DISCOVERY SERVICES

COMING INTO FOCUS

Web-scale discovery services face growing need for best practices

By Michael Kelley

> "There is a certain 'black box' atmosphere out there at the moment which is not in the collective best interests of the community."

—Bruce Heterick, VP for outreach and participation services, JSTOR and Portico

Many academic and research libraries are making significant investments in the relatively new and still imperfectly understood Web-scale discovery systems—the four leading ones being EBSCO Discovery Service (EDS) (2010), Primo from Ex Libris (2010), Summon from Serials Solutions (2009), and WorldCat Local from OCLC (2007).

There is great hope that these rapidly maturing products will not only promote information literacy strategies but also deliver what metasearch (or federated search) has failed to achieve—a Google-like interface that provides a fast, single point of entry to an institution's relevant and vetted scholarly content.

However, at the moment, even as libraries are struggling to reestablish themselves as a compelling place to start research, the three constituencies—libraries, content providers, and discovery service vendors—cannot even agree on a common vocabulary to describe what they do.

The time is ripe for innovation and collaboration, but an equitable balance of interests may prove elusive amid intense competitive battles as well as uncertainty about complex issues, such as resource coverage, depth and breadth of indexing, relevance rankings, and usage reporting, to name a few.

Nevertheless, recognizing the need for a constructive dialog around these issues, two groups have begun to explore in earnest best practices for this complex genre of software, which is based on indexed search.

The first initiative, the Open Discovery Initiative (ODI), was seeded at the American Library Association's annual conference in June 2011 in New Orleans, where Oren Beit-Arie, the chief strategy officer at Ex Libris; Jenny Walker, director of strategic initiatives for Ex Libris; and Marshall Breeding, un-

Michael Kelley (mkelley@mediasourceinc.com) is Editor-in-Chief, LJ

til recently director for innovative technologies and research at Vanderbilt University, Nashville (an Ex Libris customer), hosted an initial meeting.

As a result of this opening push from Ex Libris, the National Information Standards Organization (NISO) decided to form the ODI working group in October 2011, with Walker and Breeding as cochairs.

"We wanted to gauge interest in exploring the issues encountered with these new discovery services and in pursuing more formal standards or best practices for information providers to provide content to discovery services," said Beit-Arie at the time. "We received an overwhelmingly positive response from stakeholders, which led the group to bring the project forward to NISO."

The project is a natural fit for NISO.

"The function of NISO is to provide a neutral forum where interested parties can discuss these issues," says Todd Carpenter, the executive director of NISO. "Systems for finding and delivering electronic content, such as indexed discovery systems, are perfect examples of where industry consensus is needed and where standards development can be most useful," he says.

In tandem with the ODI effort, which conducted a survey this month and is charged with recommending best practices by May 2013, there is a parallel code of practice effort from the National Federation of Advanced Information Services (NFAIS), which represents the consolidated view of vendors and content providers.

"NISO's willingness to implement the ODI initiative only further demonstrates that there is a real need within the information community to develop a better understanding of and resolution to the discovery service issues on the table," says Bonnie Lawlor, executive director of NFAIS, who also is serving on the ODI working group.

Resource coverage and indexing

Librarians expect that the large central index that underpins discovery services will maximize awareness and usage of the library's entire collection, particularly for electronic subscriptions—which in some libraries account for 85 percent of the materials budget, according to ODI.

However, before deciding on a service, librarians find it difficult to measure how much of the content they subscribe to is covered by a given central index (a major competitive differentiator) and how deeply any included content has been indexed. In addition, measuring impact after implementation has been difficult in the absence of standardized and timely usage reports that can be easily compared.

Librarians are left nonplussed, as they sort through the noisy and competing assertions of the discovery providers. Additionally, they often make a substantial and critical investment without fully understanding what they are purchasing or the ultimate return on the investment.

At Harvard University, Laura Morse, manager of library technology services and an ODI working group member, says many researchers express concerns about the lack of visibility for what content is included in the central index. They are also uncertain whether the included content has been indexed using full text, subject headings, an abstract, or some other metadata. Morse says in an age of facets and relevance ranking, unevenness of indexing depth could omit key resources from research results.

