2016

Food for Thought: Should Libraries Partner with Nonlibrary Search Engine Providers for Their OPACs and Discovery Layers?

Michelle Wu

This paper can be downloaded free of charge from:
https://scholarship.law.georgetown.edu/facpub/2415

This open-access article is brought to you by the Georgetown Law Library. Posted with permission of the author. Follow this and additional works at: https://scholarship.law.georgetown.edu/facpub

Part of the Collection Development and Management Commons, and the Intellectual Property Law Commons
FOOD FOR THOUGHT:
SHOULD LIBRARIES PARTNER WITH NON-LIBRARY SEARCH ENGINE PROVIDERS FOR THEIR OPACs AND DISCOVERY LAYERS?

Michelle M. Wu

Since the development of integrated library systems (ILS) in the 1970s, the focus of developers and librarians has been on the “integrated” aspect. With the advances in technology, though, the time has come for libraries to consider whether a different approach would better benefit their users, at least on the public-facing functions. This article argues that the design of a user-friendly public interface to library systems is ideal for library/search engine technology company (SETC) partnerships. This approach would enable participants to harness their respective strengths while simultaneously limiting the effect of their deficiencies. While this article identifies SETCs as libraries’ likely prospective partners, the collaboration may be equally effective with independent non-profit organizations skilled in general search algorithms, so long as the organization is one that does not rely on the funding of governments or tuition revenue for survival.

1 Michelle M. Wu is the Acting Associate Dean for Administration & Finance, Associate Dean for Information Services, and Professor of Law at Georgetown University Law Center. The author wishes to thank the participants at the Seventh Annual Boulder Conference on Legal Education: Teaching & Scholarship held at the University of Pennsylvania Law School, Philadelphia, PA on July 16–18, 2015 for their helpful review and commentary of this paper.

The goal of this article is to advocate for actively seeking solutions outside of ILS/discovery platform designers (hereinafter, ILS vendors) and content vendors for the public interface to library holdings. Part I of this article illustrates why libraries and ILS vendors have been unable to develop online public access catalogs (OPACs) and discovery platforms adequate to meet users' needs. Part II describes the type of public interface that is needed to meet these needs and explains why existing efforts are inadequate. Part III argues that SECTs are ideally situated to undertake novel endeavors and better equipped than libraries to provide the necessary technological expertise to develop new platforms, and Part IV outlines concerns about library-SETC partnerships, proposes solutions, and delves into the advantages that each group would gain from working together. This paper is not meant to provide a deep dive into any of these issues but simply to make the case for exploration.

As important as what this paper covers is what it does not cover. It will not analyze the OPAC versus the discovery layer debate,\(^3\) the ineffectiveness of the ILS for accessing libraries' full holdings,\(^4\) the most effective technology available to search multiple library holdings simultaneously,\(^5\) ILS vs. library services platforms,\(^6\) or the benefits or disadvantages of ILS open source software solutions.\(^7\) While related, each of those topics would take

---

\(^3\) Dianne Cmor & Rory Litwin, *Should We Retire the Catalog?*, REFERENCE & USER SERV. Q., Spring 2014, at 213, 213–16.
significant time to cover independently. This paper assumes that the reader is at least superficially familiar with all of these issues.

**Part I: OPAC—A Vehicle Past Its Prime**

Prior to the introduction of the web, data discovery was labor intensive and primarily tackled through careful organization, whether through a classification scheme like Dewey or the Library of Congress, or through controlled vocabulary in indexes, both standalone (e.g., Index to Foreign Legal Periodical) and within publications. These tools are still valuable for researchers today, but it is undeniable that the majority of users find it easier to use blunter tools, like general online search engines. There has long been a debate about “good enough” versus accuracy in research, and while strong arguments can be made for each side, the bottom line is that even the most accurate of tools is useless if it is never used. While it would be an overstatement to say that OPACs are never used, surveys demonstrate that they are far from the preferred way to access information. OPACs and discovery layers may be capable of great accuracy, but they are generally too confusing and too limited for most users to view them as essential research tools.

