DIGITAL LIBRARIES: THE SYSTEMS ANALYSIS PERSPECTIVE

Cataloging for the masses

Robert Fox
University Libraries of Notre Dame, Notre Dame, Indiana, USA

Abstract

Purpose – The purpose of this paper is to explore methods for opening up web content to automated classification using metadata, potentially in the context of library groupware or portals.

Design/methodology/approach – Examines various web sites and meta-searching tools which provides a new means of access for users, and allow users to better document and integrate their research findings.

Findings – This paper is exploratory in nature and highlights trends in the area of library groupware, link routing, and personalized metadata usage.

Practical implications – The vast wealth of information on the web today needs to be exploited by information specialists (librarians) by assisting patrons in organizing, sharing and syndicating content from nearly any information source and empowering patrons via the use of "folksonomies" which are grass roots taxonomies, in conjunction with traditional controlled vocabularies.

Originality/value – In highlighting the as of yet untapped power of technologies such as openURL and link routing, digital librarians can assist patrons by providing services against traditional and non traditional information sources allowing resources to be organized and shared in order to increase utility. This paper examines innovative means by which this could be accomplished.

Keywords Cataloguing, Classification, Worldwide web

Paper type Conceptual paper

The modern world has been dramatically shaped by events that transpired during the late medieval era, and the echoes of undertakings that began during those great ages resonate in our consciousness and our everyday world more than we sometimes acknowledge. The library community lives out the legacies of the medieval era almost transparently when we consider issues of classification, metadata analysis and information syndication. The technological means by which these activities occur are vastly different from the fifteenth-seventeenth centuries; however, the goal and sometimes even the content of our labors remains remarkably similar.

During the seventeenth century, a great mass of texts and manuscripts had amassed in churches and monasteries which contained historical information about Christian saints. This vast wealth of information was not organized in such a way that would allow scholars and hagiographers (those who study the lives of the saints) to perform research. The haphazard state of these documents, and their subsequent reorganization, is an interesting chapter in the history of information science. Almost by pure chance, a Father Heribert Rosweyde was commissioned by the Society of Jesus to begin an endeavor to collect and "catalog" hagiographical manuscripts in the Belgian province "for the glory of the Church and the saints" (Delehaye, 1922). Father Rosweyde's project quickly became overwhelming given the large number of manuscripts available, and many years later, his work was transferred to another

OCLC Systems & Services:
International digital library
perspectives
Vol. 22 No. 3, 2006
pp. 166-172

1065-075X
DOI 10.1108/10650750610686225

Jesuit by the name of Father John Bollandus, who would continue this work for the remainder of his life. He, and those priests who joined him in his work, later became known as the Bollandists, and their task actually continues to this day. The series of texts that is the fruit of their labor, the *Acta Sanctorum*, now numbers over 68 folio volumes and is considered a critical work in hagiographical studies. Their great challenge was to organize material that spanned monasteries, abbeys, ecclesiastical libraries, and possibly hundreds of other manuscript repositories. Manuscript cataloging has not changed drastically in the past 400 years while, on the other hand, the plethora of information sources available to the average researcher boggles the mind. Indeed, it almost seems as though the number of texts and articles available to the average scholar far exceeds the situation the Bollandists faced nearly 350 years ago. And, this is complicated by the fact that the researchers who need these resources are not specialists in information science and information gathering.

Certainly, modern information science owes a debt to the Bollandists, who specialized in descriptive analysis and metadata gathering. If we can imagine the environment that they worked in, and the large number of manual tasks requiring specialized skills and knowledge, it isn't too much of a stretch to see the analogy between the formidable task of intelligently describing and cataloging thousands of manuscripts and organizing the literally millions of resources propagated every day in the digital realm and from a large variety of sources. And yet, this is not only a task of the information specialists, it is also increasingly the task of information users.

Convergence and groupware

While the momentum in the digital library realm is by and large a momentum of proliferation, dissemination and syndication, we are also seeing a movement toward convergence. Web portals and meta-searching tools are an example of this. Web services are now becoming standard extensions in vendor applications and commercial web sites, which provide a new means of access for users, and allow users to better document and integrate their research findings. The task of organizing content and discerning which items contain authoritative and appropriate information, however, is still to a large extent falling on the shoulders of the patron or user.

