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).