

Open Discovery Initiative: Promoting Transparency in Discovery



PROMOTING TRANSPARENCY
IN DISCOVERY

*A Recommended Practice of the
National Information Standards Organization*

Approved: June 25, 2014

Prepared by the
Open Discovery Initiative Working Group

About NISO Recommended Practices

A NISO Recommended Practice is a recommended “best practice” or “guideline” for methods, materials, or practices in order to give guidance to the user. Such documents usually represent a leading edge, exceptional model, or proven industry practice. All elements of Recommended Practices are discretionary and may be used as stated or modified by the user to meet specific needs.

This recommended practice may be revised or withdrawn at any time. For current information on the status of this publication contact the NISO office or visit the NISO website (www.niso.org).

Published by

National Information Standards Organization (NISO)
3600 Clipper Mill Road
Suite 302
Baltimore, MD 21211
www.niso.org

Copyright © 2014 by the National Information Standards Organization

All rights reserved under International and Pan-American Copyright Conventions. For noncommercial purposes only, this publication may be reproduced or transmitted in any form or by any means without prior permission in writing from the publisher, provided it is reproduced accurately, the source of the material is identified, and the NISO copyright status is acknowledged. All inquiries regarding translations into other languages or commercial reproduction or distribution should be addressed to:

NISO, 3600 Clipper Mill Road, Suite 302, Baltimore, MD 21211.

ISBN: 978-1-937522-42-1

...

Contents

Foreword	v
Section 1: Introduction	1
1.1 Purpose and Scope	1
1.1.1 Purpose	1
1.1.2 Scope	2
1.2 Terms and Definitions	4
1.3 Stakeholders	7
Section 2: The Evolution of Discovery and Delivery	9
2.1 Catalogs and Indexes	9
2.2 Abstracting and Indexing Resources	9
2.3 Metasearch Utilities	10
2.4 Enhanced Library Catalogs	10
2.5 Index-Based Discovery Services	11
2.6 Related Initiatives	11
2.7 Information Gathering and Results	13
Section 3: Recommendations	15
3.1 General Recommendations	15
3.2 Recommended Best Practices for Content Providers	15
3.2.1 Metadata Elements Provided to Discovery Services	15
3.2.2 Metadata Elements Provided by Content Providers to Libraries	22
3.2.3 Disclosure	22
3.2.4 Technical Formats	22
3.3 Best Practices for Discovery Service Providers	23
3.3.1 Discovery Service Content Listings	23
3.3.2 Linking	25
3.3.3 File Formats and Methods of Transfer	26
3.3.4 Usage Statistics	27
Section 4: Recommended Next Steps	31
4.1 Collaborative Discussion	31
4.2 Application Programming Interfaces [APIs]	31
4.3 Managing “Restricted” Content in Discovery Services	31
4.4 Reporting on Discovery Service Content at a Collection Level	31
4.5 On-Demand Lookup by Discovery Service Users	32
4.6 Research and Analysis to Monitor Discovery Services for Fair/Unbiased Linking	32
4.7 Further Interaction with COUNTER	32
Appendix A: Survey Summary	33
Appendix B: Content Provider Conformance Checklist	34
Appendix C: Discovery Provider Conformance Checklist	35
Bibliography	37

Document ID: NISO RP-19-2014

Foreword

About This Recommended Practice

The ODI was initiated in June 2011 by Oren Beit-Arie of Ex Libris, Jenny Walker, independent consultant working on behalf of Ex Libris, and Marshall Breeding, a library industry consultant (previously at Vanderbilt University) who, at the 2011 ALA Annual Conference in New Orleans, invited senior industry managers to meet and explore areas of mutual interest related to library discovery services. Following this initial meeting, a proposal was submitted to NISO; later in the year, the NISO Discovery to Delivery Topic Committee accepted the proposal as a new NISO work item. This Recommended Practice is the outcome of that project.

