Proposed NISO Work Item:
Open Discovery Initiative: Standards and Best Practices for Library Discovery Services Based on Indexed Search

Proposal for Consideration by the NISO Voting Membership
Approval Ballot Period: September 2 – October 2, 2011

The following proposed work item is submitted by:
Jenny Walker, Consultant, on behalf of Open Discovery Initiative meeting attendees
(See Schedule 1)

Proposal Last Modified: August 10, 2011
Approved by the NISO Discovery to Delivery Topic Committee August 23, 2011

Work Item Title:
Standards and Best Practices for Library Discovery Services Based on Indexed Search
(proposed short title: Open Discovery Initiative)

Background and Problem Statement:
The Open Discovery Initiative (ODI) aims at defining standards and/or best practices for the new generation of library discovery services that are based on indexed search. These discovery services are primarily based upon indexes derived from journals, e-books and other electronic information of a scholarly nature. The content comes from a range of information providers and products—commercial, open access, institutional, etc. By indexing the content in advance, discovery services have the ability to deliver more sophisticated services with instant performance, compared to the federated search techniques used previously. Libraries increasingly rely on index-based discovery services as their strategic interfaces through which their patrons gain access to the rapidly growing breadth of information that may be available to them. Libraries expect their uniquely licensed and purchased electronic content to be made available within their discovery service of choice. Further, they require comprehensive and clear representation of each category of content in the discovery service. In the realm of electronic material, these discovery services depend on the cooperation of information providers to provide access to metadata, and increasingly also to the full-text of information resources, in order to create effective indexes.

Several major discovery products have been released to the market in the past two years based on this new model of indexed search. Their indexes have been built based on private agreements and ad hoc exchange methodologies between information providers and discovery service creators. When licensing content, libraries need a clear understanding of the degree of availability of that content in their discovery service of choice. For example it is often not clear what articles are available, whether they will be indexed in full text, by citations only, or both, and whether the metadata derives from aggregated databases or directly through the full-text. Similarly for e-books it is not clear whether only the metadata is indexed or also the full-text.
Given the growing interest and activity in the interactions between information providers and discovery services, a group of interested parties has taken the initiative to explore ways to improve the ecosystem of library discovery services. This group is interested in establishing a more standard set of practices for the ways that content is represented in discovery services and for the interactions between the creators of these services and the information providers whose resources they represent. This informal initiative would benefit from a more formal process and the creation of standards and/or best practices.

Preliminary conversations regarding the open discovery initiative among Marshall Breeding, Oren Beit-Arie, and Jenny Walker led to an invitational meeting convened at the ALA Annual Conference in New Orleans on Sunday June 26, 2011. Invitees included individuals from the major stakeholder groups: libraries, information providers, and discovery service providers as well as representatives from NISO, the relevant standards organization, and also from NFAIS, who are in discussions with their members about discovery services (full attendance list attached as Schedule 1). The meeting was intended to gauge interest in exploring these issues in more formal fashion. It included brief statements from the perspectives of each of the stakeholder groups followed by general discussion. (A full meeting summary is attached as Schedule 2). There seemed to be general agreement among the majority of those present regarding the need for a standard way for information providers to provide content to discovery service creators, for there to be a consistent vocabulary regarding all the elements involved, that there be clarity in the business rules that apply to the content once indexed, and that there be clear descriptors regarding the extent of indexing performed for each item or collection of content and the level of availability of the content. Other areas of interest include a standard exchange of data describing what rights to the content apply within the discovery service and a standard approach to exchanging data in support of usage reports.

A follow-up conference call with NISO representatives confirmed interest in their participating in the process. NISO invited the Open Discovery Initiative to develop a proposal to the D2D Committee.