Users of digital information resources in general, and not simply academic researchers, are realizing the power behind shared peer information. In the commercial sector, we see this being capitalized on by such businesses as Amazon, who several years ago implemented the review system for the merchandise they sell. Other commercial sites such as eBay have utilized peer reviews so that purchasers and sellers can rate one another on the perceived quality of the auction transactions. Apart from the commercial realm, however, we are also seeing an increasing number of web sites that offer free services which help create and maintain interconnections between people, offer the ability to recommend and share music, and also the capacity to rate and share web URLs. In some ways, these services take advantage of the notion proposed by Yahoo! many years ago, and that is to assist web users by organizing the vast selection of web sites into manageable chunks using controlled vocabulary categories. The difference between sites such as Yahoo! and these other sites is that the former allows users to organize the recommendations according to their own terminology and also to personalize their recommendations. A commonly used term regarding these services is "groupware", because it allows people to identify

themselves, and their tastes, in order to better assist others in finding appropriate resources for their needs. In other words, they assist users in "grouping" resources under common terminology in order to connect with other like-minded people and diffuse groups.

A good example of this type of site is Delicious[1], whose implicit motto is "keep, share, discover". This site allows people to organize their favorite URLs, music, books, and other information while also allowing them to share their lists with other users who may be seeking similar items. The primary mechanism that Delicious uses is a tagging system that does not rely upon a strictly controlled vocabulary. The site instructs users to create their own tags instead of "fitting your information into preconceived categories"[2]. They further instruct people that, by creating their own vocabulary, they will "begin building a collaborative repository of related information, driven by personal interests and creative organization." As people add tags to their set of links, the access points for a given URL increase. The more people use the same vocabulary to tag a site, the higher ranked that site is in that category. For example, if the majority of people using Delicious tag a site with the term "philosophy," a user searching in the category "philosophy" or using phrases that contain the word "philosophy" will see this site ranked higher in a result set. Results can also be ordered according to current popularity, depending upon how many users put a particular link into their private collection.

Another site which provides a similar service for images is Flickr, a site that also uses the "tag" approach for classification and organization. Flickr advertises itself as "the best online photo management and sharing application in the world"[3]. With Flickr, users not only have the ability to develop a vocabulary against their own images, but they can also tag other people's images, which helps to increase the exposure of the image via access points, and enhance the related metadata associated with it. Also, in keeping with the groupware focus, Flickr encourages users to create both public and private collections in order to more easily share photos with specific groups of online users such as friends and family.

In the "offline" world, a good example of the type of application we are highlighting is EndNote. EndNote allows scholars to both retrieve and organize citations according to categories that make the most sense to them and in keeping with their research interests. The great thing about EndNote is that it not only allows you to organize, but also share groups of citations in various export formats (including XML) so that others can easily load those lists into their own instance of EndNote. Granted, EndNote is a commercial product, and is not freely available on the web. However, it demonstrates another means by which users can acquire, organize and share resources using mental categories which appeal to like minded users and within communities.

When we talk about the large amount of digital resources that need to be organized and possibly processed by interested parties, as was alluded to earlier in our discussion of the Bollandists, it might help to put this into perspective. Peter Morville, in his recently published book *Ambient Findability*, gives an estimate of the volume of new information appearing on the Internet: "Five exabytes of information. Half a million new libraries the size of the Library of Congress. That's how much new information we create in a year – 92 percent of it stored on magnetic media. It's time we shifted our focus from creating a wealth of information to addressing the ensuing poverty of attention." (Morville, 2005). And, the types of information appearing on the digital

scene is also multiplying at a geometric rate: books, articles, films, audio recordings, images, datasets, maps, etc. As these options continue to increase, the strategies people use to locate appropriate information will necessarily evolve. More measurement has been done in the last decade on information seeking behavior than has probably ever been done previously, and what has been discovered is in fact a very complex picture. The methods people are using to locate information are far more fluid then imagined. The Internet has encouraged a more multifaceted approach to information seeking, corresponding more closely to cognitive models. Again, Morville points out, the web "allows our information seeking to grow more iterative and interactive with each innovation." (Morville, 2005). In other words, for the vast volume of information available to people in various contexts to be useful, both information specialists and information users must become creative. In order to increase that success an order of magnitude, users must be able to share their creativity in the use of personalized taxonomies and ontologies with others in the context of homogeneous groups.

Mis Datos Sus Datos[4]

Obviously, the way in which people gather and use information is rapidly changing. This is true both in the academic community as well as the library patron community at large. As we've discussed, the current problem facing information technologists in the library community is not the volume of data and information sets available to our constituencies. Another problem is that users are finding what they consider useful information from sources that may or may not offer traditional web services (OAI, SRU, etc.). Therefore, using our collective expertise, our focus needs to shift to services that we can provide against our available resources that empower the user to gather, sort, catalog and share collections of resources or metadata that is meaningful to them in a fashion that provides them the greatest utility. How do we do this?