---


10 See Yongming Wang & Trevor A. Dawes, *The Next Generation Integrated Library System: A Promise Fulfilled,* INFO. TECH. & LIBR., Sept. 2012, at 76, 76–84, stating “...the library automation system, also called the integrated library system, (ILS), has not changed much for the past two decades. It finds
Searching technology and artificial intelligence (AI)\(^{11}\) have advanced rapidly since the development of the web, taking advantage of the wider availability of information and increased computing power for searching and analyzing large data sets. The graphical web's inception was in 1993\(^{12}\); Lynx and Yahoo, among others, provided directory based browsing in 1996; and Google released its search engine in 1998.\(^{13}\) In June 2014, users looked to Google over 40,000 times a second for answers to their basic informational queries.\(^{14}\)

Daily, the number of searches in Google exceeds five billion.\(^{15}\) Admittedly, these searches are not equivalent to an OPAC search. Google searches range from traditional catalog searches (e.g., Google Books) to more general informational questions (e.g., movie times), but it still illustrates the success of Google in attracting searchers at a time when libraries struggle to get users to search library OPACs and discovery layers. Google's great success, added to the continued proliferation of new search engines, should demonstrate to libraries that SETCs have been more effec-

---

\(^{11}\) Artificial intelligence in this paper adopts the definition by Stephen Arnold in Information Today: software that can extrapolate and analyze large quantities of data to perform tasks that once required human subject matter experts. Stephen E. Arnold, *Artificial Intelligence and the Promise of Better Information Access*, INFO. TODAY, Mar. 2015, at 14.


\(^{15}\) Google's daily search statistics, with approximately 5,740,000,000 a day in 2014. Statistic Brain Research Institute, *Google Annual Search Statistics*, STATISTICBRAIN.COM, http://www.statisticbrain.com/google-searches/ (last visited Mar. 4, 2016). Neither Google Books nor Amazon search numbers are available, so this more general Google search number will be used as an imperfect proxy.
tive than the library or ILS sectors in building tools (1) that anticipate users’ search habits and (2) effectively bring together disaggregated information to produce useful information to researchers. The success in the latter category relies on increasingly sophisticated AI to analyze context and relationships to identify similar terms or concepts.

Libraries have taken a different approach to their computerized systems, seeking to front-load the work through cataloging standards and controlled vocabulary. Until recently, there was little to no default relevancy ranking in most OPACs, built upon the theory that libraries should not be biasing users through their displays but instead should allow the user to determine what priority they prefer in the display. Further, even if libraries were inclined to use an algorithm for relevancy ranking, due to its long-standing commitments to privacy, detailed usage information is unlikely to be available for analysis in such an algorithm.

In other words, libraries appear to have searched for and implemented systems designed for the expert librarian while the technology industry has designed systems for the average user. Library OPACs are consistently underused as most of their users are not expert and need more guidance than library systems have been built to provide. ILS vendors, in building their OPACs and discovery layers, have made the same, though understandable, mistakes, since their direct clients are libraries, not end users. Their product specifications are tied to libraries’ demands; as long as libraries are tied to past expectations, ILS vendors will not innovate beyond the library community’s comfort levels. Even when ILS vendors innovate, they are mired in issues such as biased results

and incomplete searches. For example, ILS vendors who are also content providers have incentive and insider knowledge to make their data easily found in their discovery platforms, but it does not stand to reason that the algorithm adopted will be equally successful in mining content from other sources. SETCs, not reliant on libraries, are driven by market forces and a greater understanding of the capabilities of technology and tend to aim towards the masses. They have an uncanny ability to detect uncertainty in searching, and this is reflected in their algorithms and search results. Their search results illustrate their recognition of a title search from a search for a general idea without explicit articulation by the user of his intent. Through this balance of both the library and SETC perspectives, the most effective interface can be built for the world’s users.

**Part II: Public Interface Reimagined**

Companies that have grown and thrived are those that have anticipated their users’ needs and have sought to meet (or exceed) them, not those who try to force their users into a set pattern or practice. It is with this principle in mind that libraries should step away from ILS-vendor-created OPACs. They should instead imagine an interface that is able to search all, or most, library collections worldwide simultaneously and then condenses results

---

down to just a few pages or asks questions relevant to a query to narrow down results. While the approach below may sound similar to FRBR, it is less structured than FRBR’s concept of a national bibliographic record.

