Submitted on: 09.10.2015

What's Driving Discovery Systems? The Case for Standards

Heather Lea Moulaison
iSchool, University of Missouri, Columbia, USA
E-mail address: moulaisonhe@missouri.edu

Angela Kroeger
Criss Library, University of Nebraska at Omaha, Omaha, USA.
E-mail address: akroeger@unomaha.edu

Edward M. Corrado
University Libraries, University of Alabama, Tuscaloosa, USA
E-mail address: emcorrado@ua.edu

Abstract:

In efforts to offer the best access possible to users, libraries are increasingly interested in systems that better respond to user needs. Discovery systems provide a single-search interface to all library resources through a single index of content. A number of black box issues plague discovery at present; these could be remedied through the application of standards. Issues are related to the content that is indexed, the metadata that is provided, and the algorithms used to provide a list of relevant results. The profession should consider the adoption of the Open Discovery Initiative as a first step in applying standards to this sector. The proprietary algorithms in use are necessary, but vendors and content providers at the same time have an obligation to follow standards as a way of best addressing the black box issues that have arisen.

Keywords: discovery systems, standards, Open Discovery Initiative (ODI), metadata, access

In efforts to offer the best access possible to users, libraries are increasingly interested in systems that better respond to user needs. Complaints against web-based online library catalogs and their patron-facing interfaces have been documented since the 1990s (c.f. Borgman, 1996, etc.). Next generation catalogs that emerged after the year 2000 were much more user-friendly, but they continued to segregate library catalog data in its own silo, storing it separately from library data housed in other repositories or systems.

However, with the introduction of discovery systems starting around 2009 (Breeding, 2015b), librarians have been able to offer uniform access to a greatly increased portion of their libraries' materials. The overall goal of these systems is a single search interface providing access to all library resources (Breeding, 2014; Breeding, 2015a). Electronic library materials are housed in a variety of systems (institutional repositories, course management systems, archival finding aids, etc.), some of which are proprietary. These systems do not use the same metadata, making search more difficult for the discovery systems and providing inferior results to patrons. As discovery systems and other new technologies emerge and are adopted, the development of shared standards is necessary to ensure consistent access to all relevant library content including local digital resources and other resources not specifically linked to the discovery system provider.

In this paper, we first investigate the discovery system, defining its position and describing its upper-level workings. Next, we review the literature on discovery systems and the adoption of relevant standards before we look more deeply at the discovery system as a black box system. Finally, we conclude with a discussion and recommendations for standardizing discovery systems and their interactions with library content.

Discovery Systems Defined

Discovery systems can be defined as dedicated systems that provide access to a variety of library resources through a single search interface. Breeding (2015a) provides the most comprehensive overview of the current discovery landscape, carefully distinguishing between discovery interfaces, index-based discovery services, local index content, and several other categories of library-based search products. Although the exact definitions and lists of features of discovery services vary across the literature, there is widespread agreement on certain major features, such as the central index (meaning, the content that is indexed by the discovery provider), the single search box, relevancy ranking, and facets (Breeding, 2014; Chickering & Yang, 2014; Hoeppner, 2012; Rowe, 2010).

Working in conjunction with content providers, discovery systems enable access to traditionally siloed library content such as surrogates held in the online public-access catalog (OPAC), ebooks and ebook packages, journal indexes and full-text content, and institutional repository content. Discovery systems re-index the contents of each database, allowing for search to take place in one unified index. Ideally, the discovery system will treat all content equally during the indexing process, and will give the most relevant results from among all the databases in response to patron queries.

Library discovery systems attempt to meet patron needs by providing results from a variety of sources at the same time and allowing for patrons to limit the subsequent results by type of content, subject, or other criteria. Like *discovery*, federated searches "offer simultaneous search of diverse library resources that include the OPAC and database" (Wang & Mi, 2012 p. 231). *Discovery*, however, differs from federated searching because federated searching searches in each silo simultaneously (Wang & Mi, 2012), while discovery systems are "index-based" (Breeding, 2015b, p. 29). Indeed, with discovery systems, patrons need only assess whether the results are relevant and if the search needs to be refined in some way.