Statement of Work:

1. Project Goals:
   i. Provide effective means for libraries to assess the level of participation by information providers in discovery services, to evaluate the breadth and depth of content indexed by discovery services and the degree of availability of content to different institutions and to different users with a high degree of granularity, to include titles and issues and not just aggregations.
   
   ii. Develop a set of best practices that can help streamline the process by which information providers work with discovery service vendors. It will be important to first develop a common vocabulary that will ensure a better understanding by all parties as to what is included in content provided to discovery service vendors. Areas that could benefit from best practices include the agreements between information providers and discovery service vendors, the format of the data to be delivered for indexing (this will
likely vary depending on the nature of the content and of the provider), the method of delivery of the data and ongoing updates, and the issue of access rights to content.

iii. Streamline the interaction and communication between vendors and information providers. This reflects, for example:

a. Activation: how to best reflect libraries’ rights and subscriptions to content; ability to identify consistently the institutional subscribing unit and also the content purchasing units.

b. Unique and persistent identification of the customer (i.e. the relevant unit in the library that subscribes to content).

iv. Define models for fair linking from the discovery service to the publisher content.

v. Determine what usage statistics should be collected, for whom, and how these should be disseminated. This is a possible place to engage with COUNTER.

2. Beneficiaries:

The key beneficiaries of this initiative would be:

i. Libraries would be better able to evaluate publisher content in terms of the degree of its availability through their library discovery service of choice. In addition, libraries subscribing to discovery services would be better able to evaluate such services in terms of the degree of coverage of their content.

ii. The information providers—publishers and aggregators—contributing content to discovery services will be able to address a growing need in libraries to optimize use of the content to which they subscribe. ODI would standardize the expectations both of the library and of the discovery service provider.

iii. Discovery service vendors would benefit from streamlining of the interactions with information providers as indicated above. Discovery service vendors play a key role in serving both the libraries and the information providers, paying close attention to the needs of both stakeholder groups.

3. Specific Deliverables and Objectives:

i. Policy/business
   a. Define levels of compliance
   b. Content: Define the depth and breadth of content available to Open Discovery indexes
   c. Rights/authorization; Define level of access to indexed content (restricted to subscribers vs. open to all users)
   d. Display: Define a minimum standard of what can be displayed from information provider content and metadata (and how to protect information provider intellectual property rights)
   e. Linking: Define the incorporation of fair linking
ii. Technology
   a. Rights: determination and sharing of authorizations when content discovery is restricted
   b. Formats: content, and methods of updates
   c. Usage statistics gathering and reporting

4. Process:

Suggested approach to the process for achieving the project goals:

i. Appoint a Working Group. Note that a number of attendees at the ALA ODI meeting volunteered to work further on this initiative.

ii. Identify related standards and be critical of these in light of discovery services, but do not attempt to replicate. Some known related initiatives include the NFAIS group working on best practices that will cover the rights and obligations of all those involved in a discovery service relationship, OpenURL, OAI-PMH, and COUNTER who are currently engaged in the creation of a new version (v4) of their Code of Practice.

iii. Define a common vocabulary for use by all stakeholders

iv. Review issues and recommendations from the initial meeting of the ODI group; consider a wider consultation and agree on priorities for the Working Group.

v. It is anticipated that there would be at least two groups within ODI—one group addressing business and policy issues and the other group addressing technical issues.

vi. Develop standards/best practices based on determined priorities.

5. Partners and Participation:

The current composition of the ODI group submitting this proposal includes members from the following stakeholder groups, which all have an interest in this initiative:

   i. Libraries and library groups such as ICOLC
   ii. Information providers and representative groups such as NFAIS
   iii. Discovery service vendors

Additional participants that would likely add value to the group include members from COUNTER, JISC.
Schedule 1

ALA ODI meeting attendees

Hosts
Marshall Breeding, Vanderbilt University and independent consultant
Oren Beit-Arie, Ex Libris
Jenny Walker, Ex Libris

Attendees:
Kristin Antelman, North Carolina State University and OLE
Laura Krier, California Digital Library
Joe Lucia, University of Villanova and VUFind
Judy Luther, Informed Strategies
Nettie Lagace, NISO
John McDonald, Claremont Colleges (also on COUNTER board)
John Meador, University of Binghamton
Laura Morse, Harvard University
Michelle Newberry, FCLA and ICOLC
Kari Paulson, EBL
Roger Schonfeld, Ithaka/JSTOR
John Tagler, American Association of Publishers
Mike Teets, OCLC

