Web Scale Discovery

What and Why?

Abstract

Web scale discovery services for the library environment have the capacity to more easily connect researchers with the library’s vast information repository. This includes locally held and hosted content, such as physical holdings, digital collections, and local institutional repositories. Perhaps more significantly, web scale discovery also accesses a huge array of remotely hosted content, often purchased or licensed by the library, such as publisher and aggregator content for tens of thousands of full-text journals, additional content from abstracting and indexing resources, and content from open access repositories.

This chapter defines web scale discovery and highlights a few key concepts essential for understanding these services. For anyone who has worked a reference interview and heard a student utter, “I couldn’t find an article in the library catalog,” web scale discovery services hold tremendous potential. Extensive research on user expectations in the discovery arena, and the tools used by those seeking information—tools often disassociated from the library and often overlooking much of what the library holds and licenses—provide ample rationale for why web scale discovery is important for the library environment.

What Is Web Scale Discovery?

Connecting users with the information they seek is one of the central pillars of our profession. Succinctly put, Web scale discovery can be considered as deep discovery within a vast ocean of content. The mechanics behind Web scale discovery are not necessarily new, though a commercial application of this approach within the library environment—efficiently and, it’s hoped, effectively—is very new. While there are various approaches to Web scale discovery, this issue of Library Technology Reports will focus on what today appears to be the most common approach, which, at its heart, involves huge, centralized, preaggregated indexes searched by the end user.

Expanded further, Web scale discovery is—or certainly holds the potential to be—the evolution that libraries have long sought for information discovery. As information professionals, we all have at least a general awareness of the evolution of discovery tools within the library context. Such tools initially were print-based, such as bound handwritten catalogs, the card catalog, and works such as Poole’s Index to Periodical Literature and the Reader’s Guide to Periodical Literature. For the past several generations, such tools gradually transferred into the automated, electronic realm, with an obvious example the development and evolution of the online integrated library system (ILS) with a front-end catalog accessible to librarians and end users. These catalogs were initially available within the local library’s physical building, often through a menu-driven, text-based interface. With the development in the 1990s of the Web’s physical infrastructure and the empowerment of many to access this new environment, library vendors created HTTP Web-based online catalogs. Other evolutions included early pioneers in broad online information systems, including Dialog and LexisNexis. The 1990s ushered in growth in publisher-based electronic journal content, e-text and e-book content, abstracting and indexing databases, and full-text content aggregators looking to pull related information together within an easily accessible and searchable electronic medium. Many of these services or products were initially provided on CD-ROM or through text-based, menu-driven networked systems, which all eventually evolved into
Abstract

Debuting at the end of 2007, WorldCat Local represents the first to market web scale discovery service as defined in this report, and presently enjoys the largest install base of any web scale discovery service profiled in this report. This chapter provides a brief history, overview, and a few insights into the future development path of WorldCat Local, describes the local and remote content associated with the WorldCat Local index, and highlights some of the features, functionality, and flexibility associated with the WorldCat Local interface.

Overview

OCLC released the initial version of WorldCat Local in November 2007, following an earlier development period with trials dating to spring 2007. The experience of a pilot development partner, the University of Washington, was profiled in the August 2008 issue of Library Technology Reports. The UW pilot went live in spring 2007, and thus, for the library environment, represents the first single search discovery service combining millions of physical and electronic items within a single search result set. Approximately thirty million article-level items were intermingled with the WorldCat database in the UW pilot. In 2009, OCLC ramped up WorldCat Local and entered into additional partnerships to include substantially greater amounts of article-level content, all within an interface utilizing a single search box, relevancy-ranked results, and a back-end centralized index. Two versions of the discovery platform exist, the full-fledged WorldCat Local and the streamlined WorldCat Local “quick start.” A few of the differences are noted later in this chapter; for a more detailed comparison, see OCLC’s informative FAQ list on its website. Many of the features in WorldCat Local are available in WorldCat Local “quick start” (and, as noted below, much of the look, feel, and functionality of both versions are carried over from the WorldCat.org catalog interface). In brief, a few key options available in WorldCat Local and absent from “quick start” include integration flexibility with multiple ILSs (for example, both a local ILS and a consortial ILS, instead of a single ILS), the option to enable users to refine search results by branch location, relevancy ranking that takes into account collections from other libraries in a consortium, and resource-sharing options other than through WorldCat Resource Sharing or ILLiad. At the time of this writing, over 1,000 sites in North America and Europe have implemented either WorldCat Local or WorldCat Local “quick start.” The majority of implementations are academic institutions, though public libraries and special libraries are represented as well.