Discovery systems can be used in different kinds of libraries and their software can be created by libraries or by vendors. Discovery systems used in academic libraries include Arena (Axiell), EBSCO Discovery Service (EDS) (EBSCO), Primo (Ex Libris), Summon (ProQuest), WorldCat Local (OCLC), and others. Systems used in public libraries include

BiblioCore (BiblioCommons) and AquaBrowser Library (ProQuest) (Breeding, 2014; c.f. Spiteri & Tarulli, 2012). Some discovery systems are open source (e.g. VuFind) (Breeding, 2014), but the majority discussed currently in the literature and in this article are part of a suite of library software offerings made available through large library content and software vendors.

Discovery systems index content that is available to a library's patrons through databases owned and subscribed to by the library. Breeding (2014) notes some of the partnerships that have formed between discovery service providers and content providers. For example, EBSCO Discovery Service (EDS) has announced formal partnerships with more than 30 integrated library system (ILS) vendors. Ideally, these sorts of partnerships will ensure that library data housed in the ILS will be effectively searched through the discovery system. What remains unclear, however, is how content comes to be indexed, and subsequently retrievable, in the discovery system. At present, there are no standards to guide this process.

Black Box Issues

One of the criticisms of discovery systems is that they are basically a black box to librarians and to library patrons. Bruce Heterick, former chair of the United States National Information Standards Organization (NISO)'s Board of Directors, has been quoted as saying, "There is a certain 'black box' atmosphere out there at the moment which is not in the collective best interests of the community" (Kelley, 2012, p. 37). It is often unclear what content is included, how that content is indexed, and how the discovery system determines the order the various search results are displayed. Vendors of discovery systems have legitimate reasons for the black box approach, including the need to maintain their competitive advantage, as well as the difficulty in maintaining a transparent and useable list of the vast set of resources being indexed. This challenge is amplified because of the various degrees of quality and completeness of the metadata that discovery system providers receive from disparate sources. Nonetheless it is a concern for librarians, library patrons, and content providers alike.

Even when the discovery system vendors are vigilant and attempt to do their best to be open about of what is being indexed, the sheer enormity of what is indexed makes it very difficult to know what exactly is included. Librarians may be "uncomfortable with the uncertainty of coverage in a 'black box' discovery tool" (Somerville, 2013, p. 238). As ProQuest's John Law described it, the information about coverage may be available to librarians, but that does not mean it is practical to use (Kelley, 2012). This, combined with different ways of reporting what is being indexed by the various discovery service providers, makes it very difficult for librarians to compare discovery layers based on the content being indexed.

Black box issues do not just relate to librarians not knowing what is being indexed. There are also difficulties in knowing what metadata is being indexed for various content items. In some cases, the discovery system may have access to the full text of an article, but in other cases, it only has basic citation information such as author, article title, and journal title. Sometimes the discovery system has subject information, keywords, or abstracts as well. The metadata provided to the discovery system vendor will also vary in quality. Unsatisfactory and unpredictable results for some items can result from the uneven availability of metadata.

For competitive reasons, discovery system providers have a clear interest in protecting their proprietary relevancy ranking formulas since the relevancy ranking mechanisms could be a key differentiator of discovery systems. Breeding (2015b) asserts, "Vendors, more than ever before, face formidable competition to deliver products that deliver true innovation and

financial value" (p. 29). Librarians, however, need to have enough information about how relevancy works in their discovery system so that they can help library patrons search as efficiently and effectively as possible. This is particularly important because patrons, especially students, have a "tendency to rely only on the first page of search results and to trust the relevancy rankings of a given search engine, mak[ing] the default settings of these search systems critically important" (Asher, Duke, & Wilson, 2013, p. 477). When systems are more transparent, librarians can provide information literacy instruction that teaches students how to most effectively utilize the discovery service. It is often the case that libraries have only a limited ability to adjust the relevancy to weight certain data fields or resources that they believe are the most useful for their patrons, and sometimes libraries are unable to adjust relevancy factors at all. Thus library patrons (and librarians) must know how to use the discovery service effectively as provided, without being specifically aware of the inner workings of the black box.