By phone:
Renny Guida, Thomson Reuters
Bonnie Lawlor, NFAIS
Schedule 2

Open Discovery Initiative

Minutes of the inaugural meeting held at ALA, Sunday June 26th 2011

Summary

1. Short overview of the current landscape and background to the Open Discovery initiative (Marshall Breeding)

   a. The current environment involves decoupling of the back-end library automation systems and the front-end interfaces presented to library users for access to library collections and services.

   b. The scope of discovery interfaces is broader than what is managed in the integrated library system, to include other local repositories and digital collections and the electronic resources represented in the library subscription to information content products.

   c. Libraries—and users—have high expectations that the content available through discovery services is ‘complete’. It is hard for libraries to determine this; and there are obstacles to vendors for ensuring their services are complete. How to remove these obstacles to make discovery services easier and more powerful?

   d. Features expected by libraries for discovery services based on an aggregated index:

      i. Libraries expect to the capability for discovery services that search at the article level for e-journals.

      ii. Immediate search results, relevancy ranking, faceted browsing

      iii. Expectations are rising toward the capability to perform full-text searching of articles and not just citations

   e. The method of search has evolved from metasearch to aggregated indexes. Metasearch, based on middleware that perform real-time queries against multiple targets, is no longer perceived as able to deliver adequate search performance.

   f. A continuation of the types of structured relationships between library technology providers and content providers: OpenURL, COUNTER, SUSHI, NISO MetaSearch Initiative

   g. Current discovery products rely on an architecture based on an aggregated index

      i. Article-level indexing
ii. Publishers and other information providers allow harvesting of content for purpose of indexing by discovery systems

iii. Discovery systems do not republish articles, but use the harvested content only for indexing

iv. Discovery systems will provide links to the appropriate copy, according to a library’s active subscriptions, on the publisher’s server

v. Snippets of the article text may or may not be displayed depending on copyright or business terms

vi. Harvesting from information providers to discovery servers happens mostly through large batch transfers, with incremental updates, according to a mutually agreed frequency.

vii. The content provided by the information provider to the discovery service might include brief citations, controlled vocabulary terms, abstracts, full text, or other data or metadata.

viii. The aggregate indexes generally aim to represent the totality of the universe of potential resources available; search results may be filtered according to a profile of a library’s actual subscriptions.

h. A number of major discovery services have been produced that follow this model, including commercial products such as Summon from Serials Solutions/ProQuest, EBSCO Discovery Service, Primo Central, WorldCat Local, as well as European projects such as Ting Concept and Summa.

i. The creation of these indexes has taken place through individual agreements between information providers and discovery service providers. The mechanics of the transfers, the type of data involved, and the business terms on how the data can be used has taken place through pragmatic, ad hoc arrangements.

j. Industry would now benefit from a more standard approach. A standard protocol will make it easier for information providers to make content available for any discovery service provider. Discovery service providers will benefit from a uniform way to deal with the hundreds or thousands of potential information providers.

k. Interest focused on improving discoverability of content; should increase value of the resources. No interest in usurping the role of publishers through providing alternative paths for library users to their content.

l. No intention to replicate the work of other initiatives.

m. Overall goal is to have discovery services include all collections; with profiles per library subscriptions.
2. Benefits and values for stakeholders

   a. The library perspective (Marshall Breeding, Vanderbilt University)

      i. Vanderbilt is a user of a discovery service; Marshall thinks and writes about discovery services.

      ii. Libraries are increasingly investing in discovery services (more and more important as higher percentages of budget are e-spend). Libraries have high expectations that discovery services will improve access to library collections.

      iii. Discovery services are much stronger than previously available tools such as online catalogs, or metasearch engines. They offer a single search box that addresses all types of content, including books, electronic resources, and digital content.

