ALA American Library Association

IMPROVING WEB VISIBILITY

INTO THE HANDS OF READERS

Ted Fons

Library Technology Reports
Expert Guides to Library Systems and Services

JULY 2016
Vol. 52 / No. 5
ISSN 0024-2586

Library Technology
REPORTS
Expert Guides to Library Systems and Services

Improving Web Visibility: Into the Hands of Readers

Ted Fons

ALA TechSource
alatechsource.org

American Library Association

Library Technology
REPORTS

ALA TechSource purchases fund advocacy, awareness, and accreditation programs for library professionals worldwide.

Volume 52, Number 5

Improving Web Visibility: Into the Hands of Readers

ISBN: 978-0-8389-5974-9

American Library Association

50 East Huron St.
Chicago, IL 60611-2795 USA
alatechsource.org
800-545-2433, ext. 4299
312-944-6780
312-280-5275 (fax)

Advertising Representative

Patrick Hogan
phogan@ala.org
312-280-3240

Editor

Patrick Hogan
phogan@ala.org
312-280-3240

Copy Editor

Judith Lauber

Production

Tim Clifford and Alison Elms

Cover Design

Alejandra Diaz

Library Technology Reports (ISSN 0024-2586) is published eight times a year (January, March, April, June, July, September, October, and December) by American Library Association, 50 E. Huron St., Chicago, IL 60611. It is managed by ALA TechSource, a unit of the publishing department of ALA. Periodical postage paid at Chicago, Illinois, and at additional mailing offices. POSTMASTER: Send address changes to Library Technology Reports, 50 E. Huron St., Chicago, IL 60611.

ALA TechSource
alatechsource.org

About the Author

Ted Fons is a Principal Consultant with Third Chapter Partners, a technology and marketing consulting firm for libraries and their commercial partners. Previously, he was Executive Director of Data Services at OCLC, where he provided vision and direction for OCLC's global metadata network, including WorldCat. He participated in OCLC Research's experiments with linked data and participated directly in the BIBFRAME Early Experimenters project. Earlier he held management positions at Innovative Interfaces and also worked in academic libraries. He holds an MLS from Syracuse University and a BA in international affairs from Marquette University.

Abstract

Improving the visibility of library collections and services on the open web is one strategy in enhancing the long-term viability of libraries. The tradition of modern librarianship has been to focus on the efficiency of library workflow systems and technical processing and the accuracy of metadata against librarian-authored rules for metadata encoding. This issue of *Library Technology Reports* (vol. 52, no. 5), "Improving Web Visibility: Into the Hands of Readers," by Ted Fons, discusses actions for libraries to take with regard to content exposure systems, vocabularies, content metadata regimes, and system design approaches that will serve the convenience of the web searcher and thereby contribute to the long-term viability of libraries.

Contents

Into the Hands of Readers

Acknowledgments

Enormous thanks to the many people who gave me input, guidance, and correction on this work.

To Rob Sieracki of Ox Optimal and Semantic Web consultant Richard Wallis for help on understanding the business and technology of search on the web. To the library at The Ohio State University for its generosity in allowing an unaffiliated researcher to use its excellent collections—its library science collection was invaluable to the section on the history of library catalogs. To Patrick Sweeney of EveryLibrary for connecting me to librarians on the front lines of public library service. To Rachel Fewell and Frank Wilmot at the Denver Public Library and Erica Findlay and Emily Papagni of the Multnomah County Public Library for their insights into the joys and pains of supporting real people coming to the library from the open web. To Jeff Penka, Eric Miller of Zepheira, and Steve Potash of OverDrive for insights on content marketing and the business of web visibility for libraries. To Philip Schreur (Stanford University), Beacher Wiggins (Library of Congress), John Chapman (OCLC), and Carl Stahmer (UC Davis) for their time helping me understand their contributions to remaking library services on the web. To the metadata expert Diane Hillmann for challenging my understanding of cataloging in the context of library Semantic Web data and for giving her time generously to improving my representation of RDA.

Finally, I'd like to thank my colleagues who read early versions of this work and provided invaluable suggestions for improving its quality: Gene Shimshock (Third Chapter Partners), Merrilee Proffitt, Roy Tennant, and Jeff Mixter (OCLC Research).

Introduction

"We are in an in-between world where we have two groups of people: those ones who already go to the library and the ones who never think about the library."[1] That's how Rachel Fewell, the collection services manager at the Denver Public Library, describes her view of the landscape for libraries. This view of the world prompts these questions: What can libraries do to reach those who never think about the library? What can libraries do to most effectively reach those who sometimes think about the library? Increasing the visibility of library collections on the web is an obvious answer, but the explicit goal to make that happen has not been well defined.

A review of the history of library catalogs and library systems automation reveals a trend toward a focus on efficiency and cost savings in systems and data. There are a few bright spots of innovation in discovery, and the very earliest days of library catalogs were highly focused on the user, but the trend has been on service to ourselves instead of the convenience of or improved outcomes for the user. We see this mostly by contrast: the commercial search engines have completely disrupted the user experience of discovering information on any topic while libraries have by and large focused on internal system efficiency and high-quality metadata for print books.

There is a tremendous opportunity for libraries to connect readers to content on the web. Frank Wilmot, senior reference librarian, also at the Denver Public Library, tells the story of the library users who call their reference line asking for "that product rating chart with the black circles and red dots."[2]

After a reference interview, the answer to the question frequently turns out to be the *Consumer Reports* ratings on appliances and other consumer goods. The library often gets calls from readers who are buying a new appliance and want to see the *Consumer Reports* ratings before they buy. Wilmot explains that these readers get stopped by the *Consumer Reports* subscription requirement and immediately call the library for help. He reports that even the smallest branches have the subscription through an aggregator. On the phone, library staff can quickly connect the reader to the subscription resources, and he celebrates the successes but worries about the missed opportunities: the people who never get the benefit because they don't know about the service. The access model is simple: anybody with a library card can access the database, but that connection between the reference to the content and the full content is not made on the open web.

In the academic library environment, Roger Schoenfeld, Director, Library and Scholarly Communication Program at Ithaka S+R, summarizes the situation in his measured but direct style: "The user experience of working with e-journals and ebooks in an academic setting has failed to keep up with changing practices and preferences for how researchers now expect to access the scholarly literature."[3]

He doesn't directly say that the academic user experience should be more like the experience of the web, but it is implied that the search engines have changed the model even for researchers, and libraries have failed to keep up.

The answer to all of this is to renew the focus on the convenience and search preferences of the reader. If we remind ourselves of Ranganathan's 1931 laws, the reader is mentioned in most of them:

1. Books are for use.
2. Every *reader* his / her book.
3. Every book its *reader.*
4. Save the time of the *reader.*
5. The library is a growing organism. [emphasis added][4]

Note rule four—emphasizing the convenience of the reader. In their research on the enduring value of Ranganathan's rubric, Lynn Silipigni Connaway and Ixchel Faniel from OCLC Research argue that in today's networked information environment, where the user has many choices for information, the fourth law should be the first and that "time [is] a shorthand for convenience or almost any efficiency-based value that users ascribe to their experience with a library."[5]

Eighty-five years after Ranganathan, libraries should set clear goals about the convenience of the reader and focus on the satisfaction of the reader in the discovery process. Delivering content into the hands of the reader should be an explicit goal motivating behavior and guiding decisions. Given the reader's preference for the highly relevant and instantly informative experience on the web, it will be important to understand the rules of the web and to very explicitly change a number of aspects of culture, process, and data management. This is an important goal and the stakes are high. It is time to ask the important questions about how libraries and their partners will make this happen.

The Question: Can Libraries Improve Their Web Visibility?

Increasingly, librarians are asking the existential question: Can libraries thrive if their services aren't prominent in web search engine results? If ordinary people don't see their library's books and articles in search results, will library users disregard the library as a place to satisfy their research and leisure needs? Libraries build their collections for their readers, but if readers never find them and get them from the library, will they stop seeing the library as a place of value that should be cherished and supported?

The knowledgeable observer will take the question one step further and say that the modern library no longer features just "the collection" of books as its premier offering; the modern library offers an enormous variety of other services. What about exposing those services on the web? OCLC's 2005 *Perceptions* report helped us see clearly that the library brand continues to be "the book."[6] But libraries invest enormously in the curation and infrastructure for everything else they offer.

What is the everything else? Academic libraries offer lectures, multimedia collaboration space, exhibits, bibliographies of their scholars, digitization services, and other assistance to scholarly communication. Public libraries offer author readings and services for adults like job search instruction; for children and young adults, they offer story time, maker spaces, homework help; they connect teens to materials on sensitive topics and provide a private space for using library resources to get answers to the most difficult questions young people can ask. The list of services is long, and the need to promote those services is urgent in an environment where ordinary people are surrounded by many options for meeting their needs for serious research and leisure reading.

The pride that librarians hold in these services is manifest in any conversation with librarians today. Philip Schreur, the Associate University Librarian for Technical and Access Services at Stanford University, describes the university's explicit mandate for the library to represent the entirety of the university's scholarly output and assets. He explains: "We have a mandate to integrate all of the information that the university creates. That includes reading lists, research data sets, anonymized transactions, historical image collections, and many others; the library collections are the smallest of those assets."[7]

The intellectual and financial investment in these non-book services increases with each new budget. But these services are no more favored in search engine results than books and articles. In a search engine, any search for a best seller or a work of local historical interest will not produce a link to a library on the first page of the search results. So the problem exists with either the traditional view of the library as the source for books or the expanded view that includes the wide variety of services that libraries provide: the library's offerings are generally not prominent in search engine results. So the question remains: Can the library thrive when it is buried by sponsored results or direct links to commercial options? To understand this context, it is useful first to understand how any content is exposed on the web.

Notes

1. Rachel Fewell (Collection Services Manager, Denver Public Library), interviewed by Ted Fons by telephone, October 28, 2015.
2. Frank Wilmot (Senior Reference Librarian, Denver Public Library), interviewed by Ted Fons by telephone, November 10, 2015.
3. Roger Schonfeld, "Dismantling the Stumbling Blocks That Impede Researcher Access to E-resources," *The Scholarly Kitchen* (blog), Society for Scholarly Publishing, November 13, 2015, http://scholarlykitchen.sspnet.org/2015/11/13/dismantling-the-stumbling-blocks-that-impede-researcher-access-to-e-resources.
4. "Five Laws of Library Science," *Wikipedia*, last modified 5 February, 2016.
5. Lynn Silipigni Connaway and Ixchel M. Faniel, *Reordering Ranganathan: Shifting User Behaviors, Shifting Priorities* (Dublin, OH: OCLC Research, 2014), www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-reordering-ranganathan-2014.pdf.
6. OCLC, *Perceptions of Libraries and Information Resources* (Dublin, OH: OCLC, 2005), https://www.oclc.org/reports/2005perceptions.en.html.
7. Phillip Schreur (Associate University Librarian for Technical and Access Services, Stanford University), interviewed by Ted Fons by telephone, November 6, 2015.

Exposing Content on the Web

To understand how anything is exposed on the web, it's useful to understand how Google, the most widely used of the search engines, indexes and ranks the content that it gathers.

Google wasn't the first company on the web to provide search results across Internet content, but soon after its late 1990s debut, its search product became almost exclusively associated with searching and finding exactly the thing that the searcher was looking for. It beat competitors with colorful names like AltaVista, Yahoo!, and HotBot by providing the search tool that generally found the right thing, ranking that right thing at the top of the list of results, and doing it all in less than a second. The market quickly perceived that Google Search did that better than anybody, and Google has generally retained that position today. It has even influenced spoken languages as its success inspired the new verb *to google* as a standard way of indicating "to search for answers on the web."

Google Search Methodology

Google's science of crawling, indexing, and ranking webpages is well understood insofar as Google explains the mechanics to specialists and the general public. What follows below is a high-level view of how those mechanics work based on the information that Google makes public—it is necessarily simplified to provide the basics and to illustrate that how to make content visible on the web is a known art and science.