Another reason that transparency is important is that some librarians and content providers are concerned that a discovery service provider that is also a content provider might rank its own content more prominently. This is unlikely to be deliberate because, as Tim Collins who is now president of EBSCO has pointed out, if it was found that a discovery service provider did this intentionally, "it would be the single most obvious way for us to alienate all content partners" (Kelley, 2012, p. 38). However, because the discovery service providers who are also content providers have more knowledge of their data and are able to make sure that they supply all of the appropriate metadata to the discovery service, it still may happen unintentionally. When the ranking methods become transparent enough to librarians and to all content providers, making discovery system less of a black box, these concerns will be mostly eliminated.

Abstracting and Indexing Providers

Relevancy ranking is also one of the reasons why some abstracting and indexing (A&I) providers have been skeptical about providing their data to discovery systems providers. Content providers have a vested interest in their data being used and discoverable. If the relevancy ranking algorithm does not adequately, in their view, retrieve the data they provide, then content providers, especially those that provide A&I, have legitimate concerns that their products may be undervalued by librarians. In times of limited budgets, these A&I resources may not be renewed if they are not properly valued. Harvard University's Laura Morse explains that "undercounting of usage could cause a library to cancel a needed A&I service, as it may not be aware of the use of the metadata consumed, albeit invisibly, by the researcher" (Kelley, 2012, p. 39).

Related to this is the tricky question of usage statistics. A&I providers need to have usage data to provide to librarians demonstrating their content is being utilized inside of the black box. Because all discovery systems record statistics in their own propitiatory way, it is not clear to some A&I providers or librarians that the usage statistics adequately reflect the value that A&I resources provide to library patrons. When A&I providers hold back content from being included within discovery systems, no one wins. Since many librarians are prominently placing their discovery services on their libraries' home pages, it is likely many patrons will never encounter the lists of resources that include the A&I resources to attempt to search them. Without the rich content that the A&I resources can provide to the discovery system, the discovery system is not as comprehensive as it would otherwise be, and the patrons may not find relevant resources that they otherwise would have discovered.

Recall and Precision

The concerns of A&I providers are not unfounded. It very likely is true that a discovery system cannot search an A&I provider's content as effectively as the A&I service's native interface can. Index-based discovery systems cover a wide breadth of resources, but they may not cover those resources to the same depth as the old siloed systems. This applies not only to A&I content, but to most other electronic content as well. For patrons, discovery increases recall at the expense of precision. Breeding (2015a) notes that discovery services offer less precision than the old online catalogs, and that known-item searching can be problematic in discovery services. This is a systemic problem that may be unavoidable, because discovery interfaces draw together resources built with different metadata structures and different native search interfaces in mind, and those search techniques and structures "do not exist in the broader universe of electronic resources" (Breeding, 2015a, p. 30). In other words, to enable searching across diverse resources, discovery services must search across the lowest common denominator. This is perfectly fine for the majority of users, who will find something useful through a high-recall, low-precision search. Casting a wide net will catch something. However, specialized, high-level researchers may need the extremely precise, laser-like focus that presently can only come from the database provider's native index that is optimized for their particular set of resources.

Professional Practice and the Need for Standards

Research into discovery systems has been ongoing since their inception (Ellero, 2013; Thomsett-Scott & Reese, 2012). Assessment of discovery systems has as well. Breeding (2015a) succinctly touches on nearly every aspect of this expanding field, including issues, standards, and recommended practices; concerns related to proprietary content and open access content; the integration of discovery services with resource management systems (inclusive of integrated library systems and library services platforms) and with learning management systems; linked data as a future possibility, yet unrealized; barriers to comprehensive index coverage, with specific attention to the difficulty of including content from A&I services; and finally the recommendations and future promise of the Open Discovery Initiative (ODI) (Breeding, 2015a), discussed further below.