      iv. Need to maximize the value of library subscriptions. In some libraries 85% of budget is spent on e-resources. Libraries do not get sufficient value from this unless their users find and use the content. Some content providers are also discovery service providers which can introduce some confusion for libraries unless the arrangements between these two businesses are transparent in terms of how content is positioned.

      v. Initial feedback on discovery services if good but a better ‘before and after’ picture of this is still needed.

      vi. Libraries expect all resources to be indexed: increasingly difficult to deal with remnants of collection not represented in the discovery index. They also expect clarity on what is covered in the index: citations, abstracts, controlled vocabulary terms, full-text; and on the currency of discovery index--what is available in discovery system should closely match that of the native interface. The coverage of content by a particular discovery service is becoming a collection development tool.

      vii. Libraries expect to have flexibility between the use of discovery services and native interfaces as appropriate for different users and research scenarios.

      viii. Discussion:

            • Discovery services need to accommodate all user levels--undergraduates and postgraduates.

            a. General view that libraries tend to focus discovery service deployment at undergraduates, though Discovery is certainly
relevant (and used) by grads and researchers (especially in inter-disciplinary context)

b. Postgraduates are also expected to use discovery services and they in particular will want to know what content is included in the discovery service and what is excluded. Postgraduates also want access to the specialist tools provided by the content providers.

- It is important for libraries that discovery services can offer users a mix of what they have (through library or freely available) and of what they might have access to eg through POD, docdel, ILL etc.

- Focus on the basic components at this stage. This initiative should focus on the core components of what is needed to make content available to a discovery index (back office activities), in a consistent and comprehensive fashion, and not on specific functionality of discovery services eg relevance ranking, though in some cases the issues should be noted and acknowledged.

- It was suggested that libraries won’t subscribe to something that’s not in their Discovery service.

- COUNTER board is evaluating the impact of Discovery system use and how best to incorporate these stats in the COUNTER Code of Practice.

b. A discovery vendor (Oren Beit-Arie, Ex Libris)

i. Discovery vendors sit in the middle. They develop discovery systems that serve libraries who select and subscribe to content from multiple information providers. Libraries needs, as presented by Marshall, translate to Discovery requirements, and hence to goals of the discovery vendors. Libraries typically have base requirements in common but have a lot of varying further requirements. There is no ‘one size fits all’; libraries need to customise discovery services to fit their context and specific needs

ii. We hear from libraries a very clear need: to maximize the utilization and their investment in subscription content by enabling it in their discovery solution. Many libraries are also keen to explore additional models of access to content (“just in time”), in addition to ownership and subscription (just in case) – which impacts discovery.

iii. While we serve the libraries, we must also best represent -- and when needed, protect -- the rights of the information provider and be very aware
and sensitive to their business needs, as the eco-system has to be well balanced across stakeholders.

iv. Also recognizing the fact that there are different information resources (books, journals, audio, video, reference, newspapers etc) and different information providers (primary publishers, secondary publishers, journal publishers, ebook publishers, publishers of reference, of images, video etc. Also aggregators). They may have different needs (as well as legal restrictions). Discovery vendors need to treat the content ‘fairly’ (but what does ‘fair’ mean?).

v. The core issue: While libraries transition and expand from Meta-search (MS) to Indexed-search (IS), there is a significant discrepancy between what’s available and discoverable in each method.

vi. Today, IS is still governed by proprietary deals between discovery vendors and information providers, which results in a blurry and inconsistent “eco-system”, that under-serves libraries and users, and is becoming increasingly hard for libraries to evaluate and measure.

vii. To use an analogy, consider Linking and OpenURL. The current state of Indexed-Search resembles a state in which some Information Providers would support OpenURL linking in some of their products/content (but not in others), would enable these OpenURLs to some Link Resolvers (but not others) and would include on the OpenURL only partial metadata (not necessarily sufficient for optimal resolution) – all governed by proprietary agreements between the Information provider and the Link Resolver provider.

viii. In addition, as the IS gets richer and better (with more content) – the content that is left outside becomes more problematic for discovery by end users

ix. The goals for this initiative, from the perspective of a discovery vendor, are:

- Ensure that content selected and licensed by libraries will be enabled to discovery regardless of the method of discovery utilized by the library (e.g. MS vs. IS), or the discovery solution of choice.