"Today, indexed search is still governed by proprietary deals between discovery providers and information providers, which results in a blurry and inconsistent 'ecosystem' that underserves libraries and users and is becoming increasingly hard for libraries to evaluate and measure."

—Jenny Walker, director of strategic initiatives, Ex Libris

"This is why I think transparency for both breadth of content—that volume X of journal Y is included in the discovery service—and depth of content—for journal Y, metadata only is indexed, but for journal Z, metadata, abstract, and full text are indexed—is necessary to provide critical information for both librarians evaluating which tools are best for their researchers and for the researchers evaluating which available tool should be used to answer a particular question," Morse says.

Without such transparency, for example, it may be difficult to determine with confidence the state of the literature regarding a topic, which is a primary bibliographic step for any scholarly research.

"If a discovery service does not indicate what has been searched and how the original index was created, it is impossible to assume that one has thoroughly searched a given source or sources and that one can move on to the next," says H. Rob-

BREEDING PHOTO BY STEVE GREEN/VANDERBILT UNIVERSITY

"While each of the different discovery services has extensive and ever-growing coverage in their indexes of the body of content, the key question is which of the products might offer the best coverage for a given library's subscriptions."

—**Marshall Breeding, ODI cochair**

ert Cohen, founder and director of the Retrospective Index to Music Periodicals (RIPM) and the RIPM Online Archive. "We simply need to know what has been searched and the process that generated the data sources accessed," says Cohen, who also is professor emeritus of music at the University of Maryland.

OCLC first encountered the resource coverage issue with member libraries a few years ago.

"It's a valid concern," says Jeff Penka, portfolio director for end user services at OCLC and an observer to the NISO working group. "From the beginning, we have tried to be transparent to our members about the content available through our discovery service, WorldCat Local."

Breeding says libraries often analyze a discovery service after implementation, but "it's also important to have adequate information in advance to help select the best one for their environment."

"While each of the different discovery services has extensive and ever-growing coverage in their indexes of the body of content, the key question is which of the products might offer the best coverage for a given library's subscriptions," says Breeding, who presented on ODI at ALA's annual conference in Anaheim in June.

In addition, the depth and frequency of indexing can "make a dramatic difference" in results, Breeding says.

"Given the differences in access between indexing thin metadata versus full text, it seems helpful for libraries to know this information," Breeding says. "I've personally observed that at least some of the providers of discovery services mention in at least general terms the quantity of materials that are indexed in full text."

Most academic journal publishers will allow for searching (not displaying) of their full-text content in all discovery services. But Michael Gorrell, chief information officer for EBSCO and a member of the ODI working group, says in addition to information about the quantity of materials indexed in full text, librarians also need clarity about the degree of subject indexing involved.

"Discovery services can provide a Google-like experience in terms of fast, single-search access to a library's collection, but in academic research, the quality of results is the distinguishing factor," Gorrell says. "Full-text searching gets us some of the way there, but it needs help. Detailed subject indexing allows for a much greater level of granularity and precision when it comes to fine-tuned algorithms for relevance ranking," which Gorrell says was an important consideration since hundreds of millions of articles are being searched.

According to NFAIS, which is reviewing a draft of its proposed code, content providers have a big stake in more transparency.

"If they have agreed to supply content for use in a central index, the content providers need to know what specifically of their authorized content is actually being used and how it is being used in order to assess its ultimate findability," says Lawlor of NFAIS.

Uncertainty about such matters was a driver of NFAIS's decision to come up with recommended practices.

"There were concerns among NFAIS members on how discovery services were going to work and whether all links in the information chain were being properly represented and leveraged," says Barbara Dobbs Mackenzie, president of NFAIS and editor in chief of Répertoire International de Littérature Musicale (RILM), a comprehensive database of music literature. "It was clear that was not happening."

A "black box atmosphere"

Part of the problem is enormousness as well as transparency, according to Nara Newcomer, an assistant music librarian at East Carolina University who cohosted an Association for Library Collections and Technical Services (ALCTS) e-forum on discovery tool implementation and selection in May.