In addition to the ODI, Breeding (2015a) notes several other relevant standards and recommended practices, including Recommended Practices: Discovery Services from NFAIS, Discovery: A Metadata Ecology for UK Education and Research, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), ResourceSync (ANSI/NISO Z39.99-2014), Knowledge Bases and Related Tools (KBART) (NISO RP-9-2014), NISO Recommended Practice on Access License and Indicators (NISO RP-22-2015), OpenURL standard (ANSI/NISO Z39.88), and Digital Library Federation ILS Discovery Interface Task Group (ILS-DI).

Present Barriers to Access

Notable black box problems mentioned above that are potential barriers to access include inconsistent or incomplete metadata (Hoeppner, 2012; Kabashi, Peterson, & Prather, 2014; Somerville & Conrad, 2014); proprietary relevancy ranking algorithms (Breeding, 2015a; Hoeppner, 2012); inconsistent coverage of open access content (Breeding, 2015a; Hutchens, 2013; Somerville & Conrad, 2014); and the exclusion from the central index of portions of the library's lawfully owned or licensed content, often due to vendor unwillingness to share metadata with competitors (Breeding, 2014; Breeding, 2015a; Chickering & Yang, 2014; Kabashi, Peterson, & Prather, 2014; Somerville & Conrad, 2014; Spezi, Creaser, O'Brien, & Conyers, 2013). These could all potentially be improved with increased standardization.

Relevancy algorithms are worth a special note, because these are, as described above, proprietary in nature and treated by the vendors as trade secrets (Breeding, 2015a). Hoeppner (2012) notes that while each company's relevancy algorithm is tailored to that company's own central index, all relevancy algorithms factor in date of publication, field weighting for search terms, and search term proximity. Breeding (2014) notes keyword placement, usage metrics, and frequency among other factors typically influencing relevancy ranking. Ex Libris Primo Central is notable as the first discovery tool to offer the ability to sort by popularity, which "is considered to be a revolutionary step and a departure from traditional relevancy" (Chickering & Yang, 2014, p. 17).

However, the exact algorithms for relevancy ranking are typically not shared (Kabashi, Peterson, & Prather, 2014, p. 12), presenting that potential barrier to access. The value of non-disclosure agreements and other proprietary protections for vendors is clear, but these "should not be used to avoid . . . transparency" (Open Discovery Initiative Working Group, 2014, p. 22). Despite what vendor representatives might say, the belief that relevancy ranking favors the vendor's own products and/or excludes competitors' content is widespread (Hoeppner, 2012; Spezi, Creaser, O'Brien, & Conyers, 2013).

Standardizing Discovery

The case for the creation of standards for the use of the discovery systems community has been most eloquently put forth by Marshall Breeding and associates through their work with the Open Discovery Initiative (ODI) (http://www.niso.org/workrooms/odi/). In the United States, the National Information Standards Organization (NISO) Open Discovery Initiative (ODI) NISO RP-19-2014 is a recommended practice document addressing the technical aspects of the composition of the central indexes of discovery services (Open Discovery Initiative Working Group, 2014). Recommendations for content providers focus on core metadata elements, enriched content, information disclosure, and data exchange formats. Recommendations for discovery service providers focus on metadata for content listings, compliance with KBART recommended practice, fair linking, file formats, and encoding schema. Further recommendations are made to all stakeholders regarding metrics (Open Discovery Initiative Working Group, 2014). The ODI explicitly lists other initiatives that could potentially pertain to discovery services, including COUNTER, COUNTER Code of Practice for Usage Factors, Digital Library Federation ILS Discovery Interface Task Group (ILS-DI), International Coalition of Library Consortia (ICOLC), JISC Discovery Programme, KBART, Music Discovery Requirements, and NFAIS Recommended Practices: Discovery Services (Open Discovery Initiative Working Group, 2014).

With the ODI's white paper as a guiding document, the international library community has an obligation to continue to seek the standardization of the discovery systems sector, each library in each country in conjunction with vendors with which it works, for the reasons described above.