Controlled vocabularies and specialized taxonomies will always be with us. Sometimes the only valid way to organize information is to use an agreed-upon and commonly understood set of terms in order to increase both findability and serendipity. In order to empower patrons to use information in ways that will increase exposure and ultimately usage, though, we must also create services that allow patrons to organize information according to their own cognitive models. Almost two years ago, a collaborative study was done on the use of "groupware" and "personalized link routing", examining how information was handled across communities and offering link routing as a potential solution to information glut. This study recognized the potential crisis faced by information specialists dealing with enormous amounts of digital content, and states: "As decision-making about how to organise information expands from the centre (libraries) to the edge (users and user groups), we need to find ways to make the resources libraries provide fit more easily into a larger and more dynamic information landscape" (Chudnov *et al.*, 2004). One way to accomplish this, in effect, is to enable patrons to catalog, organize and syndicate lists of resources using vocabulary terms that they devise according to specialized needs.

The previously mentioned study focuses on "link routing", which essentially entails personalization of link resolver services, such that depending on who is using a given index, abstract service, etc., the link resolution that connects patrons to available services associated with a given resource or data object will also personalize those services so that they correspond to the patron's status, research interests, or personal

needs. Using one of the examples that the study provides, a professor using an index may use an associated resolver service to add an electronic reference to a list of courseware materials he or she manages or add the reference to their blog site. We could also add the ability to classify, store, and share electronic resources with a larger community. This would certainly involve a paradigm shift in how access is granted to collections based on such services, but such a change in our service model will "help users manage information across their diverse personal collections and information communities" and this "remains true to the core mission of libraries" (Chudnov *et al.*, 2004).

User initiated taxonomies have been labeled "folksonomies" because they originate not from professional information specialists, but because they are "grass roots" vocabularies. Morville states: "Folksonomies flourish in the cornucopia of the commons without noticeable cost. They introduce a wonderful element of serendipity into web navigation, and serve as leading indicators of interest and activity" (Morville, 2005). Folksonomies, though, seem to work best in cooperation with traditional ontologies (such as is found in RDF schemas) and taxonomies (i.e. – controlled vocabularies such as LCSH). Morville continues: "Ontologies, taxonomies, and folksonomies are not mutually exclusive … in some contexts, such as intranets and knowledge networks, a hybrid metadata ecology that combines elements of each maybe the ideal" (Morville, 2005).

The power of these hybrid approaches for information organization has been proven at such online commercial sites as Amazon, eBay, Flickr and non-commercial sites like Technorati, Wikipedia and Epinions. The ability to customize the routing, grouping and sharing of digital resources whether they be citations, images, datasets, or other digital formats combined with the ability to classify (catalog) using *ad hoc* vocabularies has the potential to greatly enhance the relevance of digital library content. The technical means to accomplish this goal may involve several steps or a combination of approaches.

Practical metadata

One common approach to implementing customized lists of electronic resources is to use web portals which allow the user to organize and personalize the digital content they regularly use, as opposed to modifying the user interface or implementing stylistic changes to their web experience. Depending on the dimension of personalization, it may be important for the portal to assist the user to personalize content based upon criteria such as time, location or interest (Schilke *et al.*, 2004). In order to take advantage of the link routing services alluded to earlier, these criteria could be used to tag digital resources so that patrons can more easily collect and inventory those resources for later usage and repurposing in the portal context. Concerning the specific implementation method certain questions may be elicited such as: At what level should link resolution and personalization services be provided? Will one generic customizable software package be used or will multiple vendor based systems be required? What metadata and openURL standards will be used to implement these services? These are not irresolvable questions, and it is entirely possible that a modular approach will be the best one. Indeed, many of these issues are already being discussed by innovative library technology professionals at such institutions as UC Berkeley, Oregon State University and Yale University.