"The sheer size of Web-scale discovery products makes it difficult to evaluate their coverage," says Newcomer, who also was part of the group that released in August the "Music Discovery Requirements" document for the Music Library Association. "Some providers make title lists available and offer to compare an institution's e-content subscriptions to the discovery product's coverage, which is helpful, but each vendor provides slightly different information and formats it differently, making comparison among providers tricky," says Newcomer.

Serials Solutions, for example, provides a list of serials titles available in Summon that is a PDF more than 4000 pages long.

"I don't think anyone's intention really is to obscure, but on

the other hand there's a lot of content," says John Law, vice president of discovery services for Serials Solutions and a member of the ODI working group, referring to the long lists of coverage. "It's all there, it's all very transparent, but is it practical?" says Law, who agreed that the variety of formats used to describe coverage makes it very difficult for librarians to compare competing services.

"We have been working with libraries through several iterations in describing 'coverage' in a way that is meaningful and useful to them," Law says. "This effort, combined with our active participation in the NISO Open Discovery Initiative, is moving us forward in the development of consistent guidelines—agreed among all three constituencies (librarians, publishers, and service providers)—for a common means for defining and describing coverage reporting."

Greater transparency and understanding of coverage would help both librarians and content providers better grasp the workings of discovery services, according to NISO and NFAIS.

"There is a certain 'black box' atmosphere out there at the moment which is not in the collective best interests of the community," says Bruce Heterick, whose term as chair of NISO's Board of Directors ended in June and who also is the vice president for outreach and participation services at JSTOR and Portico (parts of ITHAKA S+R).

"Full transparency may not always be possible, but libraries need sufficient information to enable their own decision-making, to provide appropriate user instruction, and to empower them to select appropriate customization settings for their communities," Heterick says. "On key issues such as resource coverage, relevance ranking, and routing neutrality, they frequently do not feel they have had sufficient information."

For example, a recent report called "Paths of Discovery" by Andrew Asher of the Bertrand Library at Bucknell University (and others) notes that discovery systems exert "a form of epistemological power by virtue of their relevance ranking algorithms." Being able to understand and customize relevance algorithms better could help librarians address deficiencies in the search practices of students who have a low level of information literacy and tend to trust whatever a search engine produces from its default settings.

"The critical question for librarians is therefore how to participate (or not to participate) in this process and what level of this epistemological power to exercise," the report says.

(*LJ*'s *Academic Patron Profiles* report, scheduled for publication later this month, will address some of these questions.)

Heterick says greater transparency would also allow for a more meaningful measurement of the impact discovery services have on libraries and content providers and that this, in turn, would help define areas in greater need of standardization, which leads to innovation and best practices.

"This life cycle is important, particularly in view of how much money the ecosystem collectively is investing in these systems," Heterick says.

Walker, the ODI cochair, says standardizing and clarifying the ecosystem in this way was a major goal of ODI.

"Today, indexed search is still governed by proprietary deals between discovery providers and information providers, which results in a blurry and inconsistent 'ecosystem' that underserves libraries and users and is becoming increasingly hard for libraries to evaluate and measure," says Walker, who also made a presentation in Anaheim.

"There needs to be a good balance across all constituents for discovery services to work effectively," Walker says.

Questions of neutrality

However, competitive dynamics can make striking such a balance difficult.

Although unusual, there are still some full-text information providers, such as the American Physical Society and the American Chemical Society, that refuse to make full text available for indexing to any discovery service. So, a library may be subscribing to the content but be unaware that it is not being indexed in the main discovery tool for the library (although a discovery service provider can enhance metadata on its own to help ensure some visibility).

Also, the proprietary content partnerships Walker refers to involve, among many others, ProQuest (the parent of Serials Solutions) and EBSCO, which both provide discovery services and content resources, including abstracting and indexing (A&I) products. But neither one contributes their A&I products to the other's discovery service, leading to the problem of both having to cover that material in other ways in their discovery service. Ex Libris is impacted even more, since neither EBSCO nor ProQuest contribute to Ex Libris's Primo Central index.