For example, searching for the “United States Code” in the title field in WorldCat currently yields 8,660 titles. While catalogers and sophisticated researchers understand the differences in each of these entries, the general public is likely only to be overwhelmed when all they want is access to a current version of the United States Code. Typing the same search into a general search engine today will bring up many more results, but the first two handfuls are the ones most likely to meet the needs of the searcher.

The new imagined interface would hopefully take the advantages of both entities—the AI behind search algorithms in search engines and the careful curation of content by libraries—and display it in a manner customized to the user.

In this interface, if a computer algorithm determines that the current United States Code is what the majority of users want, then that is the single title that the interface would present to the user in response to the query. If there are common alternatives to the most popular option, then those alternatives should show as questions underneath. For example:

---

18 Barbara Tillett, What Is FRBR? A Conceptual Model for the Bibliographic Universe, http://www.loc.gov/cds/downloads/FRBR_PDF (last visited Jan. 28, 2016). Note that there are discovery systems—like OCLC’s—that FRBR-izes its entries, but its current iteration of this confuses the user in displaying specific information tied to an edition for what should be a FRBR-ized record.

19 WorldCat, https://www.worldcat.org/ (last visited Jan. 28, 2016). Please note that WorldCat now does some consolidation of titles using FRBR. For example, if you do a search for a casebook title, you will get a single consolidated entry, along with a link to all editions and versions.

Do you need a different year? If so, which year?
Do you want one of the more recently updated, but unofficial codes, instead?

Answering either question would produce a different set of links for the user. However, if the user clicks on the first link, instead of answering either question, he would then receive the links to the formats of the current *United States Code* available to him or at nearby libraries. AI within the search engine would identify, through multi-point matching, relevant titles and group them together in logical formats, and default to a display in the same language in which the query was typed. The user would not see thousands of results but perhaps just four: digital-authenticated, digital-other, print, and microformat. Clicking on any one of those formats would prompt the interface to narrow the results to the ones nearby or available to this user. Nearby libraries would be determined through GPS (and would include freely available online copies from any location), and preferred libraries would be determined by a user-created profile, where each individual saves which libraries he frequents and to which he has user privileges. This display eliminates one of the most common complaints about OPACs: users getting so mired in trying to determine what differentiates one entry from the other 1,000 that they are unable to make a choice and move to the next level of research.

Expanding further, to materials more broadly desired, picture a grade-schooler searching for *Treasure Island* for a school assignment. She likely does not care what edition or printing she reads. Entering such a search in the proposed interface would combine all library holdings in all languages into a single entry, displayed in the user’s preferred language, again followed by common questions:

Do you want an abridged version? (With a link from “abridged” to a definition of the term)
Are you looking for an audio version instead?
Are you looking for a movie instead?

If she wants it in print, and has set up her profile to say that she is just interested versions in English, she would then be shown
the copy in the library nearest to her that has the item available. If the relevant library system permits external, automated actions, it would place a hold on the item for her. She would have the option to click on a button that says “show all copies available to me” if she preferred a different location. If she had clicked on the online format instead, and if three of her preferred libraries had subscriptions to the same online version, the interface would not list all, asking her to choose, but would simply send her to the first location with an available copy. Where the item is not available at any library open to her, she would be provided with links to vendors of the title in case she wants to purchase the item instead. Much as she should be able to customize the library display, she should also be able to save her preferences for vendor displays. For example, if she always prefers the lowest priced matching item, regardless of vendor, she would save that preference and that is what she would be shown first under buying options.

Admittedly, most searches are more complicated than the examples above but would still benefit from the AI that synthesizes and applies past aggregated search history to the current query. For example, a query for “Thomas Jefferson” might result in no titles being provided but simply questions informed by past searches and selections, such as:

“Are you looking for works ABOUT Thomas Jefferson?”
“Are you looking for works BY Thomas Jefferson?”

Or a query on “microbiology” could start by providing a list of the top microbiology topics/subtopics so that the user could hone the search further.

