April 2006

International Dateline -- The Importance of Metadata and Interoperability

Peter T. Shepherd

Project COUNTER, pt_shepherd@hotmail.com

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Shepherd, Peter T. (2006) "International Dateline -- The Importance of Metadata and Interoperability," *Against the Grain*: Vol. 18: Iss. 2, Article 39.

DOI: <https://doi.org/10.7771/2380-176X.4668>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

search engine for scholarly literature, and especially **Google Book Search**, a service based on the digitization of millions of books obtained from publishers and five research libraries, have drawn the most attention. But just a little clicking here and there around the **Google** Website reveals an incredible lineup of projects and services underway, some familiar, some unfamiliar, many, like **Scholar** and **Book Search**, still in a beta or somewhat experimental mode. There's blog searching, image searching, map searching, satellite searching, video searching, catalog searching, searching of university Websites. There's **Google** toolbar, **Google** desktop, **Google** messaging, **Google** shopping, **Google** blogging, **Google** software, **Google** news, **Google** mail, and more, everything designed to extend the reach and potential profitability of the company.

New services are one way to grow. Another is to create new **Google** users in all parts of the world. Today in the search engine business, or any business for that matter, this means China, whose ungainly embrace of capitalism and population of 1.3 billion will at some point, everyone figures, make it the world's largest market for practically anything. While China has well over 100 million Internet users today, Internet use is even so considered early stage, meaning that, for **Google**, there's a lot more to this country than special tofu.

Google's difficulties in China are well known. In 2000 the company launched a Chinese-language version of **Google.com**, but government-mandated filtering of search results by Chinese Internet Service Providers made the service slow, unreliable, and at times entirely unavailable, frustrating users. **Google** stood fast, rejecting self-censorship. Meanwhile, though, competitors unhampered by the **Google** value system began to work with the Chinese government. **Google**

watched as its share of the search market in China plunged, while the market share of **Baidu**, **Google's** principal competitor in China, rose from 2.5 to 46 percent. This in a market expected to more than double by 2010.

So **Google** changed course. Earlier this year the company launched **Google.cn**, a version that **Google** itself, like the China-based ISPs, will filter according to Chinese law. Has **Google** done evil here? The company has carefully explained its decision. The Internet is transforming China, for the better. But Chinese Internet users were even so underserved. More information, for them, would be better than less information. Unfiltered, imperfect access to **Google.com** will remain. **Google.cn** will tell users when results have been filtered. **Google** will protect user privacy by not hosting mail and blogging services in China.

It all sounds more reasonable than evil, **Google's** "calibrated" approach to the China problem, as one company official put it in testimony before a U.S. congressional committee. Of course calibration, as **Google** does not deny, includes business measurements as well as ethical judgments. **Google** noted in its China testimony that more than 350 million mobile phones are in use today in that country. A different **Google** official, in a different setting, recently noted that the company's single biggest business opportunity right now was cell phone advertising.

That setting was a meeting of key investment analysts. Investors, **Google** has them now. They're the people who handed over that \$1.7 billion. At last year's meeting for this group, according to the *New York Times* of March 3, 2006, the analysts were put off because the principal speaker was none other than **Charlie Ayers**, the chef, who delivered a presentation on the day's lunch menu. This year, with the **Google** stock price down 20 percent from an earlier high, the centerpiece presentation was delivered by **Google's** chief financial officer, who gave the analysts a "detailed and sober" look at the state of the com-

pany. He assured analysts that **Google's** growth opportunities remain boundless.

Was there really any other choice in China than the one **Google** made? Surely most of us don't consider that decision an "evil" one. And just as surely, as **Google** looks for ways to please investors and maintain growth, it won't be the last tough decision the company will face. Next economic slump, what happens? What about all those neat new beta services? In particular, what about the ones concerning libraries? **Google** investors will understand how important they are. Won't they? They'd lay off a few of those chefs first, wouldn't they?