- Define a common vocabulary and a standard mechanism to evaluate the availability (and “compliance”) of content to Open Discovery, so that libraries will fully know what to expect in terms of discoverability of the data they subscribe to.
• Deal with the issue of rights, authorization and restrictions: should content discovery be restricted based on subscription, or open to all library users?

• Related to -- and impacted by -- this is the trend of “Just-in-time” collection development (PDA, PPV, etc) that requires a discovery universe that is bigger than what the library subscribes to.

• To ensure that we streamline the interaction and communication between vendors and information providers, that reflects, for example:
  a. Activation: how to best reflect libraries subscription to content.
  b. Unique and persistent identification of the customer (i.e. the relevant unit in the library that subscribes to content).

• Also, how best to measure effectiveness of discovery and to share usage information with the information providers. (Consider COUNTER-type reporting)

d. Ensuing Discussion:
  • Technical barriers to the aggregation of data have largely been removed in recent years. We are now struggling with issues of policy and practice. We should not be fettered by previous standards nor by the current status of discovery services. Some standards were developed to address problems that no longer exist. Is OpenURL one of these, given that the content is now held in the aggregate? Must remember that the publisher still serves the content for discovery services! However, acknowledgement should be given to the fact that there will likely always be multiple access points to the content; what does this mean for the infrastructure?

  • What does this new library discovery world mean for A&I. Need to take cognisance that some data is open while some is not, typically A&I. What is the impact on A&I when discovery services index the full text from the publisher and then A&I for the same item? Need a trust layer for discovery services. If authorisation is needed the discovery service should handle it—index it but control the display.

c. The information providers (Bonnie Lawlor, NFAIS)
  i. NFAIS members have been discussing discovery services for 18 months. 3 of the 4 main discovery providers are NFAIS members. The mission of an A&I provider is to make sure that users find the information. In January 2010 members told NFAIS administration that they were interested in these
new discovery services. NFAIS created a survey to get member feedback (results are on NFAIS web site), solicited feedback, analyzed pros and cons. Created a subgroup - issues impacting content providers - had a roundtable meeting hosted by Thomson Reuters. Discussed both the opportunities offered to content providers by discovery services as well as the concerns that such services raised. As a result of the discussion a list of ten questions emerged. Each of the four discovery vendors provided a 90-minute webinar to answer the 10 questions from NFAIS members.

ii. NFAIS is developing a discovery service code of practice to identify rights and responsibilities of all players. Had a kickoff meeting Wednesday June 22nd and will get underway in earnest on July 11. All issues raised by others today are seen as critical. Primary publishers - nothing to lose; A&I only - concerned about authorization, ranking, branding and visibility.

iii. NFAIS members (information providers) are interested in participating in the discovery services, but there is a lot of confusion going on which translates into a lot of concerns. Bonnie sees parallels between this new delivery channel and the online services and Gateways such as EasyNete and Telebase in the 80s. Who actually provides content, how does the user know what is the breadth and depth of the content that they receive, etc? How to maintain brand identification, what is the ranking of results (each system has a different algorithm) and there is need to link to provider’s own database.

iv. Discussion

• Renny Guida, who participated from the outset in the NFAIS discovery discussions, confirmed Bonnie’s report as an accurate reflection of their group’s discussions. He added that one critical issue for Thomson Reuters (and others?) is how the value of individual subscriptions get translated through the discovery service to administrators in libraries who need to see the value they get through the subscription.

• NFAIS is not a technical standards group; their focus is on policy and business issues related to discovery services. Bonnie would anticipate an effective way forward to be a collaborative venture between NFAIS and NISO as per the current supplemental journal data work. Bonnie further noted that NFAIS does not dictate business models; her members all make individual business decisions to suit their needs. However, NFAIS does suggest best business practices in different areas.