To encourage the creation of crawlable and indexable websites, Google provides good information and even technical tools so individuals and companies can design their webpages for the best possible results. It is in Google's interest to encourage good behavior in website design so it can maximize the quality of its search results. Google also wants to minimize the amount of work it has to do to prevent bad behavior. The behavior it wants to discourage is where web designers try to game the PageRank system to advantage their own content. Google discourages that behavior through sophisticated algorithms and punitive removal of content from search results. Those removed from results have to petition Google to have their content displayed again, and Google's engineers have to be convinced that the behavior was not intentional and corrected before they are cleared.

In addition to crawling the open web, Google will seek out partners and create formal partner contracts when it wants particular content. It will insist on its technical specifications and open web protocols, but it will go beyond finding content on the open web. It will, in a sense, curate the content of its own indexes when that matches its strategic goals. Libraries tend toward cooperative arrangements with their data and resources, so this is a potential source of opportunity and a channel for libraries to expose their data.

However, the most effective way to gain visibility for any content is by following open web practices and making publicly available webpages that match Google's published best practices. The best information about how Google evaluates webpages is provided for traditional search results. Traditional search results are the list of websites and documents that appears in the middle of a traditional browser or mobile search application page. The space reserved for the traditional search results is one of three zones that make up the search results page on a full browser: sponsored links, traditional search results (central zone), and the Knowledge Card. Mobile results are different but share many of the same characteristics.

On a mobile device browser, the zones are different, but the principles for how content gets there are the same. There are also rules for the display of content in the other important parts of the Google search results page: the sponsored links and Knowledge Card. But first, it is important to understand how ranking of the traditional search results works.

The mechanics of managing results in the central zone begins with crawling and indexing and ends with page ranking. Crawling and indexing populate Google's indexes so it has words to search and links to display in results. PageRank determines how often a page has been linked to—this measure of a page's popularity is a measure of its usefulness. If lots of other webpages link to this page, it must be considered useful—perhaps even authoritative. Once Google has content in its indexes, it can compare searches to those indexes and determine what to rank in the results. To do this, the search engine compares the search to the indexes and asks somewhere around 200 technical "questions" of the page content in the indexes—these are interrogations of the indexes to determine which pages have the best results and how they should be ranked. The rules are constantly changing, and the full details of the rules are Google's most important trade secret, but Google tells all website designers that at least eight of the questions are central to the process and that they should take care to observe best practices in relation to these questions:

1. Is the page blocked? Has the webmaster put a block on the page so it can be accessed only through a browser directly and not by a crawler?
2. Does the page include videos and pictures? Multimedia content is considered good—this means it is a page that people are likely to want to stay on.
3. What is the word frequency on the page relative to the search? For example: How many times do the words *Noah's Ark* appear on the pages in the index? High word frequency on the page is a hint at relevance to the search.
4. Are the words in the search in the title section of the page? This is a technical detail of HTML writing—the TITLE section declares what a page is about. That's another hint at relevance.
5. Are the search words in the URL of pages? This is another hint. If the keywords *Noah's Ark* are right there in the URL, perhaps this entire website is about Noah's Ark; that's a hint at relevance to the search.
6. What about adjacency? Do repeats of the keywords appear close to each other? Another hint.
7. Are there synonyms for the words on the page? This hints at a deeper understanding of the topic and overall content quality.
8. What is the overall quality of the website? Here Google uses the term *spammy*. It wants to link to sites that offer what the user is looking for—whether that is buying or learning, it doesn't want to link to lists of other pages with no added value. In this area, Google might also look at how frequently a page is updated. But Google's staff warn content producers not to focus too much on this. Focus on overall quality of the content, and you will attract links and therefore value. However, a page that is updated infrequently—let's assume months or years between updates—will appear to be stale and of lower quality.[1]

Google uses the answers to these questions, and many more, to determine how useful a page is relative to the search. If the answer to many of the questions is yes, then the yes pages must be relevant. The strength of the yes (how closely the pages and the keywords match through the filter of the questions) is a measure of their relevance. But Google also uses its innovation in search: the PageRank. PageRank measures how many other sites are already linking to a page: that's the final hint at relevance and usefulness. Google's founders introduced this concept as a distinguishing feature of its search product. They developed the algorithm to create a measure of value of a page through the proxy of how often the page is linked to. In other words, how popular a page is—how many times other sites refer to it—is a measure of its usefulness. This is a key element of ranking and also has relevance in the second key element of the geography of the results page: the Knowledge Card.

Google's Knowledge Card

Because advertising is Google's chief source of revenue—users of the site don't pay money to use it, they pay with their time and exposure to advertisements placed on the results page—the company has turned to providing more and more content directly on the results pages. It is not providing just links to pages that might answer a question like, "What time does the Cincinnati Bengals game start today?" or "Where is the new Star Wars movie playing?" or "Star Wars show times?" It is providing the answers to those questions directly on the search results page. For some answers, it isn't necessary to click to the page or document—the answer is given directly in the Knowledge Card next to the traditional results. The usefulness of this Knowledge Card and its apparent durability (it has been on Google results pages for several years as of early 2016) indicates that it is worth understanding how to get content into this zone.

The Knowledge Card, sometimes called the Info-Card or Answer Panel, is the second of three important areas on the Google results pages. Of course, there are rules for how content is selected for display there as well.

There is some debate about how website managers can influence the visibility of their resources within the Knowledge Card. Richard Wallis, Semantic Web expert, describes it this way: "To get your content into the answer panel, recognizable semantic properties will prove more fruitful and effective than simple words."[2] There is a lot in that statement, and it is useful to understand more about linked data and the Semantic Web before the full value of the statement is revealed.

First a review: Wallis is saying that following the rules of the Semantic Web improves your chances of getting your content into the Answer Panel. When reviewing the rules for relevance in the central search results zone, it was clear that page rank and then words—their placement, their markup, their frequency, and so on—were key to relevance and utility. In the case of the dynamically generated Answer Panel, a different set of rules is more important. The general guidelines for following the rules of the Semantic Web are

1. quality and breadth of the internal graph
2. quality of connections to the global graph
3. recognizable markup[3]

Quality of the internal graph relates to a Semantic Web principle that states that content should be described in terms of linked references to the things you are describing. That means any reference to a person, a place, an object, or an event should include a universal reference to that thing. This includes reference to the holding organization. This is typically a Uniform Resource Identifier (URI). As with traditional library authority control, a URI provides an unambiguous and repeatable way to represent something. A URI provides both the "address" of a data item and a description of the thing it identifies. And as it is used repeatedly, systems can develop trust that the URI is reliable—it points to a site with authority and trust in describing something. If the search engine can find the same identifier on multiple pages, then it can more efficiently determine that the page is about the same thing; this is useful when matching searches to pages; an unambiguous identifier is always better than trying to match text—it produces a more confident match. So quality of the internal graph is measured in how frequently the things on a webpage are described using links to authoritative sources instead of just text—even if that text is consistent. Using links provides a better score in conforming to the rules of the Semantic Web.

Quality of connections to the global graph is enhanced by the use of global identifiers for things. An organization or even group of organizations can invent their own identifiers for things, but using existing identifiers that are already used on the web (the global web) is the approach that the search engines reward. This is a familiar concept for librarians who have created a number of widely recognized schemes for consistent description: the Dewey Decimal System for a single term describing what a published thing is about, and the many national name authority files: the Library of Congress Name Authority File in the United States, the Integrated Authority File (GND) among the German-speaking countries, and the various name authority files from the French National Library (BnF) are all efforts by local communities to describe things in a consistent way.[4] The value of consistency was always a reduction in cost in cataloging and some benefit to the reader in consistency in indexing. Somebody has already done the hard work of determining how to spell an author's name, for example, or the town he was born in, or the degree she earned at a particular university. The benefit of consistency in indexing is manifest when the user has a better chance of finding something if there are cross-references to various forms of an author's name. Furthermore, the display of results is cleaner when the persons contributing to the work are recorded consistently.

The same principle applies on the Semantic Web, but the specific incentive is to use globally recognized identifiers when they exist. And since Semantic Web description is meant for machines and not humans, it is common to use multiple identifiers for a thing. In fact, multiple identifiers can be an advantage. As with synonyms in traditional relevance ranking, it shows that a website has a deep understanding of a thing. So collecting and using multiple identifiers for a thing strengthens the breadth and quality of the internal graph and indeed the global graph when global identifiers are used. The emphasis on connections to the global graph implies that there is value in multiple sites referring to things in the same way. Semantic Web experts would say this strengthens the nodes on the graph, but the plain language way to say it is to compare it to a chorus: the more voices singing the same words in the same key and at the same volume, the stronger the impact on the audience.

Wallis's third element is recognizable markup. *Markup* in Semantic Web jargon refers to adherence to the recommended vocabulary schemes to draw Semantic Web concepts into all websites. The global search engines Google, Yahoo, Yandex, and Bing agree that consistency in markup and Semantic Web principles make their crawling and indexing work easier. They are fierce competitors, but they have found a common interest in how content should be presented. They all recommend adherence to markup specified on their website schema.org. It encourages isolating the persons, organizations, objects, places, events, and other things being described and, where possible, identifying them with global identifiers. Librarians have been very active in influencing schema.org, so the vocabulary better represents bibliographic items and the libraries that hold or offer them. By encouraging use of a de facto common vocabulary across well over 10 million sites, schema.org has introduced a broad consistency to the Semantic Web that has previously been lacking.

The dramatic increase in use of schema.org on the web hints that website developers believe it will help with indexing. It is also important to know that the Google Knowledge Card appears to draw from a set of reliable and durable sources that influence Google's own internal knowledge graph. There are many sources for Google Knowledge Card data, but DBpedia is frequently mentioned in this context. DBpedia derives its data in part from *Wikipedia*, and the direct management of the quality of that data is important for success in appearing in the Knowledge Card. Kenning Arlitsch, the Dean of Libraries at Montana State University, who has experimented deeply with managing the visibility of the library and its specialized collections, explains that DBpedia "tends to be the primary source from where Google gets is information for the Knowledge Card." Arlitsch says bluntly: "If you don't have an article in Wikipedia to draw into DBpedia, then you don't exist to Google."[5] The knowledge graph that libraries can influence directly is therefore an important part of the Semantic Web infrastructure and can't be ignored in the question of library visibility on the web.

Google's AdWords

The third zone on Google results pages is the sponsored links or AdWords. In this zone it is the business relationship with Google that determines placement. To explain how these results are displayed, Google says in plain language: "Google may be compensated by some of these providers."[6] Presumably Google's business development teams negotiate contracts with these providers and use data they provide to display results matching searches. It is reasonable to assume that all of the rules for the traditional results and the Answer Panel or Knowledge Card zones are used, but the additional factor of payment for placement is the final element that determines what is displayed in the sponsored links zone.

Libraries have an opportunity in that the rules of the web are well understood and there is an art and science around optimizing content for search engine uptake that is now more than a decade old. The challenge for libraries will be to apply those rules and change the many decades of practice in catalog and data management. With this understanding of how things are exposed on the web above, it is useful to review what users want from libraries.

Notes

1. Google, "How Search Works: Crawling & Indexing," accessed February 11, 2016, https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html.
2. Richard Wallis (Independent Structured Web Data Consultant), interviewed by Ted Fons by Skype, 23 October, 2015.
3. Ibid.
4. Ioannis Papadakis, Konstantinos Kyprianos, and Michalis Stefanidakis, "Linked Data URIs and Libraries: The Story So Far," *D-Lib Magazine* 21, no. 5/6 (May/June 2015), http://dx.doi.org/10.1045/may2015-papadakis.
5. Kenning Arlitsch and Patrick O'Brien, "Establishing Semantic Identity for Accurate Representation on the Web" (presentation, Coalition for Networked Information Fall 2014 Membership Meeting, Washington, DC, December 8–9, 2014).
6. Danny Sullivan, "Once Deemed Evil, Google Now Embraces 'Paid Inclusion,'" *Marketing Land*, May 30, 2012, http://marketingland.com/once-deemed-evil-google-now-embraces-paid-inclusion-13138.