"Investors have been successful," said one of the investment analysts, "in communicating with the management they can't be this funky renegade company." How long, one might ask, can **Google** remain **Google**? Think of all the industries that revolutionized the world through someone's invention or new idea. Working backwards in rough chronological order, think of computer software, fast food, computer hardware, television networks, airlines, processed food, motion pictures, telephone companies, electric utilities, automobiles, department stores, railroads.

That puts us back in the mid-nineteenth century, meaning that this is getting to be an old story, businesses that start with a burst of excitement and utopian promise but in time become merely backdrop to our lives. At best, these companies are invisible pieces of necessary infrastructure, like sidewalks or banks; at worst, they are objects of ridicule or even contempt. These companies grew, they made choices, they became businesses like any other.

"Don't be evil" is the famous **Google** motto. Nobody I know of thinks of **Larry Page** and **Sergey Brin** as evil men. But these are two smart guys, after all. Early on, they must have known that one day, they'd need the reminder. 🐼

International Dateline — The Importance of Metadata and Interoperability

by **Dr. Peter T. Shepherd** (Project Director, COUNTER) <pt_shepherd@hotmail.com>

In one of the more famously precious exchanges in **Oscar Wilde's** "Importance of Being Ernest," **Cecily** declares her passion for plain speaking with the cliché "When I see a spade I call it a spade," only to be put down by **Gwendoline's** cutting reply "I am glad to say that I have never seen a spade. It is obvious that our social spheres have been widely different." Thus has it been for many of us in the publishing and library world. We all rely on

metadata (data about data), we all use metadata, as a tool it is as indispensable as the spade; but it is something that, on the whole, we assume is taken care of by someone else. We have moved in widely different social spheres.

Such can no longer be the case. The rapid advance of electronic publishing has created a growing need for an expanding range of reliable metadata to support the publishing and management of information resources.¹ In

addition to descriptive metadata to characterize content, we need technical metadata for technical processes, rights metadata for rights resolution and preservation metadata for digital archiving. Occurrences of metadata vary tremendously in richness. The strategic deci-

continued on page 81

sions publishers make about metadata often concern how much to expose. The CrossRef metadata set, for example, contains only a handful of required elements. For electronic bookselling, on the other hand, the richer the metadata the better. Hence **ONIX** (the **Online Information Exchange**) standard for books specifies over 200 elements. Metadata is potentially complex and definitely strategic; it is, therefore, important for producers (publishers) and consumers (librarians) to be actively involved in the monitoring and development of metadata standards. In other words, nowadays it is not enough to see the spade; we have to agree on what the spade is, how it should be described and defined.

An Information Environment

A number of organizations worldwide, such as **ISO** and **NISO**, are responsible for the development and maintenance of metadata standards. In the UK **JISC** (the **United Kingdom Joint Systems Committee**) provides, among other things, the **JISC Information Environment**,² which is designed to provide a range of services, tools and mechanisms to enable British colleges and universities to exploit fully the value of online resources and services for teaching, learning and research. One of the main drivers behind the **Information Environment** is the need for more seamless ways to discover quality resources. The academic community needs to be able to discover resources in simpler and more integrated ways than currently available.

Metadata and Interoperability

Publishers and aggregators have an opportunity to exploit this market by using protocols and standards that will enable seamless searching of their resources. Working closely with the **Publishers and Library/Learning Solutions (PALS)** Metadata and Interoperability working group, **JISC** has funded a number of research projects to explore a range of issues related to publisher adoption of metadata standards and protocols; it is currently supporting several designed to facilitate collaboration between the academic and publishing worlds as well as to develop practical solutions for metadata and interoperability. This programme is managed by **Christine Baldwin** and its objectives are:

- explore the issues around presenting content from a range of sources by developing interoperability demonstrators
- investigate the issues surrounding the integration of publisher-initiated standards into the **JISC Information Environment**
- develop tools and guidelines to simplify the effective adoption of metadata and interoperability standards
- take forward rights expression standards, for example by building implementations

of them in particular contexts or mapping between them.