More sophisticated users could elect to filter results or lift consolidation at any point in time during the search. If the user chooses to “unfilter” results at the earliest stage, he would see all results without deduplication or could set his preferences to deduplicate only on given fields (e.g., only if title, author, and imprint fields match in their entireties). If the user instead chooses to lift the filter at a later stage, such as after clicking on the first
displayed, consolidated title, the interface would show only the materials that had been consolidated under that link. For example, it would not show video titles if default answer was a printed book. The user would then be able to search within results or filter further.

For the most sophisticated researcher, the one who knows what she wants, a customized profile would further allow her default search to present the results the way she prefers. She could build a profile that notes that she wants all results and no relevancy ranking by the system, specify that she wants materials in any of six different languages, indicate a preference for online sources over others when a title is available in multiple formats, and/or wants to exclude any fictional materials. Each noted preference would modify how results would be displayed to her each time she makes a query.

The level of control allowed within a profile would be granular, including (1) preference in formats (prefer audio over other formats), (2) libraries that she wants to search, (3) libraries to which she has borrowing privileges (used only when acquisition of title is necessary), and (4) any field (e.g., date, publisher, language) limitations. A profile would even permit for multiple sub-profiles, so that a scholar working in several different disciplines might have different specifications saved for each. In such an instance, she would have one default profile but could switch to any of the others during a search session.

While all of the examples above are simplistic, the same general approach can be adopted for comprehensive discovery services of full-text documents as well as bibliographic records. The challenges of discovery, though, include issues such as paywalls and local hosting, so will not be discussed in detail here.

Functionality of the system could extend beyond mere searching. For example, as the system matures and technologies advance, AI might even make possible an optional simulated reference interview prior to displaying results. Simulated reference would rely on the same factors used currently by search engines, analysis of big data and the most common features within ques-
tions. It might ask, “What do you need this source for?” Whether the user answers “book report” or “check citation” might change its recommendations or help to shape the results display to the item that it thinks most relevant to this particular user for a given need. Siri\footnote{Siri at a Glance, APPLE, https://www.apple.com/ios/siri/ (last visited Jan. 28, 2016).} or Cortana\footnote{Meet Cortana, WINDOWS PHONE, http://www.windowsphone.com/en-us/how-to/wp8/cortana/meet-cortana (last visited Jan. 28, 2016).} are the precursors of this type of superficial reference function.

A staff-only interface could also be available for the back-end functions that are necessary for library operations. This would likely work much as Wikipedia\footnote{Wikipedia contributors, WIKIPEDIA.ORG, https://en.wikipedia.org/wiki/Wikipedia:About#Wikipedia_contributors (last visited Jan. 28, 2016).} does, with different levels of editors and different rights associated with each level. At the highest administrative level, editors would have the ability to enter copyright information to enable “automated” rights control, where digital materials could be released en masse as their copyright term passes. Further, the system would encourage user interaction and enhancement of records by allowing editors (e.g., librarians) to link their records to other records where association is natural (e.g., new edition of a title). With user interaction, gaming and inaccurate tagging should be anticipated, so abuse or inaccurate information should be easily reportable to an administrator from any page so that an entry or edits to it can be reviewed by expert or locked if necessary.

**Part III: Why Library-SETC Is the Right Direction**

Libraries that agree that the interface described above is a good goal may still argue that, with this idea in mind: why do we not build such an interface ourselves or ask our usual ILS vendors to build such functionality into their existing systems? My answer to this relies on my background as an academic law librarian, and
therefore, the remarks below may not be accurate as applies to non-academic or non-law libraries.

First, risk is not something tolerated well by academic libraries or existing ILS vendors, and for good reason. As tuition costs rise and employment opportunities shrink, good stewardship of tuition monies means investment in ideas that have a high likelihood of success. When economies are tight, there is little room for experimentation, where failure is a very possible outcome. Students pay tuition for an education and for employment prospects; they are unlikely to tolerate long-term, high-cost investments in tools that are untested. Academic institutions have committees, focus groups, and boards all involved in decision-making, and this makes change slow and cumbersome.

Second, those holding academic positions already have pre-defined jobs. The ILS librarian may not have sufficient time or expertise to both support the existing ILS and develop a new one. Even if she did have the time, the school might not be as willing to invest the funds in this activity over others where there is a clearly defined need and an easily reached solution (e.g., ILS-vendor-supplied discovery layer).