Discovery and Fulfillment

When Steve Potash, the CEO of OverDrive, the e-book and audiobook provider to libraries, talks about making library content more visible on the web, he talks about "content marketing."[1] By that he means that libraries should understand that they have what readers want and they should market it in the most attractive and effective ways possible. For libraries to promote their content, it is useful first to understand what readers value in their offering.

What Readers Want from Libraries

Academic libraries have traditionally described what they offer to their associations and accreditation agencies in terms of counts of books, journals, and more recently, networked resources such as e-books and databases. That's a perspective that describes what the library has, but it doesn't tell us what users want or, more boldly, what content should be most visible on the web. Looking at it from the consumer perspective, it would be interesting to know what library offerings students actually use. To answer this, we can get some hints from a study called "Library Use and Undergraduate Student Outcomes."[2] That study, from 2013, used as one of its inputs the services that undergraduates actually used. This analysis was in the context of the effort to understand the relationship between library usage and academic achievement—a topic of interest to academic librarians. If we take this input as a measure of what academic library readers value, the rankings appear in table 3.1.

According to this analysis, the items of highest interest to undergraduate students are articles, and specifically articles in electronic form. The two most used library services in this study were databases that contain individual articles and articles found directly in electronic journals paid for by the library. There is also significant interest in using library computers, borrowing books, and information on the library website. The detailed findings show that interest drops off significantly after book loans.

Table 3.1. Ranking of usage of thirteen library access points by first-time, first-year undergraduate students at the University of Minnesota during the Fall 2011 semester

Service or Offering	Rank in Number of Uses
Databases of individual articles	1
Electronic journals directly	2
Workstations: PCs & laptops	3
Book loans	4
Library website	5
Bibliographic instruction course, pt. 1	6
E-books	7
Course-integrated instruction	8
Bibliographic instruction course, pt. 2	9
Reference questions	10
Workshop in library	11
Book loan from other library (ILL)	12
Peer conference	13

Source: Krista Soria, Jan Fransen, and Shane Nackerud, "Library Use and Undergraduate Student Outcomes: New Evidence for Students' Retention and Academic Success," *portal: Libraries and the Academy* 13, no. 2 (April 2013): 147–64, http://dx.doi.org/10.1353/pla.2013.0010.

From these numbers we can generalize that students who use academic libraries are primarily interested in online articles, then in using the library's computing facilities, then in borrowing books. There is interest in other library services such as bibliographic instruction and reference questions, but those are of secondary interest to library users. Recalling Steve

Definitions

Discovery: the process of finding things on a topic
Fulfillment: the process of acquiring things that have been discovered

Potash’s emphasis on content marketing when considering web visibility, it is useful to know what this important user group values in the library’s content.

There is similarly useful information about public libraries that provides answers to the question, “What do users want from public libraries?”As with academic libraries, the tradition is for public libraries to describe what they have and what they can count—these are the trailing indicators of public library offerings: counts of books, checkouts, and gate counts. But a more recent trend among public librarians is to focus on measuring value and impact of their services. In the context of understanding demand and predicting web search behavior, we can look at recent surveys to gauge value in public library services.

A Pew Internet study published in 2013 surveyed people who had used public libraries and asked them to rank what offerings were important to them.[3] The percentages in table 3.2 describe the offerings that respondents rated with the highest rating: *Very Important*.

The mix of collection and human services is far greater for public libraries than it is for academic libraries. Books and media still draw the largest proportion of interest, but the wide variety of programs and personal services are of enormous importance and a significant component of why people use a public library. Understanding the high value of books and media could be a useful guide in making decisions about which content to make more visible on the web. Similarly, public libraries have an opportunity to broadcast the expertise of their public services staff and the useful role of “library as a quiet, safe place” for their communities. All of this is useful input in considering which content to market on the web.

Table 3.2. Percentage of people ages 16+ who said that these services were “Very Important” to them and their families

Service or Offering	Rated Very Important
Books & media	54%
Librarian assistance	44%
Having a quiet, safe place	51%
Research resources	47%
Programs for youth	45%
Internet access, computers, printers	33%
Programs for adults	28%
Help applying for government services	29%
Help finding a job	30%

Source: Kathryn Zickuhr, Lee Rainie, Kristen Purcell, and Maeve Duggan, *How Americans Value Public Libraries in Their Communities* (Washington, DC: Pew Research Center), December 11, 2013, 2, http://libraries.pewinternet.org/2013/12/11/libraries-in-communities.

How People Discover What Is in the Library

At this point librarians are exhausted by being told that the library catalog is the last place users look to discover things. Countless studies in the past ten years have told them that when people begin their search on a topic, they start with a search engine. OCLC’s 2010 report on public perceptions of libraries captured the essence of all of the studies: “[By 2005] the majority of online information consumers (82%) began their searches for information on a search engine, a source they found roughly as trustworthy as a library. One percent (1%) began their searches on a library web site.”[4]

Variations on this finding have been reported over the years and all with a consistent theme: when people want to know more about a topic they start on the web.

However, those statistics speak only to the discovery process: the process that readers and researchers use to find things that match their topic. Whether they are looking for scholarly articles or topics of personal interest, search engines are a clear leader among choices for discovery. However, how people learn about things that are useful to their needs has many dimensions. Merrilee Proffitt and her colleagues in OCLC Research describe it simply: “Users increasingly have choices outside the library, and those choices are both networked and social.”[5]

Often readers discover things before they need them through media and peer networks. Scholars are inclined to share their new publications with colleagues in their disciplines. Advertising has a role as well; publishers have very sophisticated methods of pushing notices of their new titles and the tables of contents of newly published journal issues to scholars. Even Amazon has a role with its Alert Me service to tell readers when a new title is available for purchase. Amazon’s recommender services use its vast store of transaction history to recommend related titles.

But during the process of discovery where a user has a topic in mind, the user will start in a search engine or a specialized database for a particular scholarly discipline. The specialized databases range from those with a hundred years of history behind them like the Chemical Abstracts database SciFinder to ultra-specialized resources like Current Protocols in Nucleic Acid Chemistry. Many readers will develop a familiarity with databases of articles on business or cultural topics and prefer them as a starting point depending on their need.

Discovery of things that match a reader's needs is multifaceted and individualized, but the generalization remains true: in most cases it does not begin with the library catalog.

How People Get Things from the Library

That brings us to the concept of fulfillment. Fulfillment is the process of acquiring the thing that matches the reader's need. Given what we know about academic libraries, the challenge for readers there is to get the electronic article they have discovered. For public library readers it is getting a book or media that they have discovered elsewhere. For articles, the question is simply does the library have subscription access to this article? For books and media, the question is does the library have this item on the shelf?

So where do readers turn to determine availability? It depends on where they start. For articles, availability is determined by the discovery system's knowledge of the library's subscriptions. On the general web, in Google Search for example, the system will have no knowledge of the library's subscriptions, so the reader will either turn to a library system, use a pay-per-view option, or give up and find another source. Scholars with a well-defined peer network might go directly to the article's author to acquire a prepublication or published version of the article.

There is a lesser-used, but still important variation on Google Search called Google Scholar—it contains citations and an option for users to declare their institutional affiliation. When the user is starting from a system that contains only citations such as Google Scholar or something hand-crafted like a list of articles required for a course, then there are specialized tools that the library can put in place to check the library's subscriptions and provide the answer to the question upon clicking a button. When the user is in a database hosted by an aggregator or publisher and the library subscribes to the title, then the link to the full text of the article is provided immediately.

For finding the availability of books and media, the local library catalog is the most reliable system for accurate statements of availability in all library types. Many libraries use discovery systems that combine their local catalog content and selected article content, but even these systems refer to the local system in real time to determine the number of copies and disposition of the item—to really know if an item is available for lending, the local system is the "system of record."

Librarians have studied the logs of their local catalogs for many years to determine how well their search menus are configured. They have also used those logs to determine which indexes are used, how often searchers find something that matches their search, which indexes are most popular for searching, and even if there are gaps in their collections. The general trend of those studies is that known item searching is the most popular kind of search in local catalogs. Searchers tend to have a title or author in mind, and they will search the catalog to determine what the library has. This supports the generalization that people discover things outside of the catalog in many ways and refer to the catalog for fulfillment—to determine if they can acquire the thing they need. They may ask the questions, "Can I get this thing from the library? Does it have a copy available?" For articles, they use the system they are in to determine availability. If that fails them, they will use other systems or give up and find another resource that matches their need.

The gulf between discovery and fulfillment illustrates the fractured nature of the visibility of library collections today. The gulf introduces risk—risk that the reader will not be aware of the full range of fulfillment options provided by the library in local lending and engaging with the global lending networks that have been successful for decades. Clearly understanding that risk adds to the stakes in the question, "Can libraries improve their visibility on the web?"

It is clear that for books and media, the library catalog is a core asset in declaring what a library has. Given that, it is worth some investigation of the evolution of library catalogs and their historical role in telling the world what a library offers.

Notes

1. Steve Potash (Chief Executive Officer, OverDrive, Inc.), interviewed by Ted Fons by telephone, 16 November, 2015.
2. Krista Soria, Jan Fransen, and Shane Nackerud, "Library Use and Undergraduate Student Outcomes: New Evidence for Students' Retention and Academic Success," *portal: Libraries and the Academy* 13, no. 2 (April 2013): 147–64, http://dx.doi.org/10.1353/pla.2013.0010.
3. Kathryn Zickuhr, Lee Rainie, Kristen Purcell, and Maeve Duggan, *How Americans Value Public Libraries in Their Communities* (Washington, DC: Pew Research Center, December 11, 2013), http://libraries.pewinternet.org/2013/12/11/libraries-in-communities.
4. OCLC, *Perceptions of Libraries, 2010: Context and Community* (Dublin, OH: OCLC, 2011), 4, www.oclc.org/reports/2010perceptions.en.html.
5. Merrilee Proffitt, James Michalko, and Melissa Renspie, *Shaping the Library to the Life of the User: Adapting, Empowering, Partnering, Engaging* (Dublin, OH: OCLC, 2015), 5, www.oclc.org/research/publications/2015/oclcresearch-shaping-library-to-life-of-user-2015.html.

The Tradition of Library Catalogs

What follows is a review of the evolution of catalog librarianship and library catalogs. This review reveals that the tradition of library catalogs has drifted from a clear emphasis on the convenience of the reader to an emphasis on the efficiency of the systems that create library catalogs.

Starting in Babylonia

The first name recorded in the role of librarian was the Babylonian Amilanu.[1] He worked around 1700 BCE. Recording the contents of libraries was commonplace by then, so we can reasonably assume that one of his roles was to make notes on the contents of his library's collections so his readers would know what he had collected.

The task of recording the contents of libraries is more than an instinct or a compulsive tic exercised by librarians; it began as a way to broadcast to readers what is available among the stacks of materials. The tradition of open stacks of printed books is paradigmatic to modern American library users, but ancient libraries featured stacks of clay or pre-paper scrolls that resisted browsing. And even into the age of books and printed journals in the following twenty-one centuries, many private and public libraries did not allow their readers to browse the stacks. The librarian with a deep knowledge of the contents of the collection (and the collections of kindred institutions) was the guide to what the reader could borrow, and it was through an interview with the librarian that the contents of the collection were fully revealed. However, recorded catalogs were an invaluable tool for librarian and reader alike. The catalog provides a permanent record of the collection over time and changing library staff. So the recording of collections on clay, paper, and later, electronic media is more than an instinct; it has always been a valuable tool for creating a permanent memory and map of the collection.