WorldCat Local “quick start”–related FAQs


Regardless of version, the interface and discovery service for WorldCat Local is hosted by OCLC. Product support is offered through various modes (phone, e-mail, website) and available 24/7. Assuming a library has holdings within the WorldCat catalog and a FirstSearch WorldCat subscription, WorldCat Local “quick start” is included in an institution’s base subscription at no additional cost. The full version of WorldCat Local has a one-time implementation fee and is
Abstract

With a general release in mid-2009, Serials Solutions Summon is another early contender in the Web scale discovery space for libraries. Serials Solutions Summon was built from the ground up as a library Web scale discovery service. This chapter provides a brief history and overview of the Summon service, describes the local and remote content associated with the Summon index, and highlights some of the features, functionality, and flexibility associated with the Summon interface.

Overview

Serials Solution began dedicated development of its Web scale discovery solution, Summon, in 2008, building the product from scratch as a new platform. Public announcement occurred in January 2009, and after work with development partners, Summon entered general release in July 2009, making it one of the early entrants into the library Web scale discovery environment. At the time of this writing, Summon has over 120 committed customers in eighteen countries; 80 of these sites are currently live. Summon’s development focus was academic customers, and such customers make up the lion’s share of current sites. That said, the Summon discovery service is also in use at three public libraries, as well as at a statewide library system, of which hundreds of public libraries are members.

Summon is offered as a hosted software-as-a-service solution providing the Summon service and index. Annual subscription pricing relies primarily on the institution’s FTE count, but also considers other factors, such as the degree-granting status for university customers. The pricing for Summon is not impacted by the number of items included from a library’s local collections. Discounts are available for multiyear and consortial subscriptions. The annual subscription fee is inclusive and covers items such as ongoing support, inclusion of local content, access to developed APIs, and application enhancements. Serials Solutions provides updates and enhancements approximately every three to four weeks, and, because the service uses a hosted model, these updates are provided quickly to its customers. Serials Solutions support is available 24/7, and a variety of communication options are provided. Serials Solutions indicates that new customers can typically have their Summon instance live within six weeks from the start of implementation.

Content and Scope

Publisher Content

Summon currently has a very large centralized index, providing access to content sourced from a multitude of commercial databases and publishers. This material includes content from 94,000+ journals and 6,800 publishers. As of August 2010, the Summon index numbers over half a billion items. By item count, the two largest content types are newspaper articles and journal articles, though various other content types, such as books, theses and dissertations, conference proceedings, music scores, and audiovisual materials are also present. A regularly updated list of participating publishers and journal titles indexed can be accessed at the Serials Solutions website. Agreements have been made with many major content providers and aggregators; chief providers participating in Summon include ProQuest, LexisNexis Academic, and Gale (which include around 4,000 publishers).
Ebsco Discovery Services

Abstract

Ebsco Discovery Services (EDS) represents the first of the trio of web scale discovery services debuting in 2010. Built off the established EBSCOhost platform, EDS extends the platform into the web scale discovery space through a preharvested, centralized index encompassing content sourced from Ebsco databases and beyond. This chapter provides a brief history, overview, and a few insights into the future development path of EDS, describes the local and remote content associated with the EDS index, and highlights some of the features, functionality, and flexibility associated with the EDS interface.