Third, technology talent is more easily obtained in the technology sphere, as the salaries and benefits provided in the for-profit world far exceed what the average academic institution would be able to offer. Technology entities are generally more nimble and more willing to take risks, as their profits and market share depend on their ability to beat out the competition in the construction of the next, new, better product. Traditional technologies or standards to mine catalogs (e.g., Z39.50, APIs) are slow and cumbersome, and even newer techniques (e.g., creating a local database) may be less efficient than direct access. Those more familiar with technology in a wider context are more likely to be able to come up with the most effective ways to mine, crawl, and/or cache library data to speed searching.
Fourth, libraries have already made these types of attempts repeatedly over the years, with Evergreen\(^{24}\) perhaps as the most successful effort. While they have been effective in back-end record management, their public interfaces largely duplicate the concerns about vendor-provided ones: dense results that are cannot be intuitively understood and slow responses. Some interfaces, like BYU’s Blacklight and ApacheSolr, have been more successful than others, finding manual ways (e.g., tagging) to highlight the most relevant source in a search. However, if you look at their search result pages, they are still cluttered with elements that the average user does not understand or use. Kuali Ole is the most recent entrant in the market and may be the project that breaks through past failures, but until that time, the same criticisms leveled at vendor-developed interfaces apply to library-supplied ones.

To flip the question posed at the beginning of this section on its head, one might also ask why SETCs, with superior design skills and greater funds, need libraries as partners. After all, there has been some advancement in this area, as shown by Biblio Commons’ growth in public libraries.\(^{25}\) The answer to that question can be found in the history of Google Books. Google was savvy enough to partner with libraries for quality materials. Library partners were selected carefully and their collections had long-term value. Similarly, by restricting searches to library holdings, SETCs would weed out the noise of unedited materials that to drowns out relevant search results. The other lesson learned from the Google Books project is that experts are key to providing value-added tools. Google Books built an incredible store of digital information without consideration of how its use might differ from the use of other Web resources. In its earliest days, this blindspot came into glaring focus. Researchers almost immediately detailed puzzling problems such as the failure to link related items (e.g., volumes


from the same set of books) or mis-recording basic bibliographic information (e.g., publication dates). Librarians would be able to anticipate, and possibly counter, these types of problems at the outset of any project.

SETCs have the technology and the expertise to build just about anything, but when it comes to information, they are unfamiliar with core use concepts. By partnering with those who answer reference questions every day, private entities can build more responsive tools and consequently gain more market share.

In a library-SETC partnership, each additional partner creates a more useful database. SETCs provide the technological expertise and the logic crafted over decades of examining user search strategies in trillions of searches. Through this expertise, the interface can use both unofficial (user provided) and official (library-provided) information to determine relationships between items. Libraries contribute to this endeavor by making their records, and all associated coding, available to these vendors. Through mining bibliographic records provided by libraries (e.g. “reprint” in the 500 field), an AI can identify relationships within items that it may not be able to easily do otherwise. For example, the “reprint” term in 500 fields would tell the interface that this is a later printing of an earlier issued title; standard numbers (ISBN, OCLC) in the 001 field would allow the interface to recognize identical items even where individual libraries may have made edits to their local records (e.g., added subject fields) during copy cataloging. Last, but not least, where full-text data stores exist, it can compare and contrast full-text documents to determine similarity (e.g., if only the publication date is different, then it would be considered the same title for the average user).

The possibilities for partnership do not end there. Other entities (e.g., vendors) could also serve as useful collaborators, to the extent their data is open to search. For example, a search engine could use reviewers’ notes on commercial (e.g., Amazon) or resale

---

(e.g., usedaddall.com) sites to identify where something is a reissue and not a new title. This identification would help the SETC to associate related titles with each other or "deduplicate" titles that have the same content, even with different publication dates. If these non-SETC vendors decide to make their data (e.g., reviews, notes) accessible along with a field manual for their data structures, this would facilitate any integration with a search engine.