The historian of cataloging, Dorothy May Norris, tells us that the first known recorded catalog was written directly on the walls of the library of Edfu in Upper Egypt.[2] If one's goal is to broadcast the contents of the collection to readers in the library, the painted catalog is remarkably effective. This is a positive founding principle of the catalog: write down what is in the collection so your readers will know what you have—and in the Edfu case, do it in a way that broadcasts the details to all who enter the building.

The earliest librarians created rules for how to record the details of the catalog. By 700 BCE the Assyrians followed the rules set down by the Babylonians. The seventh century BCE Babylonian library in Akkad was lead by the librarian Ibnissaru who prescribed a catalog of clay tablets by subject.[3] Subject catalogs were the rule of the day, and author catalogs were unknown at that time. The frequent use of subject-only catalogs hints that there was a code of practice among early catalog librarians and that they followed some set of rules for subject assignment and the recording of the details of each item. These rules created efficiency through consistency—the catalog librarian knew how to record each item without reinventing the rules each time, and the reader knew what to expect with each visit.

It is interesting to note that catalog librarians now have at least 2,700 years of experience creating rules for how to record the details of what is in a collection. And some of the principles, such as the value of subject description, have retained value for all of that time.

The first known catalog on paper was in the library of the Ptolemies at Alexandria, Egypt, around

280–240 BCE.[4] It was written in ink on rolls of papyrus. Thus were the first hand-crafted catalogs painstakingly created and corrected as the collection was pruned and amended.

Medieval European Catalogs

In summarizing the history of medieval European catalogs, Norris describes things with this economy of words: “The first ten centuries of the Christian era tell us little of libraries or their catalogues.” She guesses that “they are still buried beneath the dust of ages and awaiting the spade of the archaeologist and the antiquarian.”[5] However, there are two notable catalogs from that era that seem to have taken inspiration from the ancient catalog of Edfu in Egypt, the catalog that was written directly on the walls. The first was engraved in marble for all to see, and the second on paper, but in verse to inspire the spirit and capture the imagination of scholars.

The librarians at the Church of St. Clement in Rome, working for Gregory the Great, took the effort to engrave their catalog in marble and wrote, in part:

> The people of Israel in the country used to offer to the Lord, one indeed gold, another silver, some also bronze, some indeed, the fleeces or skins of goats. But I, unhappy that I am, Gregorious First, Presbyter of the fostering apostolic seat, and bearing the responsibility of this blessed title, the highest client of Clement, offer to Thee, O Christ, from the treasuries, these little gifts in the time of the most Holy Zacharias, the high priest. I offer [these treasures] through Clement, thy witness and saint, by whose merits may I deserve to be free from my sins and to enter into a blessed and eternal life. Thou hast said the Kingdom of Heaven is worth all thou hast. Receive these books, Lord, I beg, as the mite of the widow—these books of the Old and New Testaments, of the Octateuch, Kings, Psalms and of the Prophets, Solomon, Esdras full of stories therein found. Seek, reader, the continuance of these syllables.[6]

This is cataloging through prayer, and the last sentences that describe the collection don’t appear to follow any scheme for cataloging rules, but certainly contribute to the art of describing a collection. Demonstrating a similarly grand approach to describing a collection, Alcuin of York used poetry to describe the books of the monastic library in the monastery of St. Martin’s of Tours in York around 782. There he wrote the verse that begins

> There shalt thou find the volumes that contain
> All of the ancient fathers who remain;
> There all the Latin writers make their home
> With those that glorious Greece transferred to
>   Rome;
> The Hebrews draw from their celestial stream,
> And Africa is bright with learning’s beam.
>
> Here shines what Jerome, Ambrose, Hilary,
>   thought
> Or Athanasius and Augustine wrought.
> Orosius, Leo, Gregory the Great,
> Near Basil and Fulgentius coruscate.
> Grave Cassiodorus and John Chrysostom
> Next Master Bede and learned Aldhelm come,
> While Victorinus and Boethius stand
> With Pliny and Pompous close at hand.
>
> Wise Aristotle looks on Tully near.
> Sedulous and Juventus next appear.
> Then come Albinus, Clement, Prosper too,
> Paulinus and Arator. Next we view
> Lactantius, Fortunatus. Ranged in line
> Virgilius Maro, Statius, Lucan, shine.
> Donatus, Priscian, Probus, Phocas start
> The roll of masters in grammatical art.
> Eutychius, Servius, Pompey each extend
> The list. Communion brings it to an end.
>
> There shalt thou find, O reader, many more
> Famed for their style, the masters of old lore,
> Whose many volumes singly to rehearse
> Were far too tedious for our present verse.[7]

There we have the first and perhaps last catalog in verse and an early admission that there might be too many things for the catalog librarian to describe—therefore only the most critical or in-demand titles are immortalized in this literary catalog. Both of these examples of early catalogs demonstrate a commitment to visibility. It’s quite possible that they are exceptional, that they demonstrated a unique drive to capture the attention of the reader and sit above a history of written catalogs less visible and available to readers. Whatever the case, they demonstrate a desire to broadcast, or market, the library’s content to the reader in the most effective means available.

The Card Catalog

Eventually the mechanization of the modern era brought the efficiencies of card catalogs. It was around 1780 that the first card catalog appeared in Vienna.[8] It solved the problems that were present in the structural catalogs in marble and clay from ancient times and the later codex (handwritten and bound) catalogs that were manifestly inflexible and presented high costs in editing to reflect a changing collection. Slightly earlier, Conrad Gessner, the sixteenth-century Swiss botanist and proto-catalog librarian, described the process of “cutting up pieces of information on paper so as to (re)arrange them more readily.”[9] Again, this was an advance over the codex approach to catalogs, which did not allow efficient sorting and resorting. The Viennese librarians of the eighteenth century took this principle one step further and efficiently put their slips in cabinets. In his book *Paper Machines*, Markus Krajewski marvels at the efficiency of this process: “What differs here from other data storage (as in the codex book) is a simple and obvious principle: information is available on separate, uniform, and mobile carriers and can be further arranged and processed according to strict systems of order.”[10]

Thus, “systems of order” are advanced from the written word in a codex into sorting and searching systems that provide massive efficiency to the catalog librarian. For the reader, the benefit is secondary. Compared to the codex, the card catalog can be created and updated much faster, and the presentation of the data is uniform across the catalog. Catalog librarians have rules for the description of bibliographic items and a highly efficient method for describing them. This science of catalog librarianship matures and becomes a significant component of investment for the library. And as collections grow and mechanized printing expands dramatically, the tasks before the catalog librarian also expand. As with the medieval librarian whose bibliographic poem ends after he tires of recording the lesser-known authors, we see the first risk—that the reader fades from focus and the maintenance of the infrastructure becomes the primary task.

In the United States in the 1870s, Melvil Dewey led the charge for scientific management of catalogs and the general library infrastructure. He also presaged the rise of union catalogs of cataloging data by a hundred years when he wrote, “Cataloging, indexing and the score of things which admit, are to be done once for all the libraries.”[11] Matthew Battles quotes Dewey’s biographer in his book *Library: An Unquiet History*: “He was convinced the best way to maximize the library’s potential was to create effectively uniform collections of quality materials and increase service efficiency by standardizing internal library procedures with common forms, appliances, and rules and systems of arrangement.”[12] And in an echo of a debate that carries on today about how much effort to put in customization of library data, Battles observes, “To Dewey, local interests and special needs were less important than the efficient movement of books into the hands of readers.”[13]

“The efficient movement of books into the hands of readers” could have easily become an operating principle of libraries, but there is little evidence that it did. The history of the coming hundred years of librarianship is one of increasing focus on efficiency and service to the infrastructure.

Library Automation

Christine Borgman, who is now the Distinguished Professor and Presidential Chair in Information Studies at UCLA, has studied the history of library automation and points out that in the United States and Europe during the 1960s, there were several forces that enabled libraries to once more dramatically improve their efficiency in catalog management: the availability of advanced computer technology, “long traditions of shared and distributed cataloging,” and “ready access to highly developed telecommunications infrastructure.”[14] All of these factors made it possible for library leaders to invest in automation of library processes and in the movement from purely paper-based systems to mainframe-based systems with significant processing power and data storage capabilities. For libraries, this meant a significant advance in the ability to store and duplicate catalog data across systems. It also meant the ability to improve the speed of some routine transactions and perhaps reduce the possibility of transaction errors.

During this period, libraries invested in the efficiency of internal workflow functions: circulation, acquisitions, serials control, and cataloging. It was also the birth of systems that allowed libraries to share catalog data at large scale to reduce costs for all in the sharing network. Borgman’s summary of the period tells us that this happened in the United States beginning in 1967 with the advent of the alphabet soup of data-sharing networks: OCLC, RLIN, and WLN.[15] Similarly, in the United Kingdom, the BLCMP and CURL networks were organized, and the PICA system in the Netherlands did the same to offer the benefits of data sharing at scale to Dutch libraries.[16] All of these systems take advantage of expanded computing power to reduce costs and calling back Dewey’s idea that cataloging “be done once for all the libraries.” In all of this, there was no significant focus on direct improvements for readers—the focus was on system efficiency and cost savings. In fact, it is interesting to observe that this period in the development of professional librarianship represented a significant investment in the industrialization of the library infrastructure. Cost savings, efficiency, reduction in transaction costs—all were designed to save the librarian effort and to meet the demands of the dramatically expanding world of published materials. Curiously absent is a direct and explicit focus on the needs of the reader in this effort.

Because bibliographic data was now being stored at a larger scale in computer systems, it quickly became clear that there would be advantages in standardizing the specifics of how that bibliographic data was stored and exchanged between institutions. Borgman explains that the late 1960s saw the birth of standard formats for the efficient storage and exchange of cataloging data.[17] The Library of Congress was the first to invest in a study and pilot of standardized machine-readable cataloging (MARC) in the mid-60s. By 1968, it had a service in place to distribute these MARC records to libraries and partners at scale.[18] Soon after, it collaborated with the producers of the British National Bibliography to produce a variant suited to the needs of the UK library market.[19] In the 1970s, the International Federation of Library Associations (IFLA) sponsored an effort to develop a system of machine readable cataloging that suited the particular requirements of European libraries that they called UNIMARC.[20] Previously, in 1969, IFLA had sponsored an important effort to finally standardize the rules for cataloging into the International Standard for Bibliographic Description (ISBD).[21] ISBD had a particular emphasis on the order of bibliographic elements and standardization of punctuation as these were essential elements for promoting uniformity on catalog cards. Clearly, global librarianship was fully invested in the industrialization of library infrastructure and in particular the efficiency of catalog building and data operations.

Almost exactly a hundred years after the introduction of the card catalog in Austria, libraries realized that these computer systems for catalog automation could be used to allow readers and not just library staff to search and discover what is in the library's collections. This happened in the 1970s for both academic and public libraries. The Ohio State University introduced the first of these catalogs in 1975 and the Dallas Public Library did the same in 1978.[22] Even with the simple non-keyword searching mechanisms that were in place at the time, libraries realized that automated systems had advantages over the physical card system for readers. Matthew Battles tells the story of the American librarian Edmund Pearson, who in 1909 fretted for the reader trying to use the old card catalog: "Harrowed individuals are seen trying to think if the name of Thomas De Quincy will be found in the drawer marked De or that labeled Qu. Then they make the choice—always wrong—and are seen, with pain only too apparent on their brows, dashing off to the other drawer."[23] The automated catalog brought the promise of eliminating those kinds of problems.

Automated catalogs evolved through the next two decades and finally offered some benefits to the reader: truncated phrase searching, keyword searching, and permuted keyword searching where the order of the search terms didn't matter. All of these improvements made searching easier and more fruitful for the reader. These were the first significant advances in catalog technology that benefited the reader in a hundred years. And in an age where the library was seen as the essential source for resources that the reader needed, that was a leap forward. As we know now, by the 2000s readers no longer see the catalog as the primary place for discovering things, but in the days when the print collection was everything and the library catalog was the primary tool for discovery, automation meant progress and improvements for the reader.