Overview

Ebsco began development of Ebsco Discovery Service (EDS) in 2008. Public announcement occurred in spring 2009, and after a beta period concluding later that year, public release occurred in early 2010. At the time of this writing, late summer 2010, approximately twenty-five customers have gone live with an EDS implementation. Ebsco indicates that new customers can generally be set up and ready to go live within eight to ten weeks. EDS is based in large part on the infrastructure and interface associated with the popular EBSCOhost platform, which debuted around 1994. Early EDS development partners were generally academic customers, though other library types, including at least one public library, have more recently begun trials of the service. EDS is offered as a hosted platform; no local installation options are available. When initially released, EDS required user authentication prior to conducting a search. In mid-2010, Ebsco released a Guest Mode option providing unauthenticated users with some limited search capabilities.

The annual subscription pricing model relies primarily on the institution’s full-time equivalent (FTE) count and level of service desired. Level of service can include factors such as the number and types of local library resources harvested and indexed (such as local digital collections and institutional repositories). Multiyear and consortial discounts are available. Ebsco provides underlying application or interface updates to portions of EDS approximately every three months. Ebsco telephone customer support is available 24/7 Monday through Friday and for reduced hours on the weekend; in addition, customers can report issues through Ebsco’s website or via e-mail.

Content and Scope

Publisher Content

At time of writing, the base index underlying the EDS service includes content from nearly 20,000 providers, in addition to metadata drawn from tens of thousands of book publishers. This base index presently includes metadata for more than 45,000 journals, more than 800,000 CDs/DVDs, nearly six million books, and more than one hundred million newspaper articles; this base index is searchable by all EDS customers. EDS includes items from several open-access repositories, such as materials from the DOAJ (Directory of Open Access Journals), OAISTER, and arXiv.org e-Prints. Ebsco creates a unique index for each EDS customer, which includes local harvested content (ILS catalog records, digital collections, etc.), the base index content, and additional content pulled from Ebsco-sourced databases for which the library has a
Abstract

Publicly released in mid-2010, Primo Central extends the Primo next generation discovery layer, released by Ex Libris several years earlier. This chapter provides a brief history, overview, and a few insights into the future development path of Primo Central, describes the local and remote content associated with Primo Central, and highlights some of the features, functionality, and flexibility associated with the Primo Central interface.

Overview

Ex Libris began development of its next-generation discovery layer, Primo, in 2005, with official public release occurring in 2007; Primo version 3 was released in spring 2010. Hundreds of libraries worldwide have implemented Primo. The Primo discovery platform harvests and indexes local library collections, such as bibliographic records, digital collection materials, and items within institutional repositories, and provides a common interface for discovery of these materials. In addition, Primo can be configured to search remote repository indexes and blend the library’s local collections with the remote index results. Primo Central, Ex Libris’s Web scale discovery component, was officially released in mid-2010. Primo Central extends the base Primo discovery experience by also searching a large preharvested central index of article-level content from a variety of publishers and aggregators. For the remainder of this chapter, Primo Central and Primo will be used interchangeably; many of the points discussed apply equally to the Primo next-generation discovery layer with or without the Primo Central service. Given that Primo Central is an extension of Primo, the interface and many of the features are the same. At the time of this writing, late summer 2010, approximately fifty customers have signed on as subscribers to the Primo Central service, with several customers already live on Primo Central. While academic libraries make up the large proportion of customers, public library customers are also present and will likely become more numerous in the future. For example, Ex Libris recently announced that the National Library of Finland, representing various library types, research institutes, archives, and museums, had chosen the Primo/Primo Central platform.

The Primo discovery layer can be hosted by Ex Libris or the local library; in either case, the central, preaggregated index associated with Primo Central is offered as a managed service and hosted by Ex Libris in a cloud environment. Primo Central is offered as a subscription service. Pricing considerations for Primo include whether an ExLibris hosted or locally installed instance is chosen, the institution’s full-time equivalent (FTE) student count, and the number of local records (such as local library catalog and digital collections records) harvested into the system. Consortial discounts are possible. Ex Libris provides minor application and interface enhancements approximately every three months; major release updates occur approximately every fourteen to sixteen months. Ex Libris customer support is available 24/7, and a variety of communication options are supported.