• Branding.
a. NFAIS members still want the brand recognition to drive usage and renewals. Discovery providers have not yet been able to show how effective these systems are – they drive traffic, but not necessarily customers.

b. Roger Schonfeld - branding implies the broader question of relationship with the user. Surveys show that the relationship between the user and the library is in freefall. Need to understand this environmental change. User relationship is migrating from the library to the publisher and potentially to discovery service. Feels like an important change.

• ‘Good enough’?

a. Bonnie reiterated that NFAIS members have not walked away from quality. Users expect quality, but also an efficient and pleasant experience from start to finish part of the whole information discovery process, period. Need a good seamless experience. This started with Google.

b. John Meador—Users want “good enough” search and care more about the user experience. However, libraries still care about the data. If you have Summon you can’t have the EBSCO data—and the reverse is also true. Because of competition we get silos and lock-in publishers. Publishers want exposure.

c. Transparency wanted re the breadth of metadata.

• Ranking of content

a. Is there such a thing as vendor-neutral relevancy? Probably not as discovery vendors compete on this issue and some libraries may specifically want control over this to promote one resource over another.

b. Kari Paulson – There are concerns with Google – how are things ranked – how do things get pushed to the top. Discover layer as hosting is problematic – need neutrality in search results. Potential for conflict if the discovery layer is also the owner of the content. Need declaration of policy, don’t want manipulation.

• A&I vs primary content
a. Will there be a time for tension between primary publishers and A&I vendors – if A&I choose not to include their content in these interfaces – deals will be made directly. Either one way or the other will get forced out.

b. Bonnie confirmed that such a discussion was not taking place at NFAIS.

c. Renny Guida reminded the group of the extent of the special indexing that is done; and that’s why libraries subscribe to these. There should be a way they are used with attribution back to the provider that did the work.

d. Joe – use case for discovery – libraries look for points of failure in these new tools. The recourse is often to recommend the old specialized tools. Is it legitimate in terms of more advanced usage or is it self preservation? In any event, special indexes are still assigned great value and I don’t think they are under threat. There is a complex environment with place for expert databases. Some present disagreed with this, considering the indexes to be not worth the cost.

- Silos being created

a. Discovery services are intended to remove silos but in fact silos are being created by the competitive nature of discovery service providers and particularly those that are both information providers and discovery vendors eg EBSCO content is not in Summon, ProQuest content is not in EDS; EBSCO withdrew content from Primo Central.

3. **Suggested areas of focus – policy/business issues as well as technical issues.**

a. Take a pragmatic approach: Identify a couple of main issues to focus on.

b. Categorize the issues and then prioritize. Identify related standards and be critical of these in light of discovery services; but do not attempt to replicate.

c. Categories:

i. Policy/business

   - content - depth and breadth of content available to Open Discovery indexes)
   - rights/authorization – define level of access to indexed content (restricted to subscribers vs. open to all users)
• Display – what can be displayed (and how to protect information provider IP)
• linking - how to incorporate fair linking

ii. Technology
• Gathering of data and updates/deletions
• Distribution methods
• Authorization/rights
• Usage

d. Suggested way forward:

i. Start on the issue of content, find the vocabulary to describe availability of content, look at ways we can share level of availability of content in the context of discovery. Maybe providers participate fully, or don't participate at all; maybe partly.

ii. Method of communicating rights/subscription information. Ability to identify consistently the institutional subscribing unit and also the content purchasing units.

iii. Usage - what should be collected and how should it be disseminated

4. Next steps

a. Time was short and there was little time for discussion on next steps.

b. All those in attendance indicated their ongoing interest. A Google group would be set up that would include those in attendance at this meeting plus those invited who were unable to attend but had expressed great interest. A survey will be created that will allow individuals to indicate priorities.

c. Continue to refer to this initiative as “Open Discovery Initiative’

d. Mike Teets to suggest a common vocabulary to help stakeholder discussions around discovery services eg tool vs service vs platform vs system vs layer.

e. Laura Morse and John McDonald volunteered to play leadership roles in taking forward this initiative.

f. Nettie Lagace to report back to NISO Exec and further discussions would be held with NISO.