The Internet

By the late 1990s, the Internet age dawned and libraries quickly saw the value of making their catalogs available to their peers and the world. They did this first through text-based catalogs available over Internet protocols like Telnet, and then in the mid- to late 1990s via the web. For readers there was little change in the features they used for searching, but the ability to access the catalog through a web browser from anywhere provided convenience and flexibility. It also improved access to library catalogs around the world. For the serious researcher this was a benefit. However, it's debatable how important it is for readers to see catalogs with materials they don't have immediate access to, but certainly for advanced scholars this was a useful change, and it marks a recognition that the web is an important venue for discovery.

The most recent advance in library catalogs that offered advantages to readers came in the mid-2000s. At that time library technologists began to follow the trends in searching on the web and the technologies available for indexing textual data. This is the same time that the search engines were demonstrating that search could be accomplished with enormous advantages for the searcher. Relevance ranking and the full embrace of keyword searching became the dominant model for searching, and the library's approach to complex keyword and phrase searching began to look more like the card catalog than a modern search interface. Library users brought these expectations with them to library catalogs, and the catalogs did not look appealing after the comparison. It was at this time that next-generation catalogs were introduced by libraries willing to experiment with new systems and entrepreneurial library systems vendors. These systems that were not based on the library's local inventory management system succeeded in introducing several new features for readers: better indexing, relevance ranking, "Did You Mean" features that mitigated the failures of the reader to consistently spell common and uncommon words,[24] and finally the introduction of integrated databases of articles. Given the enormous importance of articles to academic library users, this was a significant step forward.

However, as good as these systems were for readers, they still didn't bridge the gap between searching on the web and searching the local catalog where readers could find the full details of the collection and availability.

Notes

1. Dorothy May Norris, *A History of Cataloguing and Cataloguing Methods, 1100–1850: With an Introductory Survey of Ancient Times* (London: Grafton, 1939), 1.

[2] Ibid., 3.

[3] Ibid., 3.

[4] Ibid., 4.

[5] Ibid., 7.

[6] Ibid., 7.

[7] Ibid., 9–10.

[8] Markus Krajewski, *Paper Machines: About Cards and Catalogs, 1548–1929*, trans. Peter Krapp (Cambridge, MA: MIT Press, 2011), 37.

[9] Krajewski, *Paper Machines*, 3.

[10] Ibid.

[11] Matthew Battles, *Library: An Unquiet History* (New York: W. W. Norton, 2003), 141.

[12] Wayne Wiegand quoted in Matthew Battles, *Library: An Unquiet History* (New York: W. W. Norton, 2003), 141.

[13] Ibid.

[14] Christine L. Borgman, "From Acting Locally to Thinking Globally: A Brief History of Library Automation," *Library Quarterly* 67, no. 3 (July 1997): 215–49.

[15] Ibid., 220.

[16] Ibid., 220.

[17] Ibid., 220.

[18] Ibid., 220–21.

[19] Ibid., 221.

[20] Ibid., 221.

[21] Ibid., 221.

[22] Online Public Access Catalog, last modified 10 February, 2016, https://en.wikipedia.org/wiki/Online_public_access_catalog.

[23] Edmund Pearson, quoted in Battles, *Library*, 14.

[24] Kristen Antelman, Emily Lynema, and Andrew K. Pace, "Toward a Twenty-First Century Library Catalog," *Information Technology and Libraries* 25, no. 3 (2006): 128–39, http://dx.doi.org/10.6017/ital.v25i3.3342.

Chapter 5

The Current Landscape

Narrative descriptions of where libraries want to be relative to the reader's experience of searching on the web are difficult, if not impossible, to find, but detailed descriptions of what some libraries are doing relative to web technologies are abundant. This means that libraries are investing significantly in some of the dimensions of technology, but the community's goal and commitment to the convenience of the reader isn't articulated.

There are several important moments in the movement toward change in library catalogs. An important one was Roy Tennant's 2002 *Library Journal* article "MARC Must Die," which argued that the current data carrier MARC could be replaced by more modern carriers designed in the age of the web.[1] A more extensive treatment of the issue and an argument for the need for change from the machine-readable cataloging systems originally developed in the 1960s is in the report from the Library of Congress's Working Group on the Future of Bibliographic Control, which published its recommendations in 2008. It wrote, "The library community's data carrier, MARC, is based on forty-year old techniques for data management and is out of step with programming styles of today."[2]

The Working Group's charge was not specifically to solve the problem of raising the visibility of libraries on the web, but its work became the springboard for the central initiative around a movement in libraries to make their data more web-accessible. This became the Bibliographic Framework Initiative, and it used the Working Group's report as a base and inspiration.

The Bibliographic Framework Initiative (BIBFRAME)

The Library of Congress activity called BIBFRAME declares in its mission the goal to enable better expression of bibliographic data on the web. Its website describes it this way: "BIBFRAME provides a foundation for the future of bibliographic description, both on the web, and in the broader networked world."[3] In practice, the work is primarily focused on the process of replacing the current MARC standard for exchanging bibliographic data between library systems. The inspiration from the Working Group report to modernize the "community's data carrier" is very much alive in the work of the Library of Congress staff. The mission of the initiative makes that drive explicit by declaring that BIBFRAME is "a replacement for MARC" and that "a major focus of the initiative will be to determine a transition path for the MARC21 formats while preserving a robust data exchange that has supported resource sharing and cataloging cost savings in recent decades."[4]

The language of the BIBFRAME mission statement and the work itself continue the tradition of seeking greater efficiency in data exchange and management.

Beacher Wiggins, the Director for Acquisitions and Bibliographic Access at the Library of Congress, extends the mission and goals to a broader purpose, saying that web visibility for library collections is "one of the topmost desires of BIBFRAME."[5] His decades of experience with describing the LC's collections provides the kind of intimacy with those collections and awe for their depth that leads him to describe them as an "incredibly valuable part of the nation's intellectual and cultural patrimony."[6] However, he cautions, "There is a dormancy to the content and we render it less valuable if we don't have ready access to it."[7] The LC's primary mission is to its funder, the United States Congress, but it has long held a position of leadership in data exchange standards and the production of high-quality data to be shared among all US libraries.

Given that tradition and the technical assets the LC has today, there is a natural inclination toward a focus on replacement of the data exchange infrastructure.

The work of the BIBFRAME initiative is focused on creating what specialists call a vocabulary for expressing bibliographic data. The LC is also engaged in a pilot to experiment with creating BIBFRAME-native data. It is doing this in parallel to the existing workflows for creating the traditional MARC21 data. The goal of the project is to test the data creation and management tools it has created as part of the BIBFRAME project.

While BIBFRAME's mission and activities do not explicitly address the convenience of the reader, BIBFRAME does have a role in contributing to some of the best practices for playing by the rules of the web—specifically, the rules around the Knowledge Card component of search engine results. Given Richard Wallis's suggestion, mentioned in chapter 2, that "semantic properties will prove more fruitful and effective than simple words,"[8] it is important to express those properties in a way that the web will recognize and reward. BIBFRAME is therefore a vocabulary for libraries to express their collections on the web in a way that is generally consistent with Semantic Web best practices. Jeff Penka, Vice President for Product Management at Zepheira, the consulting company that contracted with the Library of Congress on the first version of the vocabulary, has described it as "an industry standard for libraries that can be projected into the meaningful vocabularies on the web."[9] This doesn't mean that BIBFRAME itself is not meaningful; it means that libraries are declaring their own dialect for expressing data on the web, a dialect that can be translated into the recommended languages on the web such as schema.org. The quality of the dialect will be measured by how well it can be translated without loss of meaning or intent. This is a subtle and highly technical measurement, and its success will be measured over time.

Thinking back to the practices that the search engines promote for improved relevance of content, this is the right time to raise questions about the guidance that catalog librarians use for bibliographic description. Beacher Wiggins reports that "RDA is the content standard" for the creation of bibliographic data when using the BIBFRAME vocabulary.[10]

Resource Description and Access (RDA) provides guidance and instruction for catalog librarians. It tells them how to make decisions about what a title is and if they should be concerned about the punctuation included in the title and author information on the thing being cataloged. But it also contains a set of vocabularies that can be used to express bibliographic data in a Semantic Web context. The recent history of RDA shows a transition from a ruleset focused on the traditional activities of cataloging and limited by the logistical restrictions of cataloging on physical cards; this includes things like the transcription of text from the title page to a system for recording bibliographic data, to a framework of instructions and Semantic Web vocabularies. The library metadata expert Diane Hillmann calls RDA "a coordinated set of vocabularies and guidance instructions capable of capturing the rich relationships of bibliographic entities."[11] According to Hillmann, because RDA is based on sophisticated models of entity relationships such as the Functional Requirements for Bibliographic Records (FRBR) and newer Semantic Web vocabularies, it produces data that can express rich relationships that allow discovery systems to "navigate the bibliographic space."[12]

This model is a departure from the legacy Anglo-American Cataloging Rules but has required significant revision to approach a standard that can guide catalog librarians to creating data optimized for exposure on the web. A sharp critique of the early release of RDA was expressed by Mikael Nilsson of the Knowledge Management Research Group, Royal Institute of Technology, Stockholm. He said the rules are "stenographic conventions for constructing value strings."[13] The implication is that the ghosts of catalog card production are haunting the work that is meant to modernize bibliographic description. But precisely because of those criticisms and the devastating published criticisms by Hillmann and Karen Coyle in 2007,[14] the body responsible for RDA has undertaken revisions. More recently, RDA as a whole has been described by Gordon Dunsire as "a package of data elements, guidelines and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focussed linked data applications."[15] This is a positive trend and focus on the effectiveness of RDA in producing data optimized for web exposure should continue.

Library of Congress staff are engaged in a number of activities to develop and promote the BIBFRAME vocabulary among US libraries. LC staff can be seen at professional library conferences presenting to librarians the latest changes to the vocabulary and the LC's plans for production implementation. Full production requires significant retooling of the programs and methods used by the LC's cataloging teams. This is a decades-old infrastructure with significant current investment. It will likely be a long process for the LC to switch from current systems to new systems based on the vocabulary. The LC has publicly made this commitment and regularly reports on its progress.

BIBFLOW

The BIBFLOW project, whose formal title is Reinventing Cataloging: Models for the Future of Library

Operations, is centered at the University of California, Davis and is funded to reinvent

> cataloging and related workflows, in light of modern technology infrastructure such as the Web and new data models and formats such as Resource Description and Access (RDA) and BIBFRAME, the new encoding and exchange format in development by the Library of Congress. Our hypothesis is that, while these new standards and technologies are sorely needed to help the library community leverage the benefits and efficiencies that the Web has afforded other industries, we cannot adopt them in an environment constrained by complex workflows and interdependencies on a large ecosystem of data, software and service providers that are change resistant and motivated to continue with the current library standards (e.g. Anglo-American Cataloguing Rules . . . and MARC.[16]

This mission statement captures an energetic commitment to reinventing the workflows that provide the data that describes library collections. The project's lead, Carl Stahmer, the Director of Digital Scholarship at UC Davis, is motivated to make library data more accessible on the web, saying, "Making library collection data play on the web is crucial." He cautions his library colleagues against maintaining the status quo by saying, "The idea that libraries can continue to operate as a silo alongside the open web is destructive."[17]

The BIBFLOW approach to remodeling library data is sophisticated in the sense that the project leaders want to move beyond a simple statement of what is available in the library to create "relational and comparative systems that allow us to ask different questions about how library data sets are the same or how they are different."[18] They expect to achieve this through a "good push toward the semantic web."[19]

On the question of reinventing rulesets, like RDA, that describe how library collections can be more in line with web practices, Stahmer reports that the BIBFLOW team is explicitly avoiding the "transcription fixation" of legacy description regimes.[20] BIBFLOW has not created an alternative ruleset that is specifically tuned to the needs of optimized webpages, but they are committed to experimentation to establish the "rule of the street."[21] The "rule of the street" is Stahmer's principle to use techniques that get results on the web over historical commitments to legacy models.