The NISO Open Discovery Initiative (ODI) group started its work in early 2012 to define best practices for the new generation of library discovery services. These services use an aggregated central index to enable searching across a wide range of library related resources—both licensed and free—from multiple providers. They also offer more sophisticated capabilities and faster performance than those provided by systems relying on federated search technologies. Several major discovery products, based on centrally indexed search, have been released to the market since early 2009 and have been widely deployed in libraries globally.¹

This investment by libraries in centrally indexed search as a primary means for users to discover and access licensed content brings with it new requirements for industry practices in a number of areas. To facilitate improved communications and clarity, the ODI established the following goals:

- Create ways for libraries to assess the level of content providers' participation in discovery services
- Help streamline the process by which content providers work with discovery service vendors
- Define models for fair or unbiased linking from discovery services to publishers' content
- Determine what usage statistics should be collected

Libraries expect their entire collection, including licensed and purchased electronic content, to be made available through the discovery service of their choice. When acquiring licensed content, libraries expect a clear explanation of the degree of availability of that content in that discovery service.

Based on the input from a survey done early in the project (see 2.7), the ODI group agreed to develop recommended practices in the following areas:

1. Technical recommendations for data format and data transfer, including method of delivery and ongoing updates.
2. Recommendations for the communication (automated or through reporting) of libraries' rights to distribute or display specific content (e.g., restricted to subscribers versus open to all users). These recommendations are to include technical specifications on how data will be exchanged and procedural specifications regarding update frequency and other logistical details.
3. Clear descriptors regarding the level of indexing performed for each item or collection of content and the level of availability of the content.
4. Definition of fair linking from discovery service to the published content.

¹ Marshall Breeding. "Automation Marketplace 2013: The Rush to Innovate." *The Digital Shift*, April 2, 2013. Available at: www.thedigitalshift.com/2013/04/ils/automation-marketplace-2013-the-rush-to-innovate/

5. Determination of what usage statistics should be collected, for whom, and how these data should be disseminated.

Further, the Working Group agreed to develop mechanisms to evaluate conformance with the Recommended Practice.

To work towards the development of recommended practices to address the above issues, five subgroups were formed as follows:

1. Technical formats
2. Communication of library's rights
3. Level of indexing
4. Fair linking
5. Usage stats

Early in the process subgroups 2 and 3 merged.

A further goal of the ODI was to develop mechanisms to evaluate conformance with the Recommended Practice. To help libraries better understand the position of content providers and discovery service providers, these organizations can take measures to demonstrate the extent to which they conform with the recommended practices issued by the Open Discovery Initiative. We anticipate that these conformance statements will be voluntarily issued by content providers and by discovery service providers respectively. In the absence of voluntary statements, libraries can use the presence or absence of these factors to infer conformance.

NISO D2D Topic Committee Members

The following individuals served on the NISO Discovery to Delivery Topic Committee that had oversight for this work and approved this Recommended Practice:

Kristin Antelman

North Carolina State University Libraries

Pascal Calarco, Co-Chair

University of Waterloo Library

Lucy Harrison, Co-Chair

Florida Virtual Campus

Juli Marsh

The Library Corporation (TLC)

Peter Murray

Lyrasis

Tim Shearer

University of North Carolina Chapel Hill Libraries

Chris Shillum

Reed Elsevier

Christine Stohn

Ex Libris, Inc.

ODI Working Group Members

The following individuals served on the NISO Open Discovery Initiative Working Group that developed and approved this Recommended Practice:

Marshall Breeding, Co-Chair
Independent Consultant
(formerly with Vanderbilt University)

Jenny Walker, Co-Chair
Independent Consultant (formerly affiliated with
Ex Libris)

Anya Arnold (through July 2012)
Orbis Cascade Alliance

Scott Bernier (from December 2013)
EBSCO Information Services

Jamene Brooks-Kieffer
University of Kansas Libraries (formerly with
Kansas State University Libraries)

Sara Brownmiller
University of Oregon Libraries

Lettie Conrad
SAGE Publications

Michael Gorrell (through November 2013)
EBSCO Information Services

Jeff Lang
Thomson Reuters

John Law
ProQuest

David Lindahl
University of Missouri-Kansas City Libraries
(formerly eXtensible Catalog Project)

Laura Morse
Harvard University

Michele Newberry
Unaffiliated (formerly University of Florida)

Peter Noerr (through June 2012)
MuseGlobal

Roger Schonfeld
Ithaka S+R

Ken Varnum
University of Michigan Library

Aaron Wood
Alexander Street Press

Acknowledgments

The ODI Working Group would like to offer a special thanks to the following individuals for their assistance:

- SAGE Publications, especially Elisabeth Leonard, for support of the ODI Survey.
- Lucy Harrison (Florida Virtual Campus), Bonnie Lawlor (NFAIS), and Jeff Penka (OCLC), who were observers on the Working Group and provided valuable input and feedback.
- The many individuals and organizations that responded to the information-gathering survey issued by the ODI.
- Nettie Lagace and Cynthia Hodgson from NISO for their support and advice.

