Annals of Library and Information Studies
Vol. 64, December 2017, pp. 276-279

Short Communication

Improving web scale discovery services

Nikesh Narayanan[a] and Dorothy Furber Byers[b]

[a]Systems and e-Librarian, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, E-mail: nikesh.narayanan@kustar.ac.ae

[b]Emeritus Librarian, University of Cincinnati, United States of America, E-mail: dorothy.byers@uc.edu

This article reviews the current state of web scale discovery (WSD) services and their effectiveness in providing a viable interface for initiating literature searches. Some of the shortcomings are discussed, as well as developments that are under way or necessary in order to improve the concept of single searching. Aspects discussed include indexing, relevance ranking, publication finders, linking mechanisms, and personalization of searches. The relationship between publishers and WSD providers is all-important in improving the end-user experience.

Keywords: Discovery services; web scale discovery

Introduction

The emergence of web scale discovery (WSD) services is considered a radical trend in the library information retrieval arena, and there has been rapid adoption of these services by libraries around the world. In the pre-discovery service era, users depended more on Google because library tools were not able to provide such a search environment. Web scale discovery tools have been touted as an academic Google, because these tools are able to use a single interface to integrate results from a wide range of online sources and emulate a Google-like search experience for users.

Libraries hoped these discovery services would provide an all-in-one solution for the information needs of researchers and would bring researchers back to the library. These services have now existed for over seven years and have passed the early-adopter phase. Hundreds of publications and communications on implementation, comparison, user experience, information literacy, librarian perception, and collection usage related to discovery tools have already been produced[1]. From these studies it is obvious that discovery services have had a positive impact on the search behavior of users, but the use of a discovery service as a starting point for research is inconclusive. A recent study conducted on faculty reveals that the majority of them prefer subject databases as their starting point of research, then Google Scholar and other general-purpose search engines[2]. There might be several reasons for such user behavior patterns, but in general library discovery tools must improve in functionality to attain the goals of being a credible starting point for research and bringing users back to the library. This paper covers the issues with discovery tools and areas where improvements are desired.

Discovery index coverage

A central index is the basis of a discovery service. As searches are made against the central index, the comprehensiveness and quality of information retrieval depends primarily on the coverage and quality of metadata. Currently, metadata gathering and updating are based on an agreement between WSD vendors and content providers (publishers, aggregators) and are accomplished through FTP or similar methods. Even though all the major WSD providers are making good progress in covering the maximum possible resources in their central indexes, there are major drawbacks with regard to coverage and metadata standards. These drawbacks are discussed here.

Content harvesting

Unlike Google Scholar, WSD providers do not harvest content through web crawlers; they depend on content providers to provide new data promptly. There is thus more chance for a content gap between the publisher platform and the discovery service index. Delays may occur because the publisher is late providing data, or because the discovery provider is late indexing the data, or both. A study conducted on content coverage of IEEEXplore in discovery platforms revealed that there is a significant gap between the IEEE platform and WSD providers' central index coverage[3]. The Breeding white paper[4] acknowledges content gaps in discovery services as a persistent issue and calls for content analysis tools to deal with the problem. Even though WSD providers list the covered resources, the lack of an automated mechanism to do a gap analysis of the covered resources remains an issue. In order to address this issue, WSD providers must think about technologies to harvest data directly from publisher databases and to map the data to their central index.

Coverage issues

End users do not have an option to know whether a database, journal or e-book subscribed by their institution is included in the discovery index. Discovery services normally include a publication finder portal (A to Z list of the institution's subscribed resources) to enable users to search and browse for the subscribed resources of their institution. But the publication finder does not indicate whether a particular subscribed journal/book is included in the central index of the discovery service. EBSCO Discovery Service (EDS) publication finder is a step forward in this regard as it shows a "Search within journal" search box for those resources which are covered in the EDS central index, but it still does not provide any hint about the extent of coverage for each resource in the central index.