On the question of optimization of web-based catalogs for web exposure, Stahmer reports that BIBFLOW rejects the idea of a monolithic discovery system in favor of an array of discovery systems dedicated to thematic collections and tuned to the students and scholars who need them to support their research needs.[22] This is a utilitarian approach that has a very good chance of being rewarded by the search engines. It rejects conventional thinking that massive aggregations of data will automatically attract attention by search engines and embraces the concept that high-quality data that gets traffic from affinity websites will be indexed and the pages will increase their chance of being more relevant to web searches. Stahmer provides the hypothetical narrative that "a graduate student in Malaysia builds a system that connects one of our dedicated collections using open web standards and connects that data set to many other like-configured systems thereby creating the 'best' system for research and specific queries to the data."[23] This is a bright spot in the constellation of projects around visibility on the web and reflects a sophisticated understanding of the requirements of the web.

Linked Data for Libraries and Linked Data for Production

Philip Schreur from Stanford sets the tone for the two projects Linked Data for Libraries and Linked Data for Production when he says directly, "In the future we will be working on the web."[24] To this end, he paints a vision of a distributed network of data shared by like institutions with the express goal of making it more web-accessible. This means shared databases of data built on commonly understood schemas such as BIBFRAME. It will include contributions from multiple affinity institutions with a common goal of representing a wide variety of library assets in a Semantic Web framework.

Schreur is experienced enough to know that the projects do not have a documented recipe for what a distributed data management landscape will look like. He describes this experimentation as a way to feel their way to answering his question, "How will we work on the web in a distributed way?" and acknowledging immediately that "we will not be able to control it."[25] That last comment echoes Carl Stahmer's expectation that the most effective data will be created under the "rule of the street." In the ideal narrative, libraries will experiment with different models for describing their data, and the most effective ones will evolve into a community standard. That's the paradoxical value of loss of control and rule of the street. It will be a culture shift for librarians, but the benefit is aligning with the web's effectiveness and broadcasting content.

Linked Data for Libraries (LD4L) and Linked Data for Production (LD4P) are grant-funded collaborations between libraries with a mutual interest in reinventing their bibliographic infrastructure. The participating libraries are bellwether institutions with strong technical resources, deeply knowledgeable staff, and strong funding from the Mellon Foundation. The Linked Data for Libraries project has a two-year grant for just under $1 million. Because of the participation of three prestigious institutions—Cornell, Stanford, and Harvard—knowledgeable librarians are following their efforts and watching their communications for leadership and results.[26]

The results that the projects predict are highly technical. As with BIBFRAME and BIBFLOW, the focus is on infrastructure. The project website for LD4L declares that "the goal of the project is to create a Scholarly Resource Semantic Information Store (SRSIS) model"[27] that describes a broad spectrum of library assets and follows the rules of the Semantic Web. A subpage of the project website declares that last goal: "Our larger goal is to encourage libraries, archives, and cultural memory institutions to think much more broadly about using structured information about their scholarly information resources to make those resources more discoverable, accessible, and interconnected."[28] The goal therefore is to promote the use of Semantic Web technologies in the service of making a wide variety of things more discoverable.

The project doesn't declare any specific goals relative to the convenience of the reader and search engine results. In a discussion of the question of improving the visibility of library collections on the web Schreur says: "[At the beginning of the Bibliographic Framework Initiative] we were told that was the goal."[29] But he emphasizes that the LD4L and LD4P projects are "not just moving to the web"; they plan to "play by the rules of the web" in making a broad definition of their data accessible on the web.[30]

The projects are notable for their broad view of library assets. This group seems more keenly aware of the principle that academic library users are interested in a wide range of things to support research and learning. The inclusive language of "scholarly information resources" abstractly hints at it, but when you talk to project leaders, the enthusiasm for a broad definition of things that they are responsible for exposing is evident. Schreur's enthusiasm for the mandate from Stanford University is infectious, and it is shared by his colleagues at Cornell University, who are building on their success of describing the universe of Cornell scholars in the VIVO system. Cornell's VIVO project describes not just published things, but also includes durable descriptions of the persons who authored them.[31] This positive feature of the project acknowledges that a definition of the library collection such as "books" is too narrow to satisfy the academic library reader.

During the period of active funding, the project expects to create several technical and infrastructural deliverables:[32] an ontology, a management system for the discovery and updating of the assets of each institution. Notably, it will allow import from a wide variety of local systems at each institution. These include the MARC-based library catalogs, local systems containing the institution's knowledge of its researchers—the person's scholarly outputs, awards, specialties, and so on. It will also include pathfinder systems—these are topic and curriculum-based lists of resources used by students and scholars interested in a given topic. Pathfinders are curated by subject specialists in the libraries. This commitment to a wide variety of inputs to be converted to data formats that are more readily exposable on the web reveals a commitment to a broad definition of discoverable things. Finally, for the convenience of specialists at other like-minded institutions, the project will deliver the technical infrastructure to allow other institutions using the Project Hydra content management system to discover the data in the project's main database. On the question of redefining the rules for cataloging and web discovery to optimize pages and data for search engines, the commitment is similar to the BIBFLOW. Schreur explains that they are moving away from an "emphasis on transcription" and they must "play by the rules" of the web.[33] He acknowledges that the current rulesets were built in an environment that was "designed to represent catalog cards" when collation and exact transcription were paramount.[34] Those requirements are less important now when the structure and semantics of the webpage are rewarded or punished by the search engines.

The Linked Data for Production project is a collaboration between the LD4L libraries and other institutions that have a vision for a complete transformation of their technical processes. The current academic library processes for acquiring the data for their traditional catalogs and the related databases describing persons, programs, pathfinders, and so on are generally optimized for legacy data formats designed either before the web or just not responding to any imperative to make the data discoverable on the web. This is why institutions like the Library of Congress, Harvard, Stanford, Princeton, Columbia, and Cornell are participating in an effort to redesign and retool their technical processes. Once again, the focus here is on technical processing and the efficiency of the librarian's workflow.

Integrated Library System Vendors and Bibliographic Utilities

Since the 1980s, US libraries have relied on a set of mostly commercial providers for their enterprise systems. These providers sell locally installed and cloud-hosted software that allows the library to efficiently manage its inventory, purchasing, and reporting systems. These systems also include a discovery layer that provides a view into the library's inventory of books and journals. Libraries are now augmenting these systems with a free-standing discovery layer that exposes the traditional collection and the articles that are so critical to the reader.

Twice a year, American librarians gather for a professional conference that features a panel discussion on BIBFRAME implementation that includes representatives from the library system vendors with the biggest market shares: Ex Libris, Sirsi/Dynix, and Innovative Interfaces. The panels also include representatives from the Library of Congress, the library cooperative OCLC, and Zepheira. The content from the library system providers affords a good description of their commitment to enhancing the visibility of libraries on the web. That content generally falls into two categories: a general support for the value of linked data and BIBFRAME, and a statement that changes to their systems will be considered in their future roadmaps. The enthusiasm for linked data and BIBFRAME is genuine, but the specifics in roadmaps tend to be more vague. There are some exceptions: the academic and public library vendor Innovative Interfaces highlights its partnership with Zepheira in providing BIBFRAME orientation to libraries (what is it and what experimental tools are available) and an explicit statement that it is committed to external partnerships over changes to the local system.

OCLC is the library cooperative that offers bibliographic data and a wide range of workflow and discovery services to libraries. The research and data science arm has distinguished itself by its experiments with transforming legacy bibliographic data in MARC format into the kind of representations that could be useful in an environment where libraries are playing by the rules of the web and using global identifiers. These global identifiers refer to the things that readers want to acquire from libraries such as bibliographic works and the publication history of the persons who contribute to those works. OCLC Research has produced a linked data representation of persons. Persons are defined as identities or corporate bodies that have done things like written, illustrated, edited, performed, translated, or otherwise adapted bibliographic entities. OCLC Research has done this by joining the authoritative descriptions from national libraries and other important bibliographic agencies throughout the world. It uses big-data tools and world-class data scientists to process the data into a web-accessible graph. This service creates consumable forms of the authors, editors, translators, and so on who contribute to bibliographic works. OCLC calls it the Virtual International Authority File, and the identifier, included in the data, is considered by experts in the library Semantic Web space to be the canonical identifier for persons. This status has been earned by OCLC Research's active management of the data and the reputation of the contributors for careful management of data and high quality standards. The Virtual International Authority File was created by OCLC Research in deep collaboration with national libraries and other sources of authoritative data. The identifiers are already used by the web-accessible data experiments produced by the national libraries of France, Sweden, and Spain. This data is potentially valuable because it contains authoritative descriptions of persons that can be used in local and global knowledge graphs for searching and for linking the bibliographic works that the persons created or contributed to.

The business side of OCLC provides a range of applications and traditional bibliographic data to thousands of libraries worldwide. In addition to the existing WorldCat.org site that allows crawlers to harvest its titles and uses Schema.org markup,[35] it is building a strategy for enhancing its metadata services infrastructure for a BIBFRAME future. Building on a foundation created by OCLC Research, it has begun a process of augmenting WorldCat data, including processes to model it, assign URIs, and make it suitable for use in linked data contexts. When discussing production use of its linked data assets, John Chapman, Product Manager at OCLC, explains that OCLC wants "to prove the value of the data."[36] In the fall of 2015, OCLC announced a pilot project for a new tool that allows libraries to look up data about persons. This pilot service allows producers of data—including libraries and commercial partners—to enhance their content with authoritative data about persons who have contributed to bibliographic works. Chapman points out that these persons are not limited to creators and contributors, but extend to persons named as topics or subjects of resources. He says they plan to add article authors at some point and "are aware of the need to integrate article authors into the Persons data."[37] If there is uptake on services like the Person Entity lookup service, OCLC has the opportunity to provide data to thousands of libraries and to provide the canonical identifiers that are required by the Semantic Web.

Chapman says OCLC is in close contact with the libraries in the BIBFLOW and Linked Data for Libraries projects and plans to "learn from these projects so we can draw some conclusions about efficient workflows for putting linked data to use."[38]

Ex Libris, the integrated library system vendor to academic libraries, has published its principles and roadmap related to workflows and discovery. Its published information indicates a mix of workflow changes and library catalog (discovery system) enhancements. It describes its "Key Elements of Linked Data for Ex Libris Roadmaps":

> The following principles related to linked data have helped shape the roadmap of the Alma resource management solution:
>
> - The use of linked-data format for loading and publishing bibliographic records.

- URI support for cataloging and technical services: identifying "things" based on URIs instead of simple identifiers.
- Access to linked data to enrich data displayed to staff in routine workflows.
- Support for the BIBFRAME model and ontology as they mature.

