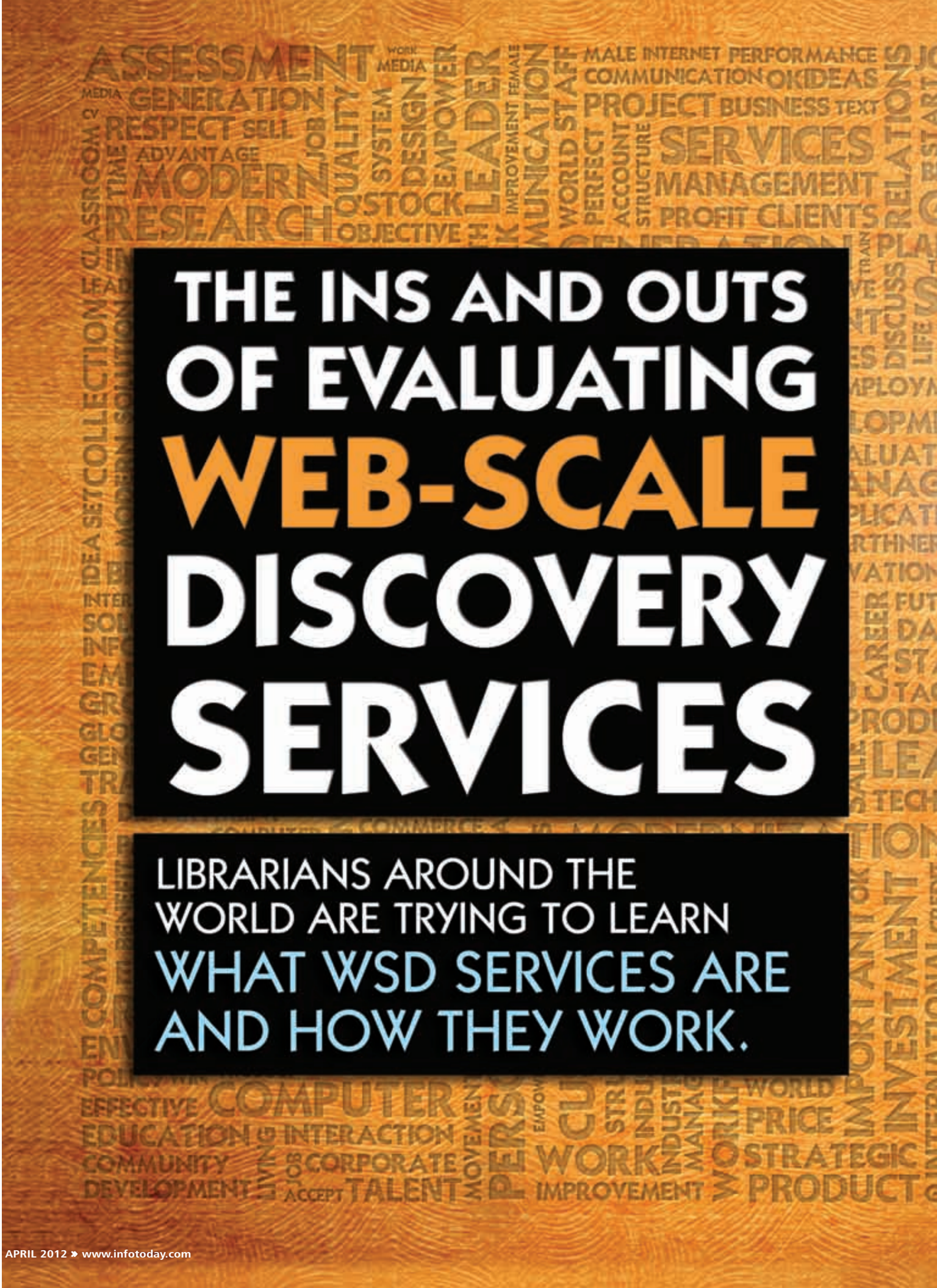


THE INS AND OUTS OF EVALUATING WEB-SCALE DISCOVERY SERVICES

LIBRARIANS AROUND THE WORLD ARE TRYING TO LEARN WHAT WSD SERVICES ARE AND HOW THEY WORK.