Content and Scope

Publisher Content

At the time of this writing, the hosted and centrally managed Primo Central index numbers approximately
Abstract

The previous chapters introduced web scale discovery and profiled a majority of the key players engaged in this space as relates to the library environment. While similarities abound, differentiators are present as well. This chapter highlights some of the differences in the areas of content coverage, metadata and relevancy, pricing, integration with other systems, and the interface. As evidenced throughout this report, each service continues to evolve at an extremely rapid pace in terms of content covered, and the features, functionality, and flexibility of the interface. While these services each hold great potential, a final note observes that web scale discovery services, at least at their present stage of development, are not the “final word” for the library discovery environment.

Web scale discovery platforms customized to the library environment, handling local library and remotely hosted aggregated publisher content are in their extreme infancy. As observed in the first chapter, features, functionality, and content scope are changing—expanding—rapidly for all players. Press releases occur often, and annual library conferences provide a showcase forum for vendors to introduce their products to potential new customers and highlight enhancements to existing customers. Vendors host presentations and panel sessions discussing the merits of their discovery service and oftentimes provide information on why they feel their offering is the best on the market. From the preceding chapters, readers will note many similarities among the discovery services, and this observation is indeed valid. Given the extremely rapid cycle of development combined with the growing openness of such platforms, this issue of Library Technology Reports wasn’t constructed as a compare-and-contrast product survey. Things change through enhancement cycles as vendors progress beyond version 1.0 and customers request new features.

Customers also create their own innovations facilitated by the openness of these platforms; as platforms become more open, libraries with technical staffing can truly customize these tools to their local environments and include additional functionality. Consider just a few examples. Claremont Colleges Library, with its Sherlock Search, has brought two discovery services together—a front-end Primo interface with harvested local resources, blended with commercial content populated through the Summon index service in the background. North Carolina State University has developed its own front-end interface, QuickSearch, which pulls content from a multitude of services. For a single search, this custom interface returns organized results including (but not limited to) commercial content—such as articles—from the Summon index, book and media materials from NCSU’s Endeca-based catalog, results from a library website search, and Did You Mean? suggestions utilizing Yahoo! Web Services.

So, acknowledging the power and creativity opening doors as never before, what are some things to keep in mind for new customers that have yet to embrace a Web scale discovery service? Here are some broad factors to consider.

Content

The ultimate goal of any discovery service, bar none, is to place content in the hands of the user or, more specifically, to discover, present, and deliver relevant content in a convenient, intuitive manner to today’s
Questions to Consider

Abstract
For numerous reasons, libraries contemplating the purchase of a web scale discovery service should very carefully complete their homework. The following set of questions can serve as a primer for those engaging in their own evaluations of the increasingly competitive library web scale discovery space. Questions are divided into several topical areas: general and background questions; local library resources; publisher and aggregator indexed content; open access content; relevancy ranking; authentication and rights management; and the user interface.

The following questions can serve as a springboard for prospective library customers seeking more information from vendors as part of their own evaluation of library Web scale discovery services.

Section 1: General and Background Questions

1. Customer Install Base
   - How many current customers do you have that have implemented the product at their institutions (i.e., the tool is currently available to users or researchers at the institution)?
   - How many additional customers have committed to the product?
   - How many of these customers fall within our library type (e.g., higher ed academic, public, K-12)?

2. References
   - Can you provide website addresses for live implementations that you feel serve as a representative model matching our library type?
   - Can you provide references—name and contact information—for the lead individuals you worked with at several representative customer sites that match our library type?

3. Pricing Model, Optional Products
   - Describe your pricing model for a library type such as ours, including initial up-front costs and ongoing costs related to the subscription and technical support.
   - What optional add-on services or modules (federated search, recommender services, enrichment services) do you market that we should be aware of and that are related to and able to be integrated with your discovery solution?

4. Technical Support and Troubleshooting
   - Briefly describe options (including hours of availability) customers have for reporting mission-critical problems and for reporting observed non-mission-critical apparent glitches.
   - Briefly describe any consulting services you may provide above and beyond routine support services (e.g., consulting services related to harvesting of a unique library resource for which an ingest/transform/normalize routine does not already exist).