennTags**. **PennTags** represents the **University of Pennsylvania**'s attempt to cut through the clutter of Web resources by showing its users what like-minded community members value. It leverages the basic concept of popular sites **del.icio.us** and **Connotea**, namely that social bookmarking can provide important cues to the discovery of web-based information. Whereas these other sites are open clubs, **PennTags** establishes some preemptive commonality among its users by limiting participation to the **University of Pennsylvania** community. The assumption is that **Penn** researchers, by virtue of their engagement at the institution, have a shared universe of interests that is distinct from the larger social bookmarking alternatives. Indeed, the project was launched as a result of the **del.icio.us** experience of two librarians, **Michael Winkler** (Library Web Manager) and **Laurie Allen** (Research & Instructional Services Librarian). Both had used **del.icio.us** and enjoyed the ability to tell the world what Websites they were reading and browsing. However, they shared a frustration at the tool's inability to work with **Penn Library** resources, notably cataloged materials, proxy services, and other items that lacked stable URLs. When Cinema Studies Professor **Peter Decherney** assigned his students a project to collect Web-based resources about a specific film, **Winkler** and **Allen** realized that to do so effectively would require an easy way for students to grab and share Web pages from both outside and within the library's walled garden. This provided them the impetus for what has become **PennTags**.



The first iteration of **PennTags** was very rudimentary. Like many **Web 2.0** applications, it was characterized by a light "let's figure it out as we go along" approach. **Michael Winkler** created the basic code over a long weekend, modified it with feedback from **Laurie Allen** and a small group of self-identified interested parties, and delivered it to **Professor Decherney** for the fall 2005 semester. His students received extra credit if they used **PennTags** for the resource collection project. Almost all of the students did so and provided feedback. This helped **Winkler** further hone the feature set and user experience.

As the next semester opened, **PennTags** was soft-launched to the greater Penn community. **Penn** students, faculty, and staff could use the tool to tag records within the library catalog, any public Web pages, full-text article links via the library link resolver, and other sources of scholarly information. The largest limitation was — and remains — the inability to tag content within databases that maintain full text (e.g., **LexisNexis**).

The library did not publicize **PennTags** except to add a muted "Add to **PennTags**" link on an increasing number of **Penn** resources. Very little marketing or support was provided. In early 2006, **Mike Winkler** and **Laurie Allen** secured library management buy-in for the creation of a small working group that met weekly to discuss **PennTags** issues and features. Many code changes and feature additions resulted from these sessions. Nearly two years into the project, the **PennTags** team has not as yet done a formal launch or rollout campaign. Even absent this type of push, nearly one thousand users have picked it up along the way (current students, faculty, and staff — a pool totaling approximately 50,000 individuals — are eligible to use **PennTags**). This grassroots validation has prompted the **Penn** library to add resources to the project. A code rewrite and a more systematic release to the **Penn** community are both in the works as a result.

The **PennTags** footprint is a light one, designed to subtly enhance the research experience. The annotations a tagger makes are viewable both within the library catalog and via the **PennTags** site (*http://tags.library.upenn.edu*). There, visitors can search or browse by tag clouds, by contributor, and also by "project," in effect an annotated bibliography on a specific subject. The **PennTags** site also contains a number of end user productivity tools, such as the ability to convert tags of interest into RSS feeds.

For materials tagged within the catalog, the **PennTags** appear alongside more formal cataloging elements. For example, a book in the catalog will include the **PennTags** post (who tagged it and what the tags are) sitting right below the more formal bibliographic information and subject headings. Tags may be just a few short keywords or rather long discussions of a resource's merits. These tags appear via **Ajax** after the page loads so as not to slow down the user experience.

The Penn library, after much discussion with the university counsel's office, decided not to gatekeep annotations. The **PennTags** user interface includes a click-through agreement that precedes a user's first post, advising him

continued on page 79

Standards Column
from page 77

meetings, will begin this fall, generously funded by a grant from the **Andrew W. Mellon Foundation**. These meetings will explore and prioritize areas in need of standardization and will improve our community's productivity and scalability.

Much like standardization helped improve efficiencies in manufacturing and other areas, standards can help the community improve the process of creating, distributing, managing, and curating information. As the pace and number of organizations that are creating digital information continuing to increase exponentially, customized and individualized solutions need to transition to standards-based so that the community can deal with this increasing volume of content.

Endnotes

[1] **Association of Research Libraries**, ARL Statistics Tables 2004-05 — available at: *http://www.arl.org/bm~doc/05tables.xls*.