By Athena Hoeppner

By now, we librarians are familiar with the single-line form, the consolidated index, which represents a very large portion of a library's print and online collection. Our end users certainly are familiar with the idea of a single search across a comprehensive index that produces a large, relevancy-ranked results list. Even though most patrons would not recognize the term web-scale discovery (WSD), it is what they have come to expect. More and more libraries are stepping up to meet their users' expectations by implementing WSD services. Librarians around the world are trying to learn what these services are and how they work, evaluating the services on the market, selecting and implementing a service, and then teaching colleagues and patrons all about it.

This is just what I needed to do in my position as the electronic resources librarian at the University of Central Florida. Based on my investigations, this article explains WSD concepts and terminology, shares findings from my interviews with major WSD vendors, and provides a template checklist, which librarians can use during their own exploration of these systems.

Caveats: While focusing on the discovery layer and central indexing components of these systems, which are of primary interest to most librarians, the article no doubt oversimplifies some WSD concepts. The article is not meant to tell you which service is best in general, which is best for your institution in particular, or even how to select the best service for your library. That said, in the space permitted, it will, however, provide enough general understanding of WSD to help readers identify issues and features that are important to their library and make their WSD investigations and evaluations more efficient.

What Are We Talking About?

Web-scale discovery is new enough that the vocabulary surrounding the service is in flux. To enable discussion and training, we need common terminology

WSD Key Concepts to Understand

Web-Scale Discovery
A preharvested central index coupled with a richly featured discovery layer that provides a single search across a library's local, open access, and subscription collections

Central Index
The collection of preharvested and processed metadata and full text that comprises the searchable content of a WSD service

Discovery Layer
The user interface and search system for discovering, displaying, and interacting with the content in library systems, such as a WSD central index

Preharvested Index
Metadata and full content systematically and periodically accumulated and processed in advance of searches; data is gathered from multiple sources and processed into a central index

Mutually Licensed Content
Content that is licensed by both the library and the WSD vendor so that the content can be searched and displayed to authorized users

and definitions. In the future, librarians and vendors may rely on consistent and standardized definitions produced by the National Information Standards Organization or some other agency. This article relies on definitions largely based on Jason Vaughn's January 2011 issue of *Library Technology Reports* published by American Library Association's TechSource and on vendor websites and presentations:

- **Web-scale discovery.** A preharvested central index coupled with a richly featured discovery layer providing a single search across a library's local, open access, and subscription collections

- **Central index.** The collection of preharvested and processed metadata and full text that comprises the searchable content of a WSD service: Central indexes typically include full text and citations from publishers; full text and metadata from open source collections; full text, abstracting, and indexing from aggregators and subscription databases; and MARC from library catalogs; also called the base index, unified index, or foundation index.

- **Discovery layer.** The user interface and search system for discovering, displaying, and interacting with the content in library systems, such as a WSD central index: Discovery layers are not new to librarians or unique to WSD services. Marshall Breeding, director for innovative technologies and research at Vanderbilt University Library, maintains a directory on his Library Technology Guides' Discovery Layer Interfaces page (www.librarytechnology.org/discovery.pl), and many of the systems are well-known to librarians as the end-user search interfaces for their OPACs.

Common features of discovery layers for web-scale discovery services include the following:

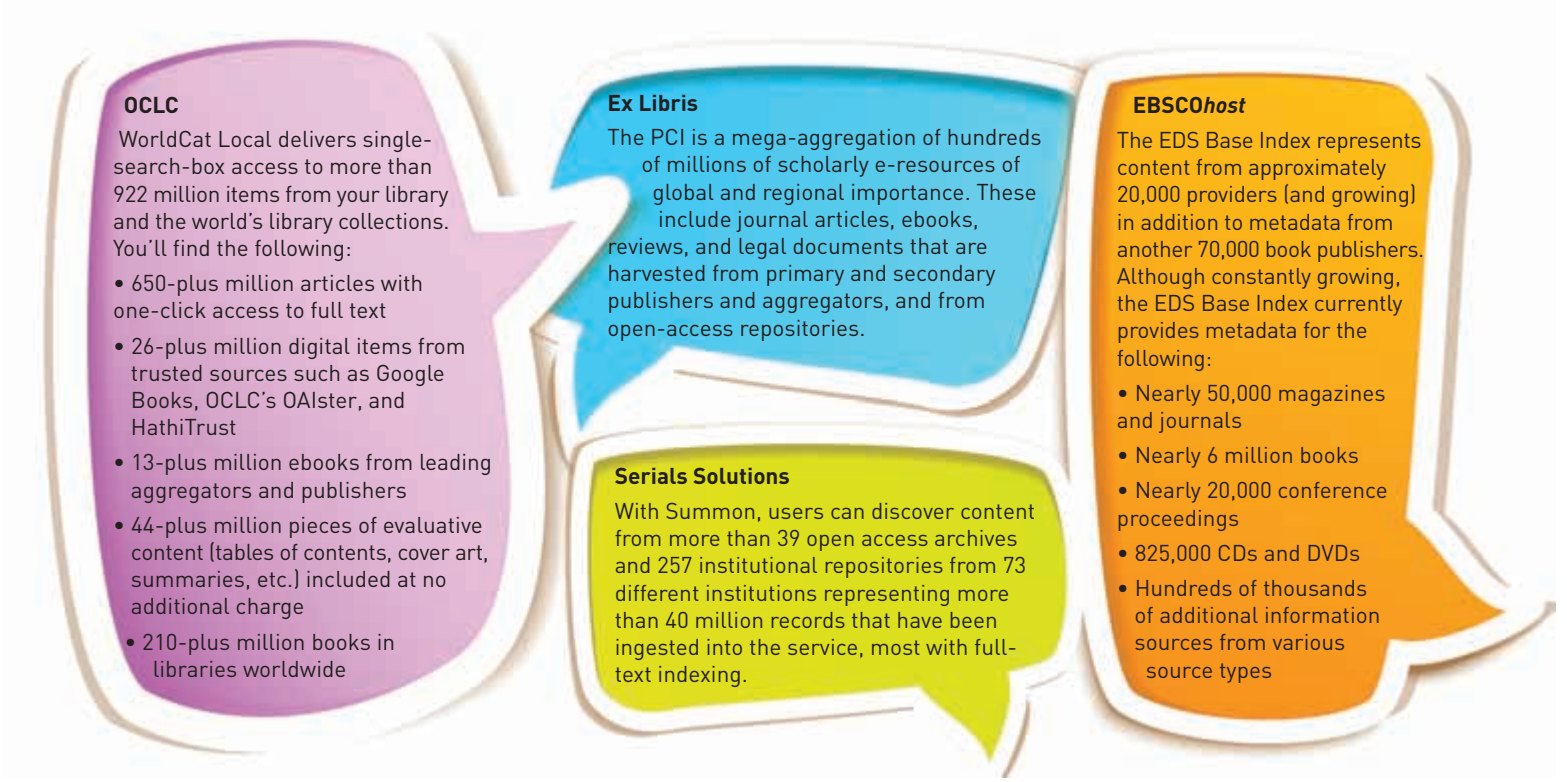
- Single search across the central index
- Fast response time
- Relevancy-ranked results list
- Facets, sort, and other tools for refining and using the results
- Connections to full text via direct links and OpenURL
- End-user accounts and features

Four vendors offer both WSD components—the discovery layer and the central index:

- EBSCO's Discovery Service (EDS) (www.ebscohost.com/discovery)
- Ex Libris Ltd.'s Primo Central Index (PCI) (www.exlibrisgroup.com/category/PrimoCentral)
- Serials Solutions' Summon (SSS) (www.serialssolutions.com/discovery/summon)
- OCLC's WorldCat Local (WCL) (www.oclc.org/worldcatlocal)

So, in brief, the central index is content, and the discovery layer is the interface. The vendors typically license the two halves

What the Vendors Say About the Content of Their Central Index



as a unified package. However, a variety of discovery layers can be used to search the central indexes from EBSCO, Ex Libris, and Serials Solutions. There are several implementations that use VuFind, Mango (the discovery interface created by The Florida Center for Library Automation), or other discovery interfaces in conjunction with the vendors' central indexes.