> The following principles have helped shape the roadmap of the Primo discovery and delivery solution:
>
> - Discovery of the underlying metadata and access to it via URIs.
> - The use of linked data by non-library applications.
> - The discovery system as the key interface to make data accessible to people and computers.
> - The use of RESTful APIs to provide support for applications based on linked data.[39]

Ex Libris's detail on BIBFRAME relates very specifically to library workflows:

> Alma will support both the export and the import of catalog records in BIBFRAME format. Thus Alma records will be part of BIBFRAME-based record workflows outside Alma. A new option will be added to the title-level export job, so existing MARC-based bibliographic records will be exportable in BIBFRAME format. Similarly, imported catalog records in BIBFRAME format will seamlessly become part of the Alma catalog, regardless of the format in which the catalog is managed. Alma will use the metadata import framework with BIBFRAME as a source format.[40]

Schema.org and Schema Bib Extend

In addition to the rules for crawling and indexing described earlier, the world's biggest search engines have declared their preference for how they want data on websites to be represented. Their preferred markup, called schema.org, is optimized for expressions of data that emphasize Semantic Web principles such as canonical identifiers to unambiguously represent things and the representation of "offers," which are the terms of purchase or lending of inventory items or services such as a car rental or movie showing. A group of librarians, consultants, and commercial vendors has quietly and effectively influenced this preference through collaboration and effective recommendations to the schema.org editors. Led by Richard Wallis, this group, Schema Bib Extend, has taken a highly pragmatic approach to inserting changes to schema.org that make descriptions of bibliographic items more precise. As with Linked Data for Libraries, the explicit mission is technical and aimed at the quality and precision of the infrastructure. The group declares its mission to "discuss and prepare proposal(s) for extending Schema.org schemas for the improved representation of bibliographic information markup and sharing."[41]

Schema Bib Extend has made suggestions that schema.org allow new properties that let a site declare that a work is a translation of another work, or that the work is a newspaper. These are seemingly obvious declarations, but they were not available in the schema.org vocabulary, and the group used its collective knowledge and experience to recommend them and a small set of other changes to the schema.org editors. Wallis describes the successes of the BibEx group:

> - Less-commercial wording—Sounds simple but was very effective (Just adding "or to loan a book" to the description of offer is a benefit for libraries)
> - Citation—Moved from an obscure place on MedicalScholarlyArticle onto the more generic and useful CreativeWork
> - Work Relationships—A lightweight version of the complex entity relationship model described by libraries
> - Periodicals—Added ability to optionally describe an article in a PublicationIssue in a PublicationVolume of a Periodical
> - Multi-volume works—Added hasPart and isPartOf to CreativeWork—much broader applicability than just multivolworks
> - Many examples of bibliographic items[42]

Finally, the most significant acknowledgment of the value of input from libraries was in the creation of the new addition to schema.org called bib.schema.org. It contains the specific additions from this group of experts and is a durable contribution to schema.org.

A knowledgeable observer might ask why BIBFRAME is necessary when the search engines have already declared a preference for a vocabulary. The reply to this suggestion from library Semantic Web experts is that it will always be necessary to have a vocabulary that is used within libraries to exchange data at a level of detail that isn't useful on the web. The additional detail would include transaction data, legacy data from the old MARC systems, and anything else that is important for the efficiency of library workflows, but not useful on the web.

Zepheira and Entrepreneurial Efforts

In a time of change, with challenges to familiar ways of working, and perceived threats to the ongoing perceived value of libraries, there emerges the opportunity to provide commercial services around web visibility.

So far, just one company has entered the market to explicitly provide those kinds of services. Zepheira, based in Powell, Ohio, won the original contract to help the Library of Congress define the BIBFRAME vocabulary. It was chosen because it was able to demonstrate familiarity with libraries and experience with Semantic Web technologies. Zepheira's marketing materials use language that is explicit about the issue of libraries' visibility on the web: "The promise of moving library assets to become visible on the web is exciting. It is also a move that will be most successful with planning and foresight into the full range of a library's operations, content, collections, and internal and external partners [*sic*] capabilities."[43]

Zepheira has been unique in seizing the opportunity to offer services that really fall under the category of change management: it explains principles and shows some experimental tools to take advantage of the desire to see what the future technical infrastructure will look like. Technical services librarians are comfortable with the focus on their processes and tools. Zepheira has essentially turned that culture into a business, helping to assuage the librarian's anxiety by explaining process and tools. It offers training services to fill that need.

To provide a community forum for talking, experimentation, and learning, Zepheira founded the LibHub service. Experimental activities involved Zepheira working with libraries to take traditional library data and transform it into web-accessible formats to allow libraries to see what their data looks like in these formats and to learn technical details along the way. Next, the group experimented with how search engines could crawl, index, and use the data.

Zepheira's second line of business, called The Library Link Network, is aimed at the issue of visibility on the web. Its technical and product leads, Eric Miller and Jeff Penka, understand the technical requirements for success on the web, and they are aware of the limitations of the current library catalogs in meeting those requirements. In response to that, they have designed a product that takes the library's traditional data and makes it available for crawling and indexing by the search engines. Their goal is to create a data set that is accessible to the search engines and a data set that is created by a set of algorithms that understand the requirements of the Semantic Web. This is a strong move toward satisfying the requirements laid out for relevance in the traditional search engine results and placement in the Knowledge Card.[44]

In the area of training libraries to understand Semantic Web concepts and the technical details of vocabularies and other Semantic Web infrastructure, Zepheira has provided training services, and more recently, the Library Juice Academy has emerged as a source for those services.

Notes

1. Roy Tennant, "MARC Must Die," *Library Journal* 127, no. 17 (October 15, 2002), http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die.
2. Working Group on the Future of Bibliographic Control, *On the Record* (Washington, DC: Library of Congress, 2008), 24, www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf.
3. "BIBFRAME," Bibliographic Framework Initiative, Library of Congress, accessed February 11, 2016, https://www.loc.gov/bibframe.
4. Ibid.
5. Beacher Wiggins (Director for Acquisitions and Bibliographic Access, the Library of Congress), interviewed by Ted Fons by telephone, 9 November, 2015.
6. Ibid.
7. Ibid.
8. Richard Wallis (Independent Structured Web Data Consultant), interviewed by Ted Fons by Skype, 23 October, 2015.
9. Jeff Penka (Vice President for Product Management, Zepheira, Inc.), interviewed by Ted Fons, 24 November, 2015.
10. Beacher Wiggins (Director for Acquisitions and Bibliographic Access, the Library of Congress), interviewed by Ted Fons by telephone, 9 November, 2015.
11. Diane Hillmann (Partner, Metadata Management Associates), interviewed by Ted Fons by telephone, 4 February, 2016.
12. Ibid.
13. Diane Hillmann and Karen Coyle, "Resource Description and Access (RDA): Cataloging Rules for the 20th Century," *D-Lib Magazine* 13, no. 1/2 (January/February 2007), www.dlib.org/dlib/january07/coyle/01coyle.html.
14. Ibid.
15. Gordon Dunsire, "RDA Data Capture and Storage," (presentation, American Library Association Midwinter Conference, Boston, MA, January 8–12, 2016).
16. "About," BIBFLOW, IMLS Project of the University of California, Davis, University Library and Zepheira, accessed February 11, 2016, https://www.lib.ucdavis.edu/bibflow/about.
17. Carl Stahmer (Director of Digital Scholarship, University of California, Davis), interviewed by Ted Fons by telephone, 17 November, 2015.
18. Ibid.
19. Ibid.
20. Ibid.
21. Ibid.
22. Ibid.
23. Ibid.
24. Philip Schreur (Assistant University Librarian for Technical and Access Services, Stanford University), interviewed by Ted Fons by telephone, 6 November, 2015.
25. Ibid.
26. Linked Data for Libraries (LD4L), main page, DuraSpace wiki, last updated February 10, 2016, https://wiki.duraspace.org/pages/viewpage.action?pageId=41354028.
27. Ibid.
28. Dean B. Krafft, "Expected Outcomes," Linked Data

for Libraries (LD4L), DuraSpace wiki, last updated September 26, 2014, https://wiki.duraspace.org/display/ld4l/Expected+Outcomes.

29. Philip Schreur (Assistant University Librarian for Technical and Access Services, Stanford University), interviewed by Ted Fons by telephone, 6 November, 2015.
30. Ibid.
31. Dean Krafft, Kathy Chiang, and Mary Ochs, "Enhancing the University's Knowledge Management Using VIVO," a presentation given at the LITA Forum, November 2014, http://connect.ala.org/node/230876.
32. Philip Schreur (Assistant University Librarian for Technical and Access Services, Stanford University), interviewed by Ted Fons by telephone, 6 November, 2015.
33. Ibid.
34. Ibid.
35. Ted Fons, Jeff Penka, and Richard Wallis, "OCLC's Linked Data Initiative: Using Schema.org to Make Library Data Relevant on the Web," *Information Standards Quarterly*, 24, no. 2/3 (Spring/Summer 2012).
36. John Chapman (Product Manager, Metadata Services, OCLC), interviewed by Ted Fons by telephone, 10 November, 2015.
37. Ibid.
38. Ibid.
39. Shlomo Sanders, "Linked Library Data: Making It Happen," *Tech Blog*, Ex Libris Developer Network, December 27, 2015, https://developers.exlibrisgroup.com/blog/Linked-Library-Data.
40. Ex Libris, Putting Linked Data at the Service of Libraries: The Ex Libris Vision and Roadmap (Ex Libris, 2015), 5, www.exlibrisgroup.com/files/Publications/LinkedDataattheServiceofLibraries.pdf.
41. *Schema Bib Extend Community Group* (accessed 23 October, 2015), https://www.w3.org/community/schemabibex.
42. Richard Wallis (Independent Structured Web Data Consultant), interviewed by Ted Fons by Skype, 23 October, 2015.
43. Readiness and Visibility Assessment Information Request Form (accessed 29 April, 2016), http://zepheira.com/assessmentinfo/.
44. Jeff Penka (Vice President for Product Management, Zepheira, Inc.) and Eric Miller (President, Zepheira, Inc.), interviewed by Ted Fons, 24 November, 2015.

Chapter 6

Steps to Take

Recall Rachel Fewell's worldview, quoted in chapter 1, that libraries "are in an in-between world where we have two groups of people: those ones who already go to the library and the ones who never think about the library."[1] It is useful to remember that there are risks to the library in this environment. What if the group that never thinks about the library grows? What if younger generations have a preference for free online search services and hold a perception that quality information resources are exclusively available on the web? To reduce that risk, the Denver Public Library has started to experiment with new ways to expose its collection data on the web. By participating in Zepheira's LibHub and Library Link projects, it is willing to invest time and money in the effort to improve its position in search results. Its goal is to try things until it finds something that works and it can determine what libraries should be doing to influence those who "never think about the library."

The way most people think about the library is probably less black-and-white than Fewell's simple two categories: those who rely on the library and those who never think about the library. In the real world, probably enough people don't think about the library much or used to think about the library more in a previous stage in their lives. The challenge is to reach the ones who do think about the library and the ones who sometimes think about the library—and to reach them when they are seeking answers outside of the library catalog.

With that more nuanced view in mind, it is worth suggesting some steps that libraries could take to improve their position in web search results. These will include both technical and organizational changes, including new business models for success.

Keeping in mind what we know about the technical requirements for appearing in search results, there are really two main approaches to the process, direct partnerships with search engines and playing by the rules, as shown in table 6.1.

Direct Partnerships with Search Engines

Generally the business model here is to pay money directly to Google to be part of its Sponsored Links program. Simply put, libraries could do this to improve their visibility. The search engines use clues to determine the searcher's physical location and identity so the search results including library holdings would show local library institutions.

It is also possible that libraries or library organizations could create direct agreements with Google to place results in the Knowledge Card section. These direct partnerships do not have to involve the exchange of money. The compensation agreement between Google and library organizations could be based on an exchange of money in either direction or some other mutually agreed business arrangement. Microsoft's search engine, Bing, has done this with the e-book provider OverDrive to place links for e-book access through local libraries, and there is evidence that it is working. Search engines experiment with different services, and they will typically drop services that do not show value. The OverDrive arrangement to include links to its e-books in Bing's Knowledge Card has been available for more than a year. That is typically a sign that the search engine sees value in the partnership.

The "rule of the street" dictates how data is represented, but it also determines which services survive

Table 6.1. Two main approaches to improving standing in search results

Direct Partnerships with Search Engines	Playing by the Rules
Business model: Pay money to Google	Business model: Follow best practices
Appears in: "Sponsored Links" section	Appears in: Search results
Appears in: Knowledge Card	Appears in: Knowledge Card

and thrive and which ones fade away. The search engines measure all of their services by effectiveness: traffic, utility, value. It is reasonable to assume that if the links to library providers in the Bing Knowledge Card were not used, the service would be discontinued. Steve Potash of OverDrive explains that this program has been in place for more than a year and the traffic is still very strong. That's an indication that "content marketing" for libraries can be effective if the data provider plays by the rules. Potash indicates that OverDrive uses "open industry standards" in its relationship with the search engines and its rule is to, whenever it can, be part of "the fabric and tools of the web."[2] That has motivated its interaction with Bing, its drive to embed library content through Semantic Web exposure, and its widget that allows libraries to embed e-book and audiobook previews into any website. These tools are driving traffic to OverDrive itself and directly to library websites.