**Part IV: Bringing Together Disparate Interests**

With library-SETC partnerships as a promising solution to the question of the user interface, we now reach the more difficult questions about the possible success of such an enterprise. If SETC entities are focused on profit, and libraries are concerned about access to information regardless of income streams, how do we make a project attractive to both sides?

In this, libraries may be the easier sell, at least initially. Access to better technologies and more sophisticated programmers, without significant charges on their established budgets? There seem to be few downsides. Their existing systems could continue to operate, so if the partnership fails, they will have lost little but time. If it succeeds, they will have a popular universal interface and greater exposure and use of their materials. Concerns would rise upon use, and these will be discussed in a later section.

Selling this to SETCs is more challenging. After all, other tools already exist (e.g., WorldCat, DPLA) and are widely adopted or funded, so prospective SETC partners may not see the benefit in developing a more sophisticated tool. Further, what do they gain from such a development? The first reservation could be overcome through user surveys, asking if existing tools are sufficient for researchers. If not, why not? If great demand exists, SETCs may be persuaded that development of a better tool will help them reach new markets. The second is trickier, as there are monetize-able elements of such a system, but the underlying mechanisms required
to monetize may be objectionable enough to dissuade libraries from signing onto such a partnership.

What are some of these monetize-able elements?

- Selling general trends in the popularity of topics, titles, and so forth to publishers looking for solid data to help them determine where they should be seeking new content or when they should reissue a title.
- Sharing trends in relationships of searches with publishers for marketing purposes.
- Where a title may not be available at any local library, linking to online stores where the user can purchase a copy. This function is one already established in many public library OPACs and Google Books.
- Charging users for a virtual library of titles; the owner would own none, but could link from the library to the title where access is instantaneous. In concept, this is similar to Amazon Kindle Unlimited, but access here would be broader as there would be access to multiple libraries simultaneously. Finding an available title among many libraries when a given user wants to read it should be possible in most cases, if current fiction (e.g., fiction published within the last two years) is excluded. Profit could be shared by the SETC and each lending library.

The remainder of this paper will discuss possible roadblocks to collaboration and possible solutions. In order for a partnership to be viable, each partner must benefit from the arrangement, either directly or indirectly. The arguments against library-SETC partnerships include the commercialization of information, loss of privacy, lack of sustainability, among others, and each of these must be addressed in turn.

**Sustainability**

Libraries have an obligation to their users and society to ensure continued access to information and the tools necessary to effec-
tively reach this information. One of the fears of relying on a SETC is the ability of the company to withdraw its support or to discontinue a very popular product that may have no or declining commercial appeal (e.g., Google Reader). One way to counter this is enter into an agreement at the outset that the product produced by the partnership belongs to both, that the code created by either is available to both, along with documentation on development at each stage, and that upon one entity’s decision to discontinue use, the product will be transferred in its entirety (interface, programming, domain name, historical documents related to the building of the tool) to the continuing entity.

Control and Governance

In order to assure a peaceful and productive partnership, control of development and implementation decisions should be determined in advance. What not likely to succeed is a large board with every participating entity having a voice or vote. In order to advance, decisions must be made quickly and efficiently, with recognition by all members that it is impossible to meet everyone’s needs. The baseline expectation, therefore, should not be one where there is unanimous approval, but rather that each iteration of the product will meet additional needs.

While space for comments should always be available and hot-washes after implementation of any change should be undertaken, limited voting rights will ensure nimbleness when needed. The final determiner of critical user needs should rest in the hands of a few, diverse libraries, as they are the ones with broad and in-depth access to users. Evergreen’s governance structure serves as an example of how this can be done.27 On the flip side, once critical needs are determined, then SETCs should have the final say on the technological design to meet the need, as they are the most knowledgeable about technology, the competition, and general user

---

needs. With this division of labor, both groups should be able to work harmoniously. SETCs would still be able to experiment beyond critical user needs but libraries would be assured of inclusion of any essential features.

**Standards**

Technology standards, not cataloging standards, should rule in this partnership. As discussed above, libraries currently draft and adopt standards to ensure that we are able to meet the specific needs of even the most expert of researchers. Whatever people want to find, by whatever features (format, title, publisher), we want to support it and ensure that they receive only what want.