Because the central index and discovery layer play separate (but interacting) roles, have distinct underlying concepts and issues, and can be acquired a la carte, it makes sense to consider them separately.

What Is in a Central Index?

True to its name, a web-scale discovery central index is very large relative to library databases of the past. Trying to pin down and compare the exact size and content of each central index is an exercise in frustration. The vendor sites give some statistics that illustrate the size, if not the scope and depth.

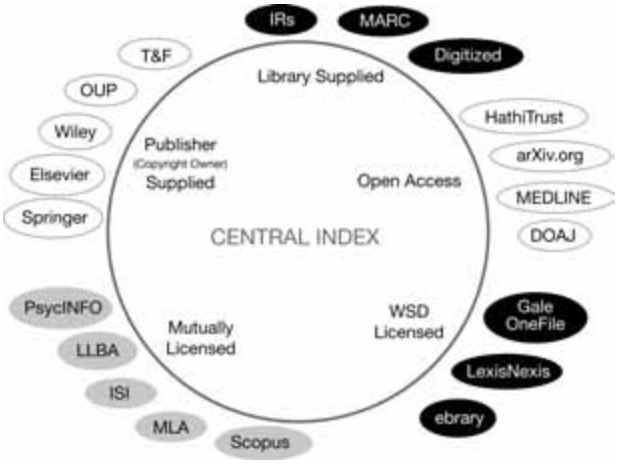
Size alone is not sufficient to evaluate a central index. Scope of the content, item types, inclusion of the full text, and richness of the metadata are all important aspects. To best serve the library, the central index should align with both the library's available content (i.e., the books, databases, and full text) and the approach to research appropriate for its users. A library serving an active group of graduate students and researchers may make different choices from a library serving community college students. That is why it is important to understand how a central index is created and how a library's collections and implementation choices will help libraries evaluate the offerings.

The image at right shows a generalized model of a central index. The bubbles outside the main circle are examples of the different data sources that might be harvested into the central index. Common methods of gathering the content include FTP and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Frequency of data loads vary, but many sources load data daily.

Content in a central index may include any or all of the following types:

- **Library supplied data.** Included are MARC and metadata from local digitized collections and institutional repositories. As part of the initial service configuration, the library may provide a complete file (or files) of the MARC, work with the service to map MARC fields to the preferred record format of the WSD vendor, identify library locations, and otherwise integrate the catalog data into the central index. Metadata from digitized collections and institutional repository content are often harvested using OAI-PMH. The library can set the frequency of MARC and OAI-PMH updates.

- **Open access and public domain data.** This includes citation metadata, indexing and abstracts, and full text from open collections and indexes.



The central index may consist of many kinds of preharvested and licensed content.

- **Publisher metadata and full text.** Included are citation metadata, author-supplied abstracts and keywords, and full text of documents. Many online journal and book publishers supply the full text of articles and chapters so this material can be *searched*, but the full text can only be displayed or downloaded if the library has rights to the content.

- **WSD-licensed material.** WSD vendors actively seek agreements with aggregators and other content providers to include content that would interest libraries and enrich the central index. For example, all four major WSDs harvest some content from LexisNexis, Credo Reference, and Gale, part of Cengage Learning. Agreements may allow the central index to include a database in its entirety or just in part. The exact deals between WSD services and database vendors are not uniform and are subject to change. Some of the database vendors are also WSD competitors. Not surprisingly, the databases owned by one vendor typically don't appear in the central indexes of its competitors.

- **Mutually licensed.** Some metadata will be loaded into the central index but can be searched and/or viewed only if the library has a subscription to the source database. For example, metadata from PsycINFO, including abstracts, subject descriptors, and other value-added content created by the American Psychological Association, will only be shown to authenticated users if both the WSD vendor *and* the library mutually license PsycINFO.