Proprietary difficulties also are arising as discovery providers explore working with integrated library system (ILS) ven-

"We have been working with libraries through several iterations in describing 'coverage' in a way that is meaningful and useful to them."

—John Law, VP of discovery services, Serials Solutions

dors, such as the deals EBSCO initiated in June with Sirsi-Dynix, OCLC, and Innovative Interfaces (which also offers a discovery product, Encore Synergy, that follows a slightly different approach). The discussions revolve around whether the ILS should serve as the front end (bringing in the discovery results) or whether the discovery service should be the front end (interacting with traditional ILS functionalities, like holds lists). The goal may be creating seamless options for libraries, but making such partnerships happen is not easy given the underlying competitive stances.

Competition also leads to questions of vendor neutrality in search results.

Heterick says he has spent several months "taking a deep dive" into the usage impact of Web-scale discovery systems on JSTOR content, and it has opened "the black box" a bit and helped frame some questions he says need to be discussed openly.

"Libraries are making significant financial investments in these Web-scale discovery systems. How are they measuring the return on those investments?" he says. "If, as research from Ithaka S+R and OCLC has shown, fewer and fewer people are actually beginning their research at the library, then what are the expected outcomes of these implementations?"

Content providers, on the other hand, are handing over content for indexing purposes without really understanding the impact on usage, the indexing itself, the relevance algorithms, or the motivations of discovery service providers, Heterick says.

"In these tough economic times, if usage is the holy grail of value measurement, then should we not expect publishers/content providers to have vested interests in making sure that their content appears as prominently in the search results of these systems as possible? Does that lead to undesirable practices?" he asks.

Tim Collins, president of EBSCO Publishing, agreed that the question of whether vendors favor their own content was something that needed to be addressed publicly.

"This is not reality, but we can understand how someone might fear this taking place," Collins says. "It would be the single most obvious way for us to alienate all content partners who participate in EDS. With EDS, we do not favor EBSCO content."

Collins went on to say that relevance ranking in EDS is centered on subject indexing and using full text to support this detailed indexing. And in light of this approach to "leveraging deep article indexing, perhaps the only bias in EDS is placed on quality in searching and surfacing the best possible results for a given query."

Law says Serials Solutions "has been working from day one to provide as much transparency as we can."

"Part of the NISO Open Discovery Initiative's charter is to define fair linking practices to regulate and address this concern," Law says.

In particular, NFAIS and NISO are exploring the question of whether discovery services are allowing libraries to decide where they would prefer their users go to access material. Without this, a library might see a precipitous drop in usage from a vendor and not realize it is because a discovery service was driving that traffic elsewhere without the library's understanding.

A draft of the NFAIS code says it is an "obligation" of the discovery service "to provide links to Subscribers/User's authorized platform(s) of choice directly from the search results retrieved by using the service."

Heterick says JSTOR is carefully monitoring the situation to ensure that expressed commitments to neutrality are fully maintained.

Usage statistics

Timely, standards-compliant usage and referral statistics can help address some of the concerns Heterick raised and were high on the wish list of content providers and librarians. NFAIS and ODI have identified such reports as an important protocol.

"At the highest level, libraries rarely have the granular usage data needed to understand and evaluate user workflows to see the impact that a discovery service is having on usage," Heterick says. "A high percentage of the usage of many licensed electronic resources, such as JSTOR, is provided via discovery driven from Google and not a library-provided system, and we are trying to work with libraries to help them explore these issues with institution-specific evidence."

The discovery vendors do provide usage statistics, but the formats vary from vendor to vendor and may not give certain

"Participation in a discovery service may not be to the benefit of all content providers, and such participation may be of the least benefit to A&I content that, because of branding policies and specific ranking algorithms, can have a low profile in some of the search results."

—**Bonnie Lawlor, executive director, NFAIS**

content providers adequate credit for their contributions, which ODI and NFAIS are trying to address.

Penka of OCLC says ODI should help establish "a framework of value across all parties."