The difficulty is that most users are not expert users. They want an answer and shy away from paging through screens and screens of listings to find what they want themselves. Whether or not librarians approve of this method of searching, it is a reality. The providers of library search technologies have not adequately responded. It is the private entities or developers who have been most responsive. The answers that users receive from search engines may not be as nuanced, expert, or even accurate as is possible through careful research, but not everyone needs the premier treatise on a given topic to be satisfied with the answer.

Libraries want reliable searches, and the rules we have adopted over the years—AACR2, LC Subject headings—are strict and require continued maintenance.\(^{28}\) The process is labor intensive and costly, and not all libraries have the resources to maintain this standard. The purpose behind this control is to ensure consistent results; users should be able to find materials about American Indians whether they use that term or "Native Americans" or "Indians of North America." SETCs have addressed the issue of controlled terms in another manner entirely, by analyzing the terms that searchers use interchangeably for each other. This latter

approach is a more sustainable one, if less accurate, as it does not rely on each library to update its records to make them findable. By dropping the requirement for strict adhesion to some standards (like subject headings), a search interface can still make titles findable despite imperfect records at the local level. In the type of system described in this paper, the aggregated nature of the search would also help local libraries in enhancing accessibility. For example, if you have two libraries holding the same book with the ISBN noted and matching in each, but one library has kept up with authority control and another library has not, the system, recognizing the titles are the same, would allow a user to find both records whether they used the old or new subject. In most instances, this type of recognition or extrapolated cataloging will not be possible without matching multiple fields to identify similar records.

As protocols are standardized and as ILSes start storing information in non-proprietary structures (e.g., SQL), querying them becomes simpler. AI should make it possible for a computer to determine related words without strict authority control. Recent developments in cataloging, such as RDA, BibFrame, or the Open Access Initiative, need not be abandoned, as these each contemplate enhancements in records that will assist in searching even if not adopted across all libraries. For example, BibFrame anticipates assigning each type of text field in a bibliographic record with a Universal Identification Number (UID) and allowing relations to be built from the mapping of a UID to records that may not conform exactly to one another but are, in essence, the same resource. In that structure, the consolidation of titles for the initial results display will be enhanced.

Simplicity

Where SETCs have excelled is in providing the user what they want as much as what it is that they need, by extrapolating from

---

general search patterns. Search engines "anticipate" what you want by showing what others with similar interests have used or bought; they consider the sequence of your queries to determine if you are narrowing or broadening your focus. When a user's search or expectation deviates from the general public's, they are more likely to get imperfect results, but for the majority of users, this tactic is adequate.

As described in the first section, the core structure of any interface should be one that defaults to the easiest search interface for the average user. Later enhancements would focus more on the expert user.

For the average user, the simpler the interface and the cleaner the results, the more likely the tool is to be used. For the default user, then, as noted above, the results would be displayed in the same language in which the search was typed, all reprints (regardless of language or publication year) would be aggregated, and the user would see only a handful of results, followed, if appropriate, by a series of questions that will prompt the user to consider whether or not the items displayed meet his need. It should recognize common terms or abbreviations like "second edition" or "2d," and adjust the results accordingly. As results are displayed, the search engine could follow with the usual set of questions or simply provide an explanation of the search results (e.g., "2d was interpreted as a desire to see only the second edition of this title. If you want to search for the terms themselves instead, please click here.")

Single Sign-on and Privacy

A customizable product is much more user-friendly than a neutral one. Unfortunately, customization comes at a cost, and that cost is a potential loss of privacy. Single sign-on technologies are now widespread (e.g., shibboleth), but most of them operate by keeping information in the cloud, accessible not only by the owner but by the technology provider. In terms of monetization, there would be great utility for SETCs to have access to and be able to sell this information, but the idea is such an anathema to the core principle
of library privacy that it is unlikely that any partnership could succeed if this became a non-negotiable requirement for SETCs.

Technologically, there is a solution that would allow for customization and still protect privacy, though it reduces efficiency for the user. Several applications (e.g., Dashlane) provide customized services and the saving of personal information while still protecting the user from any corporate intrusions. Instead of saving information in the cloud or on private servers, any personal information is stored only locally and encrypted. It would be recalled each time a search is executed but would not be saved by the providing entity.