What Might *Not* Be in Your Central Index

While the idea of a central index is straightforward—just put all the content of interest to a library into a single index—the reality is much harder to pin down. In general, the content available to any given library is affected by three factors:

- Content the WSD vendor includes in its central index, which was discussed earlier

How Index Records May Vary

Not all vendors will include all metadata from original data sources. Consider this example comparing the native record from the CSA Linguistics and Language Behavior Abstracts (LLBA) with the EDS record for the same article.

Citation: "Knee" and "Generation/People" in Indo-European: Lat. poples "knee" vs. populus "people" and Parallels. Parvulescu, Adrian. *Indogermanische Forschungen 102* (1997): 74–83.

Descriptive Content in LLBA

Abstract: A curious polysemy whereby the word for 'knee' also means 'generation' is found in a wide range of Indo-European languages; examples are cited from Latin, Romanian, Old Irish, & many Germanic & Slavic languages. ... (*excised for brevity*).

Indexing:
Indo European Languages (35400);
Semantic Change (76600);
Latin (45400);
North Germanic Languages (59130);
Polysemy (66560);
Etymology (23250);
Comparative Linguistics (13850)

Classification:
5121: Descriptive linguistics; Paleolinguistics/paleography

Descriptive Content in EDS

Subject Language:
Indo-European languages

Linguistics Topic:
Lexicology
Etymology

Subject Terms:
Anatomical terms;
as metaphor

- Overlap between the library's subscriptions and the WSD agreements with database and full-text publishers

- Customization and choices the library makes about what data to expose to users

A common question when investigating central indexes is, "Does the index include *x* database?" The answer is often one of degrees and percentages. Some databases do not supply metadata to any of the WSD services. For example, SciFinder Scholar records are not included in any of the services. In other cases, a database may be in the central index of one or more WSD service, but not in another's. This is particularly common for EBSCO and ProQuest databases.

Even when a WSD vendor does not include a database, it may have basic, citation-level metadata for many of the articles and may report that the central index includes some percentage of current journal lists; however, the coverage may miss two aspects of the database. First, titles that were formerly covered by the database may or may not be in the central index. And second, subject descriptors, abstracts, and other value-added metadata that was created by the indexer will not be in the central index.

Library customizations. The final factor that affects the content of a library's central index is the choices that a library makes during implementation. A library can choose to *not* expose some of the available content. For example, all the central indexes harvest all the metadata available from JSTOR. Many libraries do not license all the JSTOR journals. A library may set the metadata for the content as available so its users can discover it and, presumably, request it via interlibrary loan, or a library can set the metadata as unavailable by default, focusing the central index on content that will be immediately available to users.

Exploring WSD Discovery Layers

While the central index is the heart of any WSD service, the discovery layer is the public face. So long as the results list contains a smattering of good hits, users' responses to the WSD service are likely to be based solely on their experience with the interface. Like the central indexes, the discovery layers have much in common.

Branding and customizations. All four vendors support basic branding that lets the library stamp its institutional

continued on page 38 »

« *continued from page 10*

logo and colors on the WSD interface. Libraries control the basic branding through administrative modules or by asking their WSD contact to make changes. Beyond the basics, the degree of customizations supported and the ease with which they can be achieved varies. For example, in EDS, and likely the other WSD interfaces, the library must select from a small set of predefined screen layouts. PCI is notable for the degree its interface can be customized via application programming interfaces (APIs) and customer-supplied cascading style sheet), which enable more extensive options for the look and layout.

Basic Branding Options	EDS	PCI	SSS	WCL
Library Logos and Banners	Yes	Yes	Yes	Yes
Interface Colors	Yes	Yes	Yes	Yes
Customize Links in the Toolbar and Footer Links	Yes	Yes	Yes	Yes
Embed Search Forms in External Sites	Yes	Yes	Yes	Yes

Searching. All the WSD services provide the search options that librarians and users have come to expect: keyword, author, title, and subject fields. The services support Boolean, phrase, and truncation searching and incorporate spell-check. Particularly for the basic search screen, there is a great deal of consistency.