The projects funded in the current programme include:

AIMSS: Automating Ingest of Metadata on Serial Subscriptions

Information about subscriptions to electronic journals is more complicated, and less well defined than for print journals. At present there is great variety in metadata, especially as these refer to bundles of journals on offer: both the metadata and the content of packages can be volatile. Receiving regular, comprehensive and accurate information that can be automatically transferred from computer to computer is essential for keeping the records of libraries up to date. This can only happen when there is an accurate and regular source of data, an agreed mode and format for transmission and a capacity to accept and process information. The **AIMSS** project addresses these issues.

Electronic Expression of Licensing Terms: Specifying Publisher Tools and Library Benefits

As the quantity of digital resources in library collections grows, libraries have increasing difficulty complying with the widely differing licence terms applied to resources by their creators and publishers. The ability to express these terms in a standard **XML** format, link them to digital resources and communicate them to users has become a pressing need with benefits to both publishers and libraries, including:

- Increased visibility of usage rights
- Instant answers to questions on terms
- Easier analysis of licenses and their terms
- Savings in administrative time dealing with the above issues
- Increased operability between publisher, intermediary and library systems
- Improved compliance with licences

This project is designed to promote the benefits of electronic expression of licensing terms, examine the difficulties that not-for-profit and smaller publishers, including learned societies, might have in generating an **XML** version of their library licences and show how tools and services could be developed to support them.

The project will work with smaller publishers to agree the benefits to them of mapping their licenses to the **XML ONIX for Licensing Terms** format and specify the tools and services required to enable this.

In addition to the analysis of publishers' requirements and the specification of tools and services, a significant benefit of the project will be the raising of awareness among smaller publishers of

work on standards for the electronic communication of licensing terms.

Electronic Expression of Licensing Terms: XML Expression of a Publisher/Library Licence — The object of this project is to provide an **XML** representation of a publisher/library licence that would allow libraries to have an actionable record of licensing terms within their systems, linked to the appropriate resources, without having to rekey the terms into their systems. This could enable terms to be displayed to users and avoid the current situation where librarians very often have no way of answering usage questions. This would help libraries comply with licensing terms and ensure that they benefited fully from those terms.

The **APS** Journal Legacy Content

OVER 100 YEARS
of Historical
Scientific Research

14 Journals of the **American
Physiological Society**
www.the-aps.org

TOCRoSS — Table of Contents by Really simple Syndication

TOCRoSS will use existing standards to place journal tables of content (TOC) data into a library catalogue without human intervention, thus improving the accuracy of records, saving time for library administrators and adding value for library patrons. The objectives of this project are to:

- Extend the **RSS 2.0** standard to encompass relevant components of standards such as **ONIX**, **Prism**, etc. and publish the resultant documentation in the pub-

continued on page 82

lic domain

- Use the extended **RSS** standard to provide a greater level of automation of interoperability between the publisher and a library through the integrated library system
- Deliver open source software components and **XML** schemas, which will be freely available for the whole of the academic library, publisher and integrated library system supplier communities
- Produce a demonstration implementation of a service utilising the standards and software as part of the project
- Report of the impact of this project for a particular academic institution

Dictate: Distributed Content Tagging Tool for EPrints — Social bookmarking systems such as **Connotea** and **del.icio.us** allow users to organize, share and discover Web re-

sources by using “tags” (i.e., keywords) to label and share them. This approach has, in turn, lead to the concept of a “folksonomy” — the collection tags associated by a group of users with a set of content, and the relationship between them. As a way of organising information, this is distinct from, and perhaps complementary to, traditional top-down, hierarchical taxonomies. **GNU EPrints** is a popular system for creating institutional repositories. However, in common with most institutional repository software, it has yet to employ tagging to enable users to organize better and discover the content held there.