Trademarks and Service Marks

Wherever used in this recommended practice, all terms that are trademarks or service marks are and remain the property of their respective owners.

Copyright National Information Standards Organization
Provided by IHS Markit under license with NISO
No reproduction or networking permitted without license from IHS
Full document details: NISO RP-19 2014

Section 1: Introduction

1.1 Purpose and Scope

1.1.1 Purpose

The Open Discovery Initiative (ODI) aims to facilitate progress through exploration of relevant issues and the development of recommended practices for the current generation of library discovery services based on centrally indexed search. The domain of index-based discovery services involves a complex ecosystem of interrelating issues and interests among content providers, libraries, and discovery service creators.

This model of discovery relies on an index populated with metadata, full text, or other representations of the content items—such as journal articles, book chapters, e-books, research reports, reference sources, images, maps, datasets, AV materials, and other selected material—which a library provides to its users. The content comes from a range of information providers and products, such as commercial and non-profit publishers, universities and other research institutions, and many other types of organizations. The content of interest to ODI includes any materials that libraries would consider within their collection, regardless of the business model for acquisition or the type of license, such as commercially restricted or open access.

Several major discovery products have been released to the market since early 2009 that are based on the model of centrally indexed search—largely influenced by the Google search model and users’ expectations for a single, unified discovery experience. An increasing number of libraries, especially those that serve academic or research institutions, have invested in index-based discovery services. These products serve as one of the interfaces through which the library’s patrons gain access to the rapidly growing breadth of information that may be available to them. These discovery services play an increasingly strategic role in the way that libraries provide users with access to their collection, represent a growing segment of the library technology industry, and may become a factor in how libraries select content products. These factors draw attention to the discovery services arena for any improvements that might be gained through this Recommended Practice.

To work effectively, discovery services need to be as comprehensive as possible in their content coverage. Libraries expect their uniquely licensed and purchased electronic content to be indexed within their discovery service of choice. Further, they require comprehensive and clear representation of each category of content in the discovery service. Content items not represented in a discovery service present a challenge to libraries in how they might otherwise ensure that these materials are discovered and accessed. Libraries have an interest in knowing whether any content providers are excluded or underrepresented in any given discovery service.

The Open Discovery Initiative aims to facilitate increased transparency in the content coverage of index-based discovery services and to recommend consistent methods of content exchange or other mechanisms. Full transparency will enable libraries to objectively evaluate discovery services and to deal with daily operational issues surrounding these products.

Discovery services depend on the cooperation of content providers with discovery service creators to provide access to metadata or full text of information resources in order to create effective indexes. The inclusion of data in the indexes of the current slate of discovery services is based on private agreements and ad hoc exchange methodologies between information providers and discovery service creators. Index-based discovery can potentially benefit content providers through enhanced exposure of their materials. It also presents some concerns, such as enabling library patrons to bypass the specialized interfaces created by content providers, potentially reducing or eliminating branding and losing control in how content is presented to the end user. And, as libraries’ uptake of these services increases, the usage (and perceived

value) of publisher products can be greatly influenced by how successfully discovery services drive readers to a content providers' assets.

ODI investigated the need for standard protocols to make the transfer of data from content providers to discovery service creators. Consistent practices in the exchange and formats of data aim to lower the level of complexity as content flows through this ecosystem, mitigating technical issues that might hinder broader participation by content providers or potential discovery service creators.

Libraries need a clear understanding of the degree of exposure for the content that they have acquired as represented in a discovery service. This understanding is essential as libraries evaluate and select a discovery service and on an ongoing basis once it is implemented. Libraries require specific information on exactly which articles, databases, and other sources are represented; whether they are indexed in full text, by citations only, or both; and whether the metadata derives from aggregated databases or abstracting and indexing resources.