"That will give libraries, publishers, aggregators, and discovery service providers a common way to measure both exposure and access," Penka says. "The goal should be to help everyone better understand the utility and purpose of materials and how they're discovered."

Breeding, the ODI cochair, agreed with Heterick that it is important to know what portion of the library's resources can be attributed to the discovery service versus other channels, such as Google Scholar, since such metrics provide an indicator of the value of the discovery service. The focus of ODI, however, is on usage that transpires through the discovery service.

He and Harvard's Morse both say that a key issue is to be able to measure usage that not only can be attributed to the discovery service but also tracked to the content component that triggers a user's access.

Morse says that usage typically is measured in consumption of the resource itself, but there are no standardized metrics for the use of index data that leads a person to a particular underlying resource. For example, libraries could be at risk if metadata from a database led a researcher to a resource, but the researcher is sent to a different full-text source for fulfillment.

"This 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 that was consumed, albeit invisibly, by the researcher," Morse says.

Without greater transparency and more readily available information such as this, even on such things as better documentation about how best to structure and deliver metadata, librarians and content providers say it's difficult to know what's optimal.

Relevance rankings

The transparency needs extend to relevance rankings, which determine which results will float to the top and are a critical element in a given service's ultimate success or failure—along with such factors as the overall size of the index, the ability to easily include local library collections from specialized repositories, and the user interface.

But the proprietary algorithms remain a mystery to outsiders who, nonetheless, have important questions, such as: Is relevance determined based on the number of times a search term appears in a text and whether it appears in a title? Is a full-text article with many mentions of a term ranked more highly than a citation and abstract in which the term may only appear once or twice?

"There's talk of transparency, but what we've found in discussions with librarians is a desire to understand why any particular item did or did not show up in the results…a very difficult task even for the search team that engineered the algorithms," says Law of Serials Solutions. "What remains crucial is delivering the best answer to the researcher."

Walker of Ex Libris says discovery providers should share some information with content providers on how their content is ranked, as is recommended in the draft NFAIS Code of Practice, but there are also proprietary concerns.

"I see ranking as a key potential differentiator as is the case with content aggregators," Walker says.

EBSCO has published details on its website about how its ranking works (the greatest weight is given to matches on subject headings, to which Collins has alluded). While subject indexes are a key component, they are just one way people search for information, says OCLC's Penka, which also provides some details of its relevance ranking (terms in author and title fields are weighted first).

"If PsycINFO were included in the discovery services, users could believe that they are actually searching the database, when, in truth, they would only be scratching the surface. We do not wish to contribute to such a misperception."

—Linda Beebe, senior director, PsycINFO

"The more types of data we bring to the table, the more ways we can provide different 'flavors' of relevance," Penka says. "The richer the supplied data, the better experience we can provide," which he says is important to those working in specific disciplines or subject areas.

But content providers and librarians both expressed the hope that more information about the "flavors" of each service's relevance algorithm would become available.

"There is no reason for discovery providers to share the specifics of their relevance algorithms, but there is every reason for them to describe what kinds of data are treated as more important," says Ken Varnum, systems manager for the University of Michigan Library and a member of the NISO working group.

Varnum says figuring out what items in a library's holdings

are accessible in the discovery service and the extent to which they are covered is still the biggest challenge, but that understanding what goes into the relevance ranking is important.

"It is far harder to create good searches when you don't know what is being searched or how," Varnum says. "Understanding what goes into the relevance ranking, even if the specifics are kept secret, is important. Having an understanding of what is searched will make librarians feel more comfortable with the product."

Varnum says the absence of a common vocabulary may be as important a consideration as the lack of transparency.

"We have found even within our NISO ODI working group that librarians, A&I providers, and discovery providers are often talking across one another when we're describing our respective needs and offerings," he says.

One deliverable from ODI is a glossary of terms so that all parties will refer to the various entities and concepts in the same way.

A&I and discovery

A&I providers are special creatures in the discovery landscape.

These decades-old services are the progenitors of article discovery, often taking much care to create subject-specific indexing terms, based upon proprietary thesauri and bibliographic rules, which are then applied to article records.