Abstracting and Indexing (A&I) resource issues

Coverage of A&I resources is another major issue. A&I resources are very significant for researchers. Most well-known A&I service providers do not provide their value-added content to WSD providers, and users are forced to search such resources separately. The only exception is the EBSCO Discovery Service (EDS), which can integrate some third party A&I databases (EBSCO describes it as platform blending) in cases where the institution subscribes to such resources through the EBSCOhost platform. This integration is useful for users, who get the benefit of quality metadata prepared by special subject experts of the A&I databases. To benefit from this integration though, the institution must buy A&I resources through the EBSCOhost platform.

Quality of metadata

Quality of metadata provided by publishing partners is important in discovery services. Lacking or inadequate subject classifications and keywords have a dramatic negative effect on the positioning of content in relevancy rankings in discovery search services. Discovery service providers need to analyze the metadata of each publishing partner and help them to enhance their metadata quality.

Relevance ranking

Breeding's white paper on discovery service states that many librarians characterize the performance of discovery services as unpredictable and erratic in the delivery of search results[4]. Each discovery service develops its own proprietary algorithms, tools and technologies to improve relevancy. None of the discovery service algorithms is open source and libraries have only limited flexibility to tweak relevance algorithms based on their users' requirements. Discovery service providers generally make available only an overview about their ranking algorithms, which is not sufficient for libraries to understand or analyze.

EBSCO Discovery Service is a little more elaborate in describing its general approach to relevance ranking in a public document[5]. EDS gives first priority to subject headings, followed by title of the document, then author-supplied keywords, then abstracts, and least priority to search terms appearing in the full text of the article. This is a good approach but inconsistency in subject indexing provided by different publishers adversely affects this approach. Documents with more subject headings or thicker metadata would definitely get the chance to come on top. For example, as an aggregator EBSCO provides value-added subject headings to its aggregated content and also makes use of its own subject indexes. When such content is included in EDS, these contents have a better chance to get placed on the top. Publishing partners' metadata with lacking or inadequate subject headings and keywords would have a negative effect on positioning content in relevancy-based search results.

ProQuest describes its static and dynamic ranking[6] in a general way in its public document on relevancy, but does not provide a clear view about the weight given to each field as is described in the EBSCO public document. The ProQuest document mentions that it takes into consideration citation counts from Web of Science and other sources. It is indeed a good feature but the drawback is that old articles with more citations come on top compared to newer articles with fewer citations. In order to eliminate this problem, the citation count-based relevance ranking should be a separate option for users and not included in the general relevance ranking algorithm.

Another ranking option is usage-based recommendations. Ex Libris has initiated this by incorporating a feature called ScholarRank[7] to inform relevance ranking factored into associations derived from its bX Recommender service using SFX open URL link resolvers.

Another area where discovery services need to improve is to make search results more comprehensive by harvesting different versions of the same document as Google Scholar does. Primo Central and Worldcat Discovery services are doing better in this regard by incorporating FRBR features (based on entities and relationships for describing information objects) in discovery metadata.

Publication finder (A to Z list) of library resources

The knowledge base (database of publications) is an integral part of discovery services and all major discovery services maintain and update a comprehensive list of journals, books and other databases. Libraries can select and customize their subscribed resources from this knowledge base to set up the institution's publication portal. A publication finder (A to Z) list of the institution's resources serves mainly two purposes. First, it is the basis for limiting the search results to the subscribed resources of the institution. Second, end users can search and browse the list for publication information and link directly to a publisher's portal. Browse and search features vary depending on the WSD provider. They generally include features such as subject and title browsing. EBSCO Discovery Service facilitates one step further by providing a "search within journal" function for those journals which are indexed in discovery service. But none of the discovery services provides a combined search facility within a set of journals or databases selected by the customer from the A to Z list. Such a pre-search limiting facility would be very useful for researchers to find resources from the more relevant sources of their choice. Another enhancement option is to incorporate advanced browsing options within the A to Z resources list by including journal ranking. Ranking features such as the H-index and impact factor would help users to limit to prestigious journals from their subject areas. Discovery service providers could collaborate with journal ranking service providers such as SCOPUS or Journal Citation Reports from Web of Science to provide a ranked list of journals in subject areas. Collaboration with the open access tool SCImago journal ranking is another option for non-subscribers of SCOPUS or Web of Science.