It is reasonable to believe that Google might prefer aggregations of library data to minimize the number of individual agreements and data harvesting efforts, but that will work only if the aggregated data satisfies its data quality standards—that is, if the data is of very high quality and has reliable links to fulfillment options. Google doesn't divulge specifics of its data quality management techniques, but search engine optimization experts estimate that link accuracy must be above 95 percent for Google to accept data from a partner. Link performance that falls below that threshold will not be surfaced in results in any of the zones under any agreement. If libraries are going to aggregate their data, they would have to commit to data quality standards equal to or exceeding the data quality standards they apply to their local catalogs.

To maximize the position of libraries in the Knowledge Card section of search results, libraries will have to keep in mind Richard Wallis's exhortation, mentioned in chapter 2, that "semantic properties will prove more fruitful and effective than simple words."[3] This means a commitment to the current best practice for vocabularies—schema.org and bib.schema.org—and a deep commitment to the concepts of internal graph and global graph. In practice, this will mean following the Semantic Web principle that any reference to a thing (a person, a place, a concept, an event) should use identifiers that are used elsewhere in the local or global graph. This is the principle of universal identifiers. These could be in place for Works, Persons, Events—entities that are well described already by libraries. Using those identifiers across all library catalogs could be recognized by Google as a system of interlinking and a collective display of confidence among libraries in the value of the links. In this model all library catalogs would become a kind of community graph that sits somewhere between the local graph and the global graph. It is the best hope for libraries that manage their data locally and aren't typically referred to by other websites. In other words, because other sites on the web don't typically refer to library webpages, all libraries should refer to the same links and therefore refer to each other.

Play by the Rules

Google is explicit on the business model for the traditional search results: it does not exchange money for improved position in the relevance-based results. Therefore, libraries wishing to influence the position of their data in the traditional search results must follow the best practices that are recognized across the web. Some of those rules will create a challenge for libraries based on current practices. The rules could create challenges either because their systems are not optimized for web crawling or because the rules for bibliographic description are optimized for systems developed long before the web and before the convenience of the reader became paramount. Some examples:

- *Is that page blocked?* It is common practice for current web catalogs to be blocked from crawling. Changing this practice could improve results.
- *Adjacency, word frequency, and synonyms.* The current rules for bibliographic description are optimized for earlier catalog systems that focused on traditional sorting and subject indexing, not keyword retrieval and search engine optimization. Libraries could review current practice and establish new best practices to optimize bibliographic descriptions.
- *Data quality and frequent page updates.* The data quality regimes currently in place for bibliographic data are based on a workflow that focuses first on subject expertise (original cataloging done by subject experts in publishing companies and libraries) and then data sharing at scale. Bibliographic records are shared by consortia and in subscription-based bibliographic utilities. The model is increasingly "update once," which

> is positive for local library efficiency, but negative for search engine optimization. Some of the highest quality websites now use crowdsourcing for data management, which produces frequent updates and improved quality over time.

Finally, there is the issue of PageRank. Keep in mind that PageRank is Google's measurement for the number of times a page is referred to by other sites. This presents a significant problem for libraries. The solution to this problem lies in the same principles that will benefit libraries in the Knowledge Card region of search results and in the same recommendations made for improvements to results there: use canonical identifiers and create a community graph among libraries.

Montana State University

"Clearly Google had no idea that we existed."[4] This is Kenning Arlitsch's summary of the visibility of the Montana State University Library before it began to "play by the rules" to enhance visibility. There is evidence that the "play by the rules" approach can work for libraries, and Arlitsch and his colleague Patrick O'Brien, the Semantic Web Research Director, have experimented thoroughly to prove it to themselves. Arlitsch and O'Brien have gotten results in two areas of web visibility: the visibility and accuracy of Google search results for the library as a physical entity, and the visibility of digital collections of interest to specialized researchers. The work on the visibility of the library entity itself is the most persuasive.

After documenting clearly that Google had a poor definition of the library and inaccurate details about location and contact information in the Knowledge Card, the MSU team went about fixing the problem. Armed with a knowledge of Semantic Web principles, the team knew that Google is using the Google knowledge graph drawn from DBpedia to show results in the Knowledge Card. Arlitsch says, "We know how to fix this problem." So the team went about improving the *Wikipedia* article on the Montana State University library and saw immediate benefits. The quality and therefore utility of the Knowledge Card information for the Montana State University Library improved.

Arlitsch and O'Brien have presented and written widely on their experiments with institutional and collection visibility. In many ways their books and articles serve as how-to guides to playing by the rules.

Library Collaborations

Given the technical and business model requirements for significant improvement in search results, library associations or even commercial support organizations could provide a number of specific actions for libraries:

- Data aggregation to allow frequent data quality updates and crowdsourcing of improvements—even nonexpert update of the data following the *Wikipedia* model
- Data quality monitoring with an eye to optimizing data for search engine best practices
- Promotion of canonical URIs to promote the growth of the community graph
- Negotiation with search engine companies for agreements on data harvesting and commercial terms for exchange of value
- Monitoring current developments in data presentation and Semantic Web technology
- Negotiation with local library system providers for technical changes to local catalogs

Libraries that recognize the risks of poor performance in search engine results should review the readiness of current library associations and support organizations and be prepared to inject these roles into those institutions or seek new ones that respond to their needs.

The Role of BIBFRAME

Efforts like BIBFRAME to modernize and, more specifically, prepare library data for the web are a positive step forward. However, to focus entirely on the data container is to continue the pattern that focuses on internal processes instead of the needs of the reader. The entire ecosystem of linking, data quality, data aggregation, and formal relationships with search engines must be of equal importance, or the risk of continued poor performance in search engine results will continue.

Defining Success

Success for libraries on the web must follow the path of the disruptive influence of relevance ranking and comprehensive indexing of the open web on search and discovery: expedited access to relevant results. web searchers reacted positively to that development because the service was convenient and the perception of usefulness was high.

If libraries can make their collections and services more visible on the web, then libraries should experience a cumulatively positive effect of each connection between search, discovery of the library's assets, and links to fulfillment sponsored by the library. Each moment of discovery and link to fulfillment should contribute to the overall positive value proposition of the library and its offerings. Recognition of the importance of the convenience of the reader and responding to the individual content preferences of the reader will be key elements in achieving that success.

Measuring that success is an important aspect of monitoring progress in satisfying the reader. There are essentially two levels of success that can be measured:

- *"Above-the-fold" results.* Simply summarized, this level of success means that a reader searches for a topic and the library's offerings (books, articles, events, services) appear on the first page in the traditional results, the sponsored links, or the answer panel. This can be measured by regular sampling and by measuring the number of links to the local system from the search engine origin. A dramatic rise in links to fulfillment are a good proxy measurement for highly relevant results.
- *Improved relevance.* This level of success means you have improved overall relevance, but without achieving above-the-fold results. This is also measured by clicks to local fulfillment and increased engagement with non-book and non-article library services.

The distinction between measuring above-the-fold results and general improved relevance isn't arbitrary, it's a matter of degrees. Above-the-fold results are extremely difficult to achieve, but easy to measure. Incremental improvements in relevance and clicks through to fulfillment are more readily achievable and are also easy to measure.

In a heterogeneous environment like the community of library catalogs, achieving above-the-fold results will take tremendous commitment to a declared goal and significant technical, cultural, and organizational change. Given that the first item—an explicit, widely documented goal to improve the visibility of libraries on the web through relevance in search results—is not evident, progress toward this goal is difficult to predict. Defining goals and defining success will be important steps along the road to progress.

Are Libraries Doing the Right Things?

The arc of this review is to answer the original question, "Can we improve the visibility of libraries on the web?" The response can be summarized like this:

The earliest library catalogs, broadcast on the walls of the earliest libraries, were designed exclusively for the convenience of the reader. The history of the development of library systems, and catalogs in particular, features an increasing focus on the efficiency of process without an explicit drive toward the convenience of the reader or focus on the efficiency of getting things into the hands of the reader. The rules for improving relevance in library search engines, with an example focus on Google, are well known and achievable with dedicated action. Libraries are taking action on making their data more accessible on the web, with the focus almost entirely on vocabularies and new systems for storing that data. In that work are some steps that will help improve the visibility of libraries on the web:

- Development of Semantic Web vocabularies that recognize the need for a way to express library assets in the language of the web (BIBFRAME)
- Experiments with expressions of important entities like Persons and Works and the corresponding canonical identifiers (various OCLC services)
- Experiments with new workflows to replace the existing MARC21 workflows and the beginnings of a recognition that library assets extend beyond books (LD4L, LD4P, and BIBFLOW)
- Initial offerings from entrepreneurs that provided conversion of legacy data to data expressed in the web's vocabularies and complementary data hubs to host that data and make it available to search engines following the search engine rules (Zepheira)

However, some of the requirements for improved relevance on the web are not evident in the current efforts toward visibility on the web. Some examples of gaps in current activities, showing other requirements that libraries should be addressing, include the following:

- No evidence of an overall, well-articulated goal of making things convenient for the reader by making library collections and services more visible on the web.
- No widespread and action-motivating commitment to "follow the rules" established by the search engines. This would involve changes to local catalogs and the development of alternative hubs for linking and indexing, changes to the shared rules for descriptive and subject cataloging, commitment to shared canonical identifiers, commitment to linking to other library catalogs, and a generalized commitment to change things that are ingrained today, but must be changed tomorrow as the rules of the web change.
- No evidence of focus on exposing the things that are most highly valued by academic library readers: articles and e-journals. This would involve a change in business models and licensing by the publishers. Achievable, but only with significant coordination and collective commitment.

Addressing the gaps described above would enhance the prospect of improving the visibility of library collections and services on the web.

An even shorter summary of the arc reads: Libraries started with a focus on the reader, then shifted to a focus on the librarian; now it's time to focus on the reader again. Libraries aren't doing the wrong things, but they aren't doing enough of the right things to make a positive impact in the near future.

The imperative for libraries today is to renew the focus on the reader. Just as the search engines have done, libraries must articulate a goal to focus on the convenience of the reader and recognize that readers benefit from a wide variety of library collections and services, beyond just books. Libraries should develop a new language of focus on the reader, recognize a new hierarchy of library assets of interest to the reader, and make a commitment to follow the rules of the web. All of these things will produce inevitable improvements in library service and benefits for the user. And even if the highest goal of above-the-fold search results is not widely achieved, some improved service to the reader and improved satisfaction of the reader will be worth the effort.

Notes

1. Rachel Fewell (Collection Services Manager, Denver Public Library), interviewed by Ted Fons by telephone, October 28, 2015.
2. Steve Potash (Chief Executive Officer, OverDrive, Inc.), interviewed by Ted Fons by telephone, 16 November, 2015.
3. Richard Wallis (Independent Structured Web Data Consultant), interviewed by Ted Fons by Skype, 23 October, 2015.
4. Kenning Arlitsch and Patrick O'Brien, "Establishing Semantic Identity for Accurate Representation on the Web" (presentation, Coalition for Networked Information Fall 2014 Membership Meeting, Washington, DC, December 8–9, 2014).

Notes

Notes

Notes

Library Technology REPORTS

Upcoming Issues	
August/ September 52:6	**The Knowledge Base at the Center of the Universe** by Kristen Wilson
October 52:7	**Google Analytics and Google Tag Manager** by Tabatha Farney
November/ December 52:8	**Social Media Optimization: Improving the Visibility and Use of Library Resources** by Scott W.H. Young and Doralyn Rossmann

ALA TechSource

alatechsource.org

ALA TechSource, a unit of the publishing department of the American Library Association