Basic Search Feature	EDS	PCI	SSS	WCL
Spell-Check/Did You Mean?	Yes	Yes	Yes	Yes
Phrase/Proximity	Yes	Yes	Yes	Yes
Stems/Truncation	Yes	Yes	Yes	Yes
Keyword by Default	Yes	Yes	Yes	Yes
Single Search Entry (Simple Searches)	Yes	Yes	Yes	Yes

Advanced discovery. The advanced discovery options, such as fields available in the advanced search form, limiters, and facets, reveal differences among the services. In general, the differences fit into two categories:

- Different approaches to accomplish the same result
- Special fields or facets that capitalize on the central index's strong content areas

Where one service will provide a field on the advanced search screen, another may use a limiter or facet on the results page to narrow the results set. In addition, the specific labels for features and fields vary. When comparing the systems, consider whether a system has a way to produce results with given criteria, rather than looking for exact search fields, limiters, and facets.

Advanced Search and Limit	EDS	PCI	SSS	WCL
Fields and Boolean Searching	Yes	Yes	Yes	Yes
Full Text	Yes	Yes	Yes	Yes
Scholarly/Peer Reviewed	Yes	Yes	Yes	Yes
Language	Yes	Yes	Yes	Yes
Item Format/Content Type	Yes	Yes	Yes	Yes
Limit to Library Owns/ Available in Library	Yes	Yes	Yes	Yes

Facets	EDS	PCI	SSS	WCL
Author	Yes	Yes	Yes	Yes
Database/Content Provider	Yes	Yes	No	Yes
Format/Content Type	Yes	Yes	Yes	Yes
Language	Yes	Yes	Yes	Yes

The discovery layers each capitalize on the specific content strengths of their central index. SSS has limiter options to exclude newspapers and dissertations from the results, which is a good match for its significantly larger set of news content and exclusive rights to include ProQuest dissertations and theses. WCL has an abundance of book information and offers limiters for fiction, nonfiction, biographies, etc. Facets, in particular, reveal the different content strengths. WCL has several facets useful for books that are based on standard MARC fields, such as audience, genre, and geographical area. SSS has options for handling news articles and lets users select facets to be excluded from the results set. EDS has a range of potential facets coming from its abstract and index databases. PCI lets the implementing library choose any field to use as a facet.

Relevancy and refinement. With a large central index comes large results sets. Relevancy-ranking and post-search refinement tools are more important than ever. Every WSD service uses sophisticated relevancy algorithms with weighting systems to surface what they hope are the best hits for any given search. Each vendor has a relevancy algorithm designed with its own central index in mind, but they all consider factors such as the following:

- Currency/date of publication
- Inclusion of the search term in the title, subjects, or author fields
- Proximity of the terms to each other in the results

The default discovery layer interfaces have limited options for the customer to tweak the relevancy. PCI and WCL

both allow customers to increase or decrease the prominence of the library's local collections.

End-user features and Web 2.0. All the discovery layers give end users tools for organizing, saving, and exporting results. Web 2.0 and social media features are making an entrée into discovery layers but are not widely offered yet. PCI and WCL support tagging, user ratings and reviews, integration with personal accounts, and sharing with external sites, such as Facebook.

End-User Features	EDS	PCI	SSS	WCL
Formatted Citation Export	Yes	Yes	Yes	Yes
Formatted Citation Display	Yes	Yes	Yes	Yes
Persistent Folders/Lists	Yes	Yes	Yes	Yes
Shared Folders/Lists	Yes	Yes	No	No
User Accounts	Yes	Yes	Yes	Yes

Enhancements. There are a variety of ways that results are enhanced. All the discovery layers have options for including book jackets and for showing Thomson Reuters Journal Citation Reports rankings and citation counts if the library subscribes to the appropriate services from Thomson's Institute for Scientific Information. Widgets can be used to bring in additional functionality, such as ask-a-librarian chat.