Discussion and Future Work

Discovery systems contracted through a vendor logically will use different proprietary algorithms in providing responses to patron queries. These algorithms are important, especially given the nature of short queries entered by library users (cf. Moulaison, 2008). Wang and Mi (2012) remind us, "Today's users form their information seeking behaviors by using Google and other Internet search engines" (p. 229), and libraries are doing their best to adapt. Breeding (2015b) tells us that public libraries in particular are eager to adapt to the demands of their patrons and "are ready to upgrade or replace incumbent products with ones

better able to fulfill current realities and expectations" (p. 29). In a single-search box library environment, Georgas (2014) notes that 79.3% of participants used keyword or phrase queries and 65.5% of participants used natural language queries. Only 10.3% of participants used a Boolean operator (the "+") and one used the native AND feature of the interface; none used quotation marks.

Short, search engine-like searches in library databases have little in the way of context about the users and their information needs, and give the library algorithms little to go on in presenting lists of results. Unlike search engines, which have a vast data pool from which to draw in compiling their lists of results, libraries do not. "Vendors of library resources do not track individual IPs and user searches in an attempt to produce (or guess at) more relevant results for specific users" (Georgas, 2013, p. 179). The discovery system algorithm, accordingly, must make sense of a limited query given very little context. We acknowledge the difficulty that this presents for these vendors in the creation of their systems.

Vendors of discovery systems, at the same time, need to understand that when patrons find the best resources through their discovery systems, everyone wins. Just as the Google web search engine produces different results from the Bing search engine, it is understandable that different results in different systems will be ranked more highly according to the algorithm in use by a library's chosen discovery system. The algorithm used, however, must be applied to contents in a way that is equitable and uniform across libraries and across collections. A concerted effort must be undertaken to include metadata for categories of materials not yet universally indexed by all discovery systems, such as local content contained within course management software, content management systems, library websites, archival finding aids, institutional repositories, and so on. Vendors need to be incentivized to produce the best set of results possible while taking into consideration the largest number of resources in a way that is unbiased.

Discovery system vendors are not the only ones who should be encouraged to adhere to standards; content providers and librarians also have a role to play. Content providers and potentially A&I providers must agree to provide necessary metadata and data for their content to be searchable in a variety of discovery systems, adjusting their tracking methods to suit this new environment. At the same time, librarians must work with providers of the disparate systems they use for local content, ensuring those systems' metadata is complete, correct, and can be harvested by the discovery system. To do so will be no small task, but requiring vendors and content providers to adhere to standards is a necessity in the current electronic search environment. Individual libraries spend too much money on content for that content not to be equitably made available through the library's discovery system.

Conclusion

In this paper, we defined and described discovery systems and made note of the black box-related issues they present to libraries and their users. Librarians have assessed these systems, and in the process of evaluating, have made a case for the continued need for standards development in discovery systems. Although the work carried out by the ODI is laudable, that work is far from finished, and requires grass-roots support to make an impact. Libraries around the world must communicate to their vendors and content suppliers the importance of standards in this emerging area of library technology.

Increased adherence to standards would facilitate the gathering of a greater array of diverse resources into common indexes, remove some of the mystery and confusion of unpredictable

relevancy ranking results, and improve consistency of search results from different systems. Standardization benefits libraries by bringing a greater portion of libraries' resources into discovery system indexes. Standardization benefits vendors by lessening the chances of a customer dropping a vendor's product because of incompatibility with the library's other products. Finally, standardization benefits library users by improving their access to the resources they need. Libraries must work with vendors and with other libraries, and likewise vendors must work with libraries and with other vendors, to ensure that the application of standards is mutually beneficial. Embracing the work of the ODI and heeding their recommendations for future development offers a sound path toward these goals.

References

Asher, A. D., Duke, L. M., & Wilson, S. (2013). Paths of discovery: Comparing the search effectiveness of EBSCO Discovery Service, Summon, Google Scholar, and conventional library resources. *College & Research Libraries, 74*(5), 464-488.

Borgman, C. L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science, 47*(7), 493-503.

Breeding, M. (2014, January). Library resource discovery products: Context, library perspectives, and vendor positions. *Library Technology Reports, 50*(1), 5-58.