Their content is provided via sophisticated search interfaces that these providers say offer more versatile features as well as more precise and thorough results than any discovery service.

All the discovery providers say they fully appreciate how access to A&I databases is critical for deep and serious research and a key differentiator for libraries in their struggles to compete against search engines like Google.

Nevertheless, many A&I providers are wary when it comes to the discovery services and decline to add their bibliographic databases to the services, fearing doing so could threaten their lifeblood. A&I providers, also, are the focus of some intense competitive jostling in a landscape where the balance of interests is already quite delicate.

"Participation in a discovery service may not be to the benefit of all content providers, and such participation may be of the least benefit to A&I content that, because of branding policies and specific ranking algorithms, can have a low profile in some of the search results," says Lawlor of NFAIS. "I know of several A&I providers who have chosen not to participate currently in a discovery service for that very reason."

"A major repercussion is that a database could be perceived as not providing relevant hits and a subscription could be canceled as a result," says Lawlor. "The same can happen if branding is lost or obscured and users do not know who is providing the content that they find."

Breeding was sympathetic to the A&I providers' concerns.

"The producers of these products work hard to ensure that they work well when used through their own interfaces," Breeding says. "When the data that underlie the A&I products are incorporated into a broader discovery service, it may or may not appear with the kind of weighting that would apply within the original interface, and it may not be clear to the researcher that the citation came from their product."

Breeding says part of the ODI group's work will be to foster a better business environment.

"The model of index-based discovery should not subvert the interests of any of the stakeholders," he says.

At the ALA conference in Anaheim, for example, Serials Solutions announced it was making tailored usage statistics available to ensure such content was valued appropriately, which would supplement other existing features, such as a database recommender.

Many A&I providers, nevertheless, remain convinced that the best decision for their business and for academic research itself lies in either remaining completely apart from discovery services or engaging them in a limited way.

"It is not clear to all those who do participate that there has been a positive impact on their business," Lawlor says.

Lawlor estimated that there are 61 A&I providers that are members of NFAIS (including those with their own platforms and/or that also offer full text), and she says that based on members' comments "a minority of pure A&I providers participate."

Among humanities database providers, the MLA International Bibliography is a notable exception, participating in all the services, but NFAIS's Mackenzie says "they were the first to jump in all the way and possibly the only one."

For example, the American Psychological Association (APA) was among the early publishers providing full-text content (PsycARTICLES, PsycBOOKS, and PsycCRITIQUES) to all four major services (as well as three open web services); however, the organization has declined to add its bibliographic databases, including PsycINFO, to any of them.

"If PsycINFO were included in the discovery services, users could believe that they are actually searching the database, when, in truth, they would only be scratching the surface," says Linda Beebe, a member of the ODI working group and the senior director for PsycINFO. "We do not wish to contribute to such a misperception."

APA does, however, make PsycINFO fully searchable via EDS as a result of some unique leverage EBSCO has in this space. EBSCO has a significant A&I business, befitting its emphasis on subject indexing and its acquisition of the H.W. Wilson Company a year ago. Many third-party A&I providers' content is also subscribed to via EBSCOhost, and EBSCO has been able to leverage that preexisting relationship to allow for searching of this hosted content in its discovery service (via "platform blending").

"EDS is the only one that is integrated with a delivery platform," Beebe says. "We can be assured that only authenticated users who have licensed access to the content are seeing the PsycINFO record, and we can be assured that it is stimulating use, not substituting for it."

Lawlor, of NFAIS, says in addition to authorized user access, branding, ranking, and the lack of informative usage reports, there has also been concern about coverage reports provided by some discovery services ("Discovery service X covers Y percent of database Z"), which Lawlor says have "caused a lot of confusion in the business."

"I know that some NFAIS members did question the validity of the numbers and requested that marketing pieces about their specific database be removed," she says.

In the end, there was hope that the combined efforts of NISO and NFAIS may help dispel some of the difficulties and tensions.

"We are hopeful that the ODI will provide a forum for discussion, identify best practices, and develop a common framework that will support maximum value to all involved," says Penka of OCLC. ■