SSS and EDS both have enhancements that inform users about relevant external databases that may have content that is not in the central index. The SSS recommender service identifies relevant databases based on the user's search. EBSCO*host* Integrated Search broadcasts the searches to preselected databases and shows the number of hits. Users can click on the database to go to the native interface and see the results. Both these enhancements help overcome the absence of some key databases from the central index content.

Factors to Consider When Evaluating WSD Systems

When considering your options think about how important each of these factors is to your library and rate it on a scale of 1 to 5, with 5 being critical.

Central Index	1	2	3	4	5 (critical)
Mutually Licensed Content: Does the central index help expose the rich metadata from your A&I subscriptions?					
Item Types: Does the central index include item types important to your library? Are the results full of item types you do not want?					
A&I content: Is content from specialized indexing searchable?					
Full-Text Indexing: Does the central index include the backfiles for journals and have searchable full text for a variety of sources?					
Discovery Layer					
Relevancy Ranking: Does the relevancy expose the kind of material important to your users?					
Search/Discovery Tools: Do the basic/advanced options, facets, and limiters enable your users to find the content they seek?					
End-User Features and Controls: Do the folders/lists, tagging, exporting citations, and other features support user interaction with the content?					
Interface customizations: Can you control the branding and layout, include widgets, and use APIs?					
Results Enrichment: Is it possible to enhance the results with book jackets, external information, and recommenders?					
Overall Usability: Do users understand the interface and make good use of the options?					

Considerations and Choices

All the central indexes are of a size and scope unprecedented in library products prior to the advent of WSD services. Any WSD service is likely to be very well-received by patrons. Picking the one that is best may come down to the library's preferences, philosophies, and existing vendor relationships. Of course, there are additional factors that come into play for libraries when selecting a web-scale discovery service. The back-end options, existing business relationships, and pricing will most certainly influence the decision. As libraries evaluate the options, they may find it useful to create a weighted template they can use of the features they consider most important.

Have a Look for Yourself

Visiting several WSD service implementations from the major vendors is a good way to experience the commonalities, observe the differences, and discover features that may have been added since this article was written. When perusing implementations, keep in mind that each institution customizes its central index and discovery layer. Any individual site does not show the full range of possibilities available from any vendor. The implementations shown above provide good examples of libraries that use a single vendor for the central index and discovery layer and have fairly out-of-the-box implementations. Customizations and branding rely on standard options rather than significant changes through APIs and programming, and they allow guest searching. Unauthenticated users can enter searches and see a results list.

As I noted at the start, space did not permit this article to be either exhaustive in its explanation of WSD services or comprehensive in its comparison of features and functionalities of the main WSD vendors. I do hope, however, that you have found it to be an interesting and informative overview of the core issues to consider in reviewing and selecting the service that you need for your library.

Athena Hoeppner *(athena@ucf.edu) is the electronic resources librarian at the University of Central Florida.*

Sample WSD Implementations

EBSCO's Discovery Service (EDS)

University of Georgia
www.libs.uga.edu

The University of North Carolina–Greensboro
http://library.uncg.edu

Okanagan College, Canada
www.okanagan.bc.ca/administration/students/library.html

Bournemouth University, U.K.
www.bournemouth.ac.uk/library/resources/mySearch.html

Ex Libris' Primo Central Index (PCI)

The University of Iowa
www.lib.uiowa.edu

Curtin University, Australia
http://library.curtin.edu.au

Northwestern University
www.library.northwestern.edu

The University of Manchester, U.K.
www.library.manchester.ac.uk/searchresources/librarysearch

Serials Solutions' Summon (SSS)

Oregon State University
http://osulibrary.oregonstate.edu

University of Michigan
www.lib.umich.edu

North Carolina State University
www.lib.ncsu.edu

Senate House Libraries, University of London, U.K.
http://external.shl.lon.ac.uk/summon

OCLC's WorldCat Local (WCL)

University of Washington
www.lib.washington.edu

The University of Arizona
http://universityofarizona.worldcat.org

Bethel University
http://bethellibraries.worldcat.org

York St John University, U.K.
http://library.yorksj.ac.uk

Copyright of Computers in Libraries is the property of Information Today Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.