Breeding, M. (2015a). *The future of library resource discovery: A white paper commissioned by the NISO Discovery to Delivery (D2D) Topic Committee.* Baltimore, MD: National Information Standards Organization. Retrieved from http://www.niso.org/apps/group_public/download.php/14487/future_library_resource_discovery.pdf

Breeding, M. (2015b). Library systems report 2015: Operationalizing innovation. *American Libraries*, 28-41.

Chickering, F. W., & Yang, S. Q. (2014, June). Evaluation and comparison of discovery tools: An update. *Information Technology & Libraries, 33*(2), 5-30.

Ellero, N. P. (2013). Integration or disintegration: Where is discovery headed? *Journal of Library Metadata*, *13*(4), 311-329.

Georgas, H. (2013). Google vs. the library: Student preferences and perceptions when doing research using Google and a federated search tool. *portal: Libraries and the Academy, 13*(2), 165-185.

Georgas, H. (2014). Google vs. the library (Part II): Student search patterns and behaviors when using Google and a federated search tool. *portal: Libraries and the Academy, 14*(4), 503-532.

Hoeppner, A. (2012, April). The ins and outs of evaluating web-scale discovery services: Librarians around the world are trying to learn what WSD services are and how they work. *Computers in Libraries, 32*(3), 6-10.

Hutchens, C. (2013). Open access metadata: Current practices and proposed solutions. *Learned Publishing, 26*(3), 159-165. doi:10.1087/20130302

Kabashi, A., Peterson, C., & Prather, T. (2014). *Discovery services: A white paper for the Texas State Library and Archives Commission.* Austin, TX: Texas State Library & Archives Commission. Retrieved from https://www.tsl.texas.gov/sites/default/files/public/tslac/lot/TSLAC_WP_discovery__final_TSLAC_20140912.pdf

Kelley, M. (2012). Coming into focus: Web-scale discovery services face growing need for best practices. *Library Journal*, *137*(17), 34-40.

Moulaison, H. L. (2008). OPAC queries at a medium-sized academic library: A transaction log analysis. *Library Resources & Technical Services, 52*(4), 230-237.

O'Hara, L. (2012). Collection usage pre- and post-Summon implementation at the University of Manitoba Libraries. *Evidence Based Library and Information Practice, 7*(4), 25-34.

Open Discovery Initiative Working Group. (2014). *Open Discovery Initiative: Promoting transparency in discovery. NISO RP-19-2014.* Baltimore, MD: National Information Standards Organization. ISBN: 978-1-937522-42-1. Retrieved from http://www.niso.org/apps/group_public/download.php/13388/rp-19-2014_ODI.pdf

Rowe, R. (2010). Web-scale discovery: A review of Summon, EBSCO Discovery Service, and WorldCat Local. *The Charleston Advisor, 12*(1), 5-10. doi:10.5260/chara.12.1.5

Somerville, M. M. (2013). Digital age discoverability: A collaborative organizational approach. *Serials Review*, (4), 234-239. doi:10.1016/j.serrev.2013.10.006.

Somerville, M. M., & Conrad, L. Y. (2014). *Collaborative improvements in the discoverability of scholarly content: Accomplishments, aspirations, and opportunities: A SAGE white paper.* Los Angeles, CA: SAGE. Retrieved from http://www.sagepub.com/repository/binaries/pdf/improvementsindiscoverability.pdf

Spezi, V., Creaser, C., O'Brien, A., & Conyers, A. (2013). *Impact of library discovery technologies: A report for UKSG.* London: UKSG Global Forum. Retrieved from http://www.uksg.org/sites/uksg.org/files/UKSG_final_report_16_12_13_by_LISU.pdf

Spiteri, L. F., & Tarulli, L. (2012). Social discovery systems in public libraries: If we build them, will they come? *Library Trends, 61*(1), 132-147.

Thomsett-Scott, B., & Reese, P. E. (2012). Academic libraries and discovery tools: A survey of the literature. *College & Undergraduate Libraries, 19*, 123–143. doi:10.1080/10691316.2012.697009

Wang, Y., & Mi, J. (2012). Searchability and discoverability of library resources: Federated search and beyond. *College & Undergraduate Libraries*, *19*(2-4), 229-245. doi:10.1080/10691316.2012.698944