The object of this project is to add bookmarking, tagging and recommendation features to institutional repositories running **GNU EPrints**, by extending that software to enable it to integrate with the existing social bookmarking systems. This should enable more effective organization and discovery of content held in **EPrints** institutional repositories.

Something Useful to Read on the Train

Gwendoline never travelled without her diary because “One should always have something sensational to read on the train.” For those of us who are more interested in useful rather than sensational reading, I recommend leaving the diary at home and instead taking advantage of wireless connections to catch up on the very practical research being sponsored by **JISC**. Further information on the projects described above, as well as the other metadata and interoperability projects currently supported by **JISC** can be found on the **JISC** Website.³ All are due to report their findings by 31 July 2006; some have already completed their work and the resulting tools are being applied in practice.

References

1. Metadata Demystified: A Guide for Publishers; Brand, Daly and Meyers; NISO/The Sheridan Press (2003).
2. http://www.jisc.ac.uk/about_info_env.html
3. http://www.jisc.ac.uk/index.cfm?name=programme_pals2

Standards — Libraries, Data Providers, and SUSHI: the Standardized Usage Statistics Harvesting Initiative

by **Adam Chandler** (Information Technology Librarian, Library Technical Services, Cornell University Library, 107B Olin Library, Ithaca, NY 14853-5301; Phone: 607-255-5760) <alc28@cornell.edu>

and **Tim Jewell** (Director, Information Resources, Collections and Scholarly Communication, University of Washington Libraries, Box 352900, Seattle, WA 981195-2900; Phone: 206-543-3890) <tjewell@u.washington.edu>

The sheer amount of usage data now available to librarians who manage electronic resources represents a true “embarrassment of riches:” while the information is extensive and necessary to their operations, keeping up with it all has become a major problem. E-resource management systems now available or in development will provide tools to store and manipulate the data but may not offer much help with “intake.”

The **Standardized Usage Statistics Harvesting Initiative (SUSHI)** discussed below is aimed at overcoming this problem, and while librarians and ILS vendors have already demonstrated strong interest in it, there are reasons why data providers like e-journal and database publishers and vendors should also embrace it. For example, as awareness and interest in **SUSHI** build within the library community, data providers that become early supporters and adopters can demonstrate their vision and responsiveness to the market, while those that don't may be rightly per-

ceived as lagging behind. And while there will be development costs associated with supporting **SUSHI**, those costs should be minimal and may lead over time to the need to invest less in developing custom usage data interfaces. Lastly, since the **SUSHI** protocol has already garnered support by **COUNTER**, it could conceivably become a requirement for **COUNTER** certification; work done now to implement it will position vendors and publishers to retain or establish certification.

COUNTER: Opportunities and Challenges

Under the skillful leadership of **Peter Shepherd** and the **COUNTER Board**, **COUNTER** is securely fixed among publishers and librarians as the standard for reporting online usage.¹ However, on the demand side of the fence, there are serious impediments to making use of the **COUNTER** reports. The formatting of the delimited reports (either comma separated or in **Excel**) is widely variable across publishers,

which makes processing them automatically difficult or impossible. Though steps are being taken to make the reports more consistent, **COUNTER** reports are not yet available in a consistent data “container.” The administrative cost of individual provider-by-provider downloads is also high. To make complete use of the reports, library staff must log into each and every one of their different providers, navigate to the right page, then download the appropriate report to their hard drive. Then it is up to the library staff member to consolidate the individual report with all the others and for all the other months.

The **Standardized Usage Statistics Harvesting Initiative (“SUSHI”)** represents a “Web services” approach to solving the **COUNTER** consolidation problem.

SUSHI as COUNTER Companion

In response to the difficulty librarians are experiencing in making use of the **COUNTER** delimited reports, last summer a small group of librarians and vendors assembled and sketched out an automated protocol for moving **COUNTER XML** reports from providers

continued on page 83