In the operation of an index-based discovery service, many different factors contribute to how it presents and orders results and how it connects users to content resources. For any given item of content, multiple metadata elements contributed from different content providers may be indexed by the discovery service. For a journal article, for example, its full text might be contributed by the primary publisher, citation data from the provider of an aggregated database, and abstracts or controlled vocabulary terms may be provided by yet another provider. Content providers are motivated to contribute to discovery services in order to gain more access from the patrons associated with the libraries that implement the discovery services. It is therefore important to each type of content provider that its contributions are appropriately recognized. If a record contributed by an abstracting and indexing (A&I) service, for example, leads to the selection of a full-text resource from another provider, how does the A&I service gain benefit from the discovery transaction? A subgroup of ODI on Fair Linking was established to explore and make recommendations on these issues.

The Open Discovery Initiative recognized and aimed to address perceptions regarding bias and concerns about the possibility of bias in discovery services. Special concerns surround the possibility of bias when discovery services are owned by the same corporate parent as content products or services. Concerns also arise through exclusive arrangements or other business relationships made by a discovery service with a content provider that might introduce bias. Some of these recommended practices were developed with the intent of helping discovery services providers mitigate concerns that exist in the community about conflicts of interest and other relationships that introduce bias. By explaining the nature of their business connections with related content providers and third parties alike, and affirming the neutrality of their discovery offerings, these services will be positioned to reassure both libraries and content providers about the nature of their practices.

1.1.2 Scope

In broad terms, ODI focuses primarily on the issues related to the composition of the central index associated with these discovery services and not with the design of user interfaces. The initiative does not seek to intrude into areas of proprietary innovation that distinguish each of the discovery service products.

The arena of index-based discovery spans many different issues, some of which lend themselves to a more open and standard treatment, while others remain in the realm of product development. The Open Discovery Initiative recognizes that even among the issues that might potentially benefit from its attention, some rank at a higher priority and that others may need to be addressed through possible follow-up activities.

In Scope

A primary area of interest for ODI involves the arena of content coverage of the discovery services represented in their central index. The content used to create the indexes of a discovery service comes

from a range of information providers and content products, such as commercial and non-profit publishers, universities and other research institutions, and many other types of organizations. The content of interest to ODI includes any materials that libraries consider within their collection, regardless of the business model for acquisition or the type of license, such as commercially restricted or open access.

This initiative aims to address the following questions in the realm of content coverage:

- The quantity of content that a provider makes available to discovery service providers relative to its total offerings
- The form of that content, such as whether it consists of citation-level metadata or if it also includes full-text representations
- Whether the discovery services operate in a way that results presented to the user do not favor or disfavor items from any given content source or material type
- The specific metadata fields provided within metadata records
- The specific metadata fields indexed in the discovery services
- Whether any controlled vocabularies or ontologies are included
- How abstracting and indexing products relate to discovery services
- How branding of content providers is presented in a discovery service

This initiative aims to address the following technical issues:

- The transfer mechanisms or protocols by which data are delivered from content providers to discovery service creators
- The format in which the records are delivered by content providers to discovery service creators

One area of focus deals with the definitions of the metadata delivered by content providers to discovery services, as well as the data made available to licensed customers. A perceived lack of transparency across these data flows prompted the need to develop definitions of the data points and to clarify what metadata or data elements are made available to which parties and under which conditions. For example, a content provider might allow certain metadata elements to be included in the search index for retrieval purposes, but not allow those elements to be displayed in the final user interface. Conversely, libraries might not have understanding of what elements from which full text or abstracting and indexing services are available and in which circumstances. Some elements might be displayed to authenticated users, and some not, but definitions of these distinctions are sometimes vague, if they are described at all.

Another topic concerns factors related to whether or not a discovery service functions with a bias toward certain databases or content products based on business relationships rather than user needs or library preferences. It was deemed important to propose practices that facilitate the presentation of unbiased links to a user following the execution of a search through a discovery service and in support of this objective to ensure transparency about discovery service practices.

Out of Scope

This initiative does not address issues related to performance or features of the discovery services, as these are inherently business and design decisions guided by competitive market forces.

Aspects of index-based discovery not deemed within the scope of ODI include:

- **User interface issues** – The user interface components of a discovery service may depend on the content indexes, but are out of scope for this ODI initiative.

- **Relevancy ranking** – The specific methods that a discovery service employs to order items in a search result set may fall within the realm of proprietary technologies used competitively to differentiate commercial offerings.