Full text linking mechanisms

Discovery services make use of various linking mechanisms to connect the user to the full text of an article. Such mechanisms include OpenURL-based link resolver software, custom links provided by publishers, DOI, etc. In the initial period of development, discovery services depended entirely on link resolvers. Breeding's report on link resolvers[8] indicates some of the reasons for the failure of link resolvers. Also many small-scale publishers are not using OpenURL standards, and in some cases link to the issue or journal level rather than the article level, which of course frustrates users. Custom linking is comparatively better, where participating publishers provide a direct linking solution to the discovery service, using either a Digital Object Identifier (DOI) or other direct unique identifier. However, linking failure could arise in all these cases as discovery services do not check the actual existence of the full text through an automated mechanism like Google bots. In order to eliminate broken links, discovery service providers must find a solution to check periodically the existence of indexed documents and to remove broken links. This is a tricky issue as discovery providers use publisher-provided data and not data based on crawling as with Google Scholar.

Personalization

Even though web scale discovery solutions provide some limited personalization options, they are not sufficient for an advanced user. Some of the desired personalization options missing in discovery services are:

Creating profiles and limiting the search within favored resources

As of now, none of the discovery services has the facility to create profiles and limit searches within selected resources. Some of them provide subject profile search options but end users do not have any control to include their own wish list in the subject profile. They must depend on their institution's selection. Most of the subject profiling provided by discovery services is constituted by including some relevant databases of a particular subject; individual journal/book profiles are not possible.

User-defined relevancy

Currently discovery service providers employ their own proprietary relevance algorithm. Some discovery service providers allow customers to increase the relevancy of their local resources such as catalog and institutional repositories. Discovery providers can tweak the algorithm, but the subscribing institution has to request again if they need further modifications. It would be desirable for subscribing institutions to be able to modify the ranking algorithm to suit their users.

Conclusion

Web scale discovery services have had a positive impact on user ability to search across multiple databases in the academic realm. Since the advent of WSD, providers and publishers have taken a variety of approaches to improve the user experience. However as this paper discusses, further improvement is needed to make discovery systems the first point of departure for scholars searching the literature. Publishers and WSD providers alike must cooperate to enhance indexing, ranking, limiting, linking, and personalizing. Innovations implemented thus far bode well for further improvements that will lure researchers back to the library.

References

1. Renaville F, Discovery tools, a bibliography, Available at: https://discoverytoolsbibliography.wordpress.com (Accessed 04 Oct 2017).
2. Wolff C, Rod A and Schonfeld R, Ithaka S+R US Faculty Survey 2015, April 2016.
3. Zhu J and Kelley J, Collaborating to reduce content gaps in discovery: What publishers, discovery service providers, and libraries can do to close the gaps, *Science and Technology Libraries,* 34 (4) (2015) 315–328.
4. Breeding M, The future of library resource discovery, *Information Standards Quarterly,* 27 (1) (2015) Available at: http://www.niso.org/sites/default/files/stories/2017-10/NR_Breeding_Discovery_isqv27no1_0.pdf (Acccessed on 09 May 2016).
5. EBSCO, Discovery relevance ranking, Available at: https://www.ebscohost.com/discovery/technology/relevance-ranking (Accessed 23 March 2016).
6. SUMMON, Relevance ranking in the summon service, Available at: http://media2.proquest.com/documents/Summon-RelevanceRanking-Datasheet.pdf (Accessed 23 March 2016).
7. Exlibris, The Primo ScholarRank Technology: Bringing the most relevant results to the top of the list, Available: http://www.xlgps.com/article/244495.html (Accessed 09 May 2016).
8. Breeding M, Knowledge base and link resolver study: General Findings, 2012. Available at: http://www.kb.se/dokument/Knowledgebase_linkresolver_study.pdf (Accessed on 09 May 2016).