This approach is unreliable for two reasons. First, even if the information is not stored locally, if it is used to generate results each time a search is run, the search engine has the ability to capture the criteria at the time of search. It may not be able to link the search to an individual, but it would still be able to identify patterns of people who have similar preferences.

Second, there is debate among scholars as to how much information is needed to identify a person; recent scholarship argues that you need only three pieces of information to be able to reliably tie a person to them.\(^{30}\) If true, even if the SETC did not retain individual information, if subpoenaed for general information on search patterns at given times, the SETC (or law enforcement) may be able to identify the person who executed the search.

The further risk of a library-SETC partnership is that the SETC itself would be able to mine and use the data gathered from users for purposes not intended in the granting of initial access. As they are independent entities, not subject to review by outsiders, no matter what guarantees they make, there is no certainty of privacy. Further, a SETC might more easily give into federal authorities when the users that are impacted are not the majority of their users.

There is no ideal solution to the privacy/utility conflict, whether in libraries or elsewhere online. Risks associated with

using external service providers have been well documented, but as illustrated by security breaches at the United States’ Office of Personnel Management, the University of Maryland, Indiana University, and many others, maintaining data in-house does not guarantee any greater protection of privacy. Further, many libraries have already made the move to SaaS (software as a service) for their ILSes, where storage is in the cloud and security is provided by a vendor. Libraries and users will need to determine what balance, if any, they are willing to strike to make their collections more usable.

**Other Concerns with Monetization**

Information in the aggregate can be very useful, and as noted above, an aspect of the project that may be appealing to SETCs is the ability to analyze user patterns and interests, and then selling the analysis to content vendors. While this outcome could be beneficial to the public—more titles in areas of high interest—it could also have a deleterious effect, stifling scholarship in currently unpopular areas. It could also result in a glut of poor-quality publications in popular areas just to take advantage of current market conditions. Whether this would be considered enough of a conflict of interest to libraries to scratch such a project is not known.

A related concern is the skewing of information. Search results are not necessarily complete or unbiased. Not only do they reflect the general biases and misperceptions of past searches within the search algorithm, but search engines have been shown to alter results based on the searcher or to disproportionately weigh certain results for profit. Libraries have aimed to be a neutral provider of

---

information, to allow each researcher to conduct an impartial inquiry on a subject without undue influence. Search providers have no such mission, so any concerns about preferential treatment—even if that preferential treatment were in the favor of certain libraries—would need to be discussed and understood in advance.

**Other Library Concerns**

Library staff occasionally debate whether schools might use technology as an excuse to eliminate libraries and their employees. After all, reference questions—though certainly not the same quantity of them—were once answered by reference librarians, and the types of in-person reference questions at libraries has changed in consequence. However, it is important to note that the number of reference queries has not notably decreased since search engines have made their appearance. The types of questions may have changed, but the nature of the shift only demonstrates how necessary the librarian is to in-depth research. Librarians are not only receiving more complex questions at the reference desk, but they have seen an increasing demand for research instruction. Increased, better exposure to libraries’ collections should only contribute to this growth.

In addition, this collaboration should result in lower library costs, because the current costs of ILSes are in part driven by research and design for improving the public interface. With every publisher developing its own proprietary interface, libraries pay

---


multiple times for the same function (e.g., searching). If it becomes possible to develop a set of international standards which vendors would be willing to adopt, libraries and publishers alike could eliminate these interface costs, licensing to libraries only the data. There would be no need then for libraries to assess interfaces as part of their collection development decisions; instead, they would select only content, as all content would be mined by a single interface.

Conclusion

It should be evident that this paper advocates for an approach that is unusual for libraries, as libraries tend to adopt expensive platforms, standards, and models only once they have been fully tested against our existing tools and collections. It is time to abandon that approach. In building for the future, libraries cannot forever look to the past; they must be aware of it and respect it, and they must eventually be able to accommodate it, but they should not build specifically to it if they want true advancement. They must be willing to launch intentionally an incomplete product and seek improvements or additional features over time. The partnership outlined is not perfect solution and is not aiming to be one. However, I believe it is a better solution than what we face now and is one that speaks to the generations to come.