Web-scale discovery: Utopian dream or dystopian nightmare (or maybe something in between)?

William Breitbach
Dean of Library Services & Educational Technology
Shasta College

Presented at the California Academic & Research Libraries 2016 Conference
March 31 – April 2, 2016
Costa Mesa, California

Abstract

An early depiction of the ideal discovery platform was imagined by Vannevar Bush in 1945. Eighty years later, information professionals are still struggling with the challenge of creating a single search interface. The most recent rendition of the dream, web-scale discovery, arrived on the scene in 2009 with much promise and exuberance. The apex of excitement peaked around 2012, but the value of such systems remains unclear. After six years of continued development, are these systems delivering the promised value?

Introduction

Libraries are adopting web-scale discovery services with increasing frequency. These tools offer library users a single entry point to the library's print and subscription content through a single pre-aggregated index. Like innovations that have come before, these services are part of the continuing story of library automation, which began long ago. Vannevar Bush's famous article *As We May Think* illustrated an early vision of improved discovery through an automated retrieval system (Bush, 1945). Bush imagined a device called the Memex that would aggregate all personal books, papers, and notes and the content in the Memex would be linked through an associative index. The efficiency of the device at retrieving and linking information together would fuel creativity and have a significant impact on human knowledge creation.

The Memex is one of many visions of library automation. An earlier vision included the Book Wheel. The Book Wheel was designed by Agostino Ramelli and was meant to provide easy access to the large and heavy books of the 1600s. Neither of these devices were ever built, but they both had the goal of improving access to information. In some ways, the goals of these early visions of information systems are not unlike the goals of web-scale discovery. They are all part of the continuing story of information professionals trying to make meaning out of a complex system.

Web-scale discovery has four major goals: a single point of entry into a library's collections; an enhanced user experience with an easy-to-use interface; support of unmediated discovery; exposure to unseen collections; and increased efficiency at discovering content. These are indeed exciting propositions, and if the goals were manifest, they could fuel meaningful discovery of information.

Aaron Tay (2014) suggests that these systems, like other technological innovations, are subject to the Hype Cycle. In the Hype Cycle, an innovation is triggered and followed by early development. Excitement for the technology reaches its apex at the "peak of inflated expectations." This point is followed by a phase of disappointment called the "trough of disillusion" as reality about the prospects for the product set in. This is followed by a phase of development and improvement in the product, during which people begin to identify use cases. The final phase of the cycle is called the "plateau of productivity," in which the product is widely seen as functional and useful. Anecdotally, web-scale discovery reached the peak of inflated expectations around 2012. This is supported by the fact that there were 72 scholarly works published that year (Ellero, 2013), while a total of 53 scholarly articles were noted between 2007 and 2012 (Richardson, 2013). The author has not come across estimation in the number of articles published since, but there was clearly an escalation of interest in web-scale discovery in 2012.

Another heuristic that may be useful in thinking about the excitement around web-scale discovery is irrational exuberance. Irrational exuberance is a term used to explain the stock market bubble of the 1990s, in which the value of tech stocks was inflated. These inflated stock values were followed by a major slump in the market. Whether we are in the slump of disillusionment or have experienced irrational exuberance, web-scale discovery is far from delivering the dream of a highly efficient content discovery platform.

Metadata Matters

As Web 2.0 was approaching maturity, David Weinberger (2007) published a book called *Everything is Miscellaneous*. Weinberger argued that conventions of classifications such as the Library of Congress classification system and the Dewey Decimal System are arbitrary constructs that create categories for physical items. This categorization was necessary when items existed in a physical space, but the conceptual organization breaks down and becomes unnecessary in the digital world. In Weinberger's conception, the text of the content becomes the metadata. Moreover, users will create meaning and fuel discovery by organizing content with folksonomies.

Although Weinberger made some interesting arguments and generated important conversations about discovery, content is not miscellaneous. On the contrary, all information resources have context and that context matters. Traditional information discovery systems contextualize information by two primary and important mechanisms – subject databases and specialized metadata. Web-scale discovery systems break down both of these content contextualization mechanisms without replacing them with a viable alternative.

As web-scale services mix and merge the data in order to normalize it into a single index and interface, the value of the metadata is diminished. Ultimately, the link between discipline specific metadata and content is broken and conceptual boundaries between disciplines are blurred (Breitbach, 2012). In a sense, web-scale discovery systems defeat the important purpose of subject specific databases.

Subject-specific databases like Sociological Abstracts or PsycINFO serve a very important function as they manifest the conversation that is occurring within discourse communities. Discourse communities are communities of shared meaning and practice. They use and understand similar language, have shared rhetorical practices and values. They also value similar types of evidence (Elmborg, 2003). When instructors give undergraduates an assignment, they are implicitly asking them to participate in a discourse community. In fact, the process of earning a degree is a process of becoming familiar with and interacting in a discourse community. Subject specific databases help users search for information within their discourse community. Without these tools, boundaries in the literature between individual disciplines become blurred. While web-scale discovery services do not prevent users from interacting with discourse communities, they do not currently offer enough support for this critical component of information literacy (Breitbach, 2012).