Daniel Chudnov, a librarian who works for the Yale Center for Medical Informatics, has introduced a unique method for implementing link routing and custom classification of resources, found almost anywhere on the web. As presented at the recent code4lib conference held in Corvallis, Oregan[5], Mr. Chudnov described a web based API called unAPI, the purpose of which is to "enable web sites with HTML interfaces to information-rich objects to simultaneously publish richly structured metadata for those objects, or those objects themselves, in a predictable and consistent way for machine processing"[6]. The primary advantage to unAPI is that it affords digital resource users the flexibility of obtaining metadata about resources from almost any source in several available formats and/or metadata standards so that the metadata can be utilized as the user sees fit. In the preliminary specification, unAPI is described as being composed of three parts: a. a URI microformat, b. an HTML based autodiscovery link referring to unAPI services regarding digital objects on individual sites and c. a set of HTTP interface functions. We will take a brief look at each of these components in order to more fully understand the scope of the proposed standard.

A URI microformat, as described on the microformat web site, is "designed for humans first and machines second...microformats are a set of simple, open data formats built upon existing and widely adopted standards"[7]. Essentially, microformats are descriptors for content embedded in XHTML which can extend the semantic content of HTML such that it can be machine extracted. For example, using unAPI, a span can be embedded in a page which describes a digital object:

< span class = "unapi-uri" title = "info:wos/987654321" />

This indicates that a particular Web of Science citation is an unAPI accessible digital object.

Secondly, each object identified as unAPI accessible needs to have an autodiscovery link associated with it which represents a link where unAPI requests can be directed:

< link rel = "meta" type = "application/xml" title = "unAPI" href = *http://myspace.com/unapi/* />

The HTTP interface works very similarly to OAI-PMH or SRU/SRW. The command set is different, but Chudnov has proposed three types of commands corresponding to a request for a list of available metadata formats, a URI command which will return available metadata formats for a particular digital object with the associated links for each metadata object (using the HTTP 300 status code – Multiple Formats), and finally a command which corresponds to the retrieval of a particular digital object identified by a URI and an available metadata format for retrieval. Metadata formats may include OAI-PMH Dublin Core, MODS, METS, HTML, tab delimited, JPEG, GIF...the formats are really only limited to the expectations of the user population.

Again, the advantage to the user is flexibility, increased utility, and multipurposing of data. Chudnov has suggested that application layers could be added which would provide a service layer on top of any existing framework. Using an link routing layer such as unAPI within an existing application API, functionality could also be added which assists the user in categorizing the incoming metadata and possibly publishing that information into multiple contexts such as courseware systems, portals, etc. as was outlined earlier. The metadata/data formats can also include options for granularity which will fit the needs of various patron populations, and allow them to

quickly and comprehensively catalog their digital resources. As stated earlier, this may include a combination of *ad hoc* "tagging" which can then be fed back into a larger context, systematically controlled vocabulary, or a combination thereof. Other services added to the ingest layer may include a rating system, record overlay, record conversion and other value added features. For internal usage, a method like this may also assist digital librarians in enhancing application data exchange and interoperability.

Offering services such as these in a transparent way to digital library patrons can bring new relevance within the information science field, by helping the patrons to more easily integrate the vast wealth of digital content into contexts that are meaningful to them. In providing a framework in which grass root taxonomies combine with the time tested wisdom of traditional vocabularies, along with a means to easily retrieve, store, and use information from almost any data source via link routing and established metadata standards, we can greatly assist patrons in overcoming information glut while adhering to a traditional mandate of librarianship in the digital world.

Notes

1. See Delicious at http://del.icio.us
2. Instructions taken from the Delicious web site: http://del.icio.us/help/tags
3. See Flickr at www.flickr.com
4. Play of words on Spanish saying "*mi casa, su casa*" which means "my house, your house". *Datos* in English means "data".
5. For information on the content of the cod4lib conference and the associated listserv, please see www.code4lib.org/.
6. See www.code4lib.org/files/unapi_revision_1-14.html for information regarding unAPI.
7. See http://microformats.org/about/ for more information on microformats.

References

Chudnov, D., Frumpkin, J., Weintraub, J., Wilcox, M. and Yee, R. (2004), "Towards library groupware with personalized link routing", *Ariadne*, Issue 40, July, available at: www.ariadne.ac.uk/issue40/chudnov/

Delehaye, H.S.J. (1922), *The Work of the Bollandists Through Three Centuries 1615-1915*, Princeton University Press, Princeton, NJ, p. 8.

Morville, P. (2005), *Ambient Findability*, O'Reilly Media Inc., Sebastopol, CA, pp. 44-45, 60, 138, 139.

Schilke, S.W., Bleimann, U., Furnell, S.M. and Phippen, A.D. (2004), "Multi-dimensional-personalisation for location and interest-based recommendation", *Interest Research*, Vol. 14 No. 5, pp. 379-80.