Web-scale discovery services attempt to mitigate some of these problems by creating faceted browsing interfaces. These facets can cover broad subject areas as well as specific metadata descriptors. That said, a cursory exploration of these metadata facets shows that they are deeply flawed as a product of the data normalization process that forces content records into a single index. To compensate for the lack of a robust metadata infrastructure, it appears that web-scale discovery services have created algorithms that favor recall over precision. This prevents any given search from missing important content at the expense of including massive results lists. The effect is that even the most naïve search strategies will produce a list of results. This encourages users to select items that are only tangentially related to the initial topic of interest and reinforces poor search strategies. In these cases, it would be better for the index to display zero results, as seeing no results would either force the user to re-think their search strategy or ask someone for help.

Most users access library resources because they have been asked to by an instructor. If one assumes that, in most cases, users are looking for l0 or fewer articles on a given subject, then how do users benefit from being provided thousands of articles on a wide variety of topics from a single query? The present author argues that users do not benefit in such a scenario. Users will likely ignore content beyond the first page or two and may end up selecting content that is only marginally related to their topic. If a user simply needs a few articles on a given subject, the most efficient resource for them to use is a subject-specific database.

Competing with Google Fallacy

The notion that libraries compete with Google is a fallacy. Google is an advertising company whose main purpose is to sell ads for other products and service providers. Libraries have fundamentally different goals and objectives, so discussing the relationship between libraries and Google as competitive is less than helpful. Moreover, libraries could not win such a search/discovery competition with Google, and adopting a web-scale discovery service is unlikely to draw users away from Google. In the sample searches the author conducted in preparation for the current conference, Google and Google Scholar out-performed a web-scale discovery service in terms of locating subject-relevant content. This is an anecdotal report, but systematic comparative research would be a welcome addition to the conversation.

Before you Buy

At the beginning of this paper, four major goals of web-scale discovery were identified. Before libraries subscribe to these services they should ask if these goals (or a similar set of goals) are being met. They do (for the most part) create a single point of entry into the library's collections. They do not, however, support an enhanced user experience if they are not providing precise subject-relevant search results. They do a poor job supporting unmediated discovery because the search results are not precise enough, typically presenting thousands of results per query. They may be exposing unseen collections, but if those collections do not appear on the first few result pages, they are unlikely to be accessed. These systems may or may not save the end user time. Again, robust task performance studies that compare web-scale discovery services to subject-specific databases are needed.

Since web-scale discovery already exists (i.e. Google), might libraries be better off leveraging Google Scholar for their discovery layer rather than attempting to create their own service (Askey, 2014 and Kortekaas, 2012)? This approach acknowledges the reality of modern search preferences without spending funds on services that may not provide value commensurate with the cost. Libraries would continue to maintain subject level databases and leave the discovery layer up to a free service that is already widely used. In this scenario, librarians would continue to emphasize subject level resources but also make sure users were able to access material discovered through Google Scholar via Open URL, among other methods.

Libraries that pursue either approach should remember that traditional library vendors are like Google in that their search algorithms are essentially black boxes. We cannot know how results lists are ultimately generated with enough specific detail and how commercial interests influence the algorithms that produce the search results. That said, libraries may feel at ease using Google Scholar since there is no known commercial element to the service as Google Ads are absent from the Google Scholar interface. The black box nature of discovery search algorithms is certainly a concern (Breeding, 2015), but that is beyond the scope of this paper.

Since funding is fungible, libraries should consider the tradeoffs they make when purchasing a discovery layer that does an inadequate job at helping users access subject-specific content. Even a small institution is likely to spend $10,000 or more annually for a discovery tool. The author encourages libraries to consider the impact those funds could have on improving student success in other ways. For example, the funds could be dedicated toward professional development for instruction librarians; used to improve website design; used for modernizing library instruction classrooms; or used to improve services for online students. As student success is the ultimate goal, the question of how to spend scarce resources is worth critical reflection.

References

Askey, D. (2014). Giving up on discovery. Retrieved from http://taiga-forum.org/giving-up-on-discovery.

Breeding, M. (2015). The future of library resource discovery. *A white paper commissioned by the NISO Discovery to Delivery (D2D) Topic Committee*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.675.3000&rep=rep1&type=pdf

Breitbach, W. (2012). Web-scale discovery: A library of babel? In *Planning and implementing resource discovery tools in academic libraries* (641-649). IGI Global: Hershey, PA.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, *176*(1):101-108.

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

Elmborg, J. K. (2003). Information literacy and writing across the curriculum: Sharing the vision. RSR. *Reference Services Review, 31*(1), 68–80.

Kortekaas, Simone (2012). Thinking the unthinkable: A library without a catalogue – reconsidering the future of discovery tools for Utrecht University library." *Presented at the LIBER General Annual Conference.* Retrieved from http://www.uttv.ee/naita?id=12538

Richardson, H. A. H. (2013). Revelations from the literature: How web-scale discovery has already changed us. *Computers in Libraries, 33*(4), 12–17.

Tay, A. (2014). Four possible Web Scale Discovery Future Scenarios. Retrieved from http://musingsaboutlibrarianship.blogspot.com/2014/12/four-possible-web-scale-discovery.html#.Vvrov-IrJpg

Weinberger, D. (2007). *Everything is miscellaneous: The power of the new digital disorder*. New York: Times Books.