Further, the demands of relevancy ranking—necessary for satisfying user expectations in keeping pace with open web search applications—require continuous enhancement of supporting technologies and algorithms. As such it would be both impractical and an impediment to require that service providers continuously publish the highly dynamic and substantially detailed workings of their search relevancy algorithms. Therefore, the ODI Working Group concludes that, while disclosure of the broad aspects of a relevance ranking algorithm is encouraged, there should be no expectation that a discovery service provider would explain in any level of detail the ranking algorithms it applies. (As noted above, however, whether rankings result in the bias of results presented to the end user is within scope.)

- **APIs exposed by discovery services** – Several discovery service providers offer access to their discovery index through an Application Program Interface (API). This is a set of protocols that allows a computer program to query the index and receive search results. Some libraries build their own interfaces to the search index; others use the API to embed data retrieved from the search index into other applications, like pull lists of citations into course sites or other web pages.

The ODI Working Group determined that APIs were out of scope for this initial foundational stage of the initiative. This decision was made somewhat reluctantly based on the growing need for libraries to have access to data via API. Because the provision of an API to customers is largely a business decision by the vendor, it was felt the existence of an API could not be suggested as a best practice. Additionally, the desire to recommend standardization of API results would have added significant complexity to the workgroup's recommendations. Thus, the current committee concluded that best practices concerning discovery service APIs should be deferred until a later round of work. See Section 4.2 for recommended next steps.

- **Fair linking** – While the concept of “fair linking” was identified as a highly shared value among all stakeholders, it was determined that fully qualifying how links are presented and ordered fairly is beyond practical scope and intrudes into areas that stakeholders consider to be of a proprietary nature.
- **Protocols for data exchange** – In the arena of the technical mechanisms involved in the transfer of data between content providers and discovery services, the ODI concludes that the existing protocols and methodologies previously defined provide adequate options and that it is not necessary to create a new protocol specifically for use within the discovery services ecosystem. There has been much work and standards development in the area of file formats, schemas, naming practices, transport mechanisms, etc.

1.2 Terms and Definitions

In order to ensure consistent use of terminology, one of the early tasks of ODI was to create a glossary of definitions relevant to the field of investigation. The terms as used in this recommended practice, have the meanings indicated below.

Term

Definition

Search Models and Related Concepts

central index	The result of storing and indexing content in a central location. Disparate content sources are aggregated with consistent formatting, indexing, and ranking algorithms. Also referred to as <i>centrally indexed search engine</i> .
federated search	Method for searching multiple disparate content sources with one query. Federation of searching involves restructuring queries to appropriate forms and sending them to multiple remote search engines. Results are coordinated and displayed to the user.
metasearch	Method for searching multiple disparate content sources with one query. Metasearch is sometimes synonymous with federated search; however, metasearch can also describe searching across disparate content sources stored at a single location.
full-text search	Method for searching the text of the works rather than metadata about the work.

Data Definitions

full text	The complete text of the work. The text may be in display format or be a simple character representation of the complete text of the work.
index only	Limitation of rights of the search service or aggregator to only allow indexing of portions of the content with limited or no rights to display content to users. Most often applied to full text requiring the user to be directed to another source for display capabilities.
snippet	A small portion of the text of the work, often including text immediately around terms matching the query. Used for display of search results to give the end user context of the retrieved result.
basic metadata	Metadata excluding textual descriptions extracted directly from the work including abstracts, tables of contents, and sample pages.
enhanced metadata	Metadata that is augmented by textual descriptions extracted directly from the work including some or all of abstracts, tables of contents, and sample pages

Methods of Data Exchange

harvest	Method of extracting indexing and/or full text from remote web-accessible sites for the purpose of providing search and/or display from a different location.
syndication	Method of pushing content to remote indexing, abstracting, or display services.
RSS (rich site summary)	Standard method for advertising the availability of frequently published content that includes metadata, publication date, and authorship information.
AtomPub (Atom publishing protocol)	Standard method for publishing syndication feeds similar to RSS. AtomPub provides a more robust method than RSS for publishing beyond blog and webpage entries.

Term

Definition

screen scraping

Simplest method of harvesting content that places no technical burden on the publishing site. The human-readable text is extracted and indexed, formatted, and searched by various engines.

linked data

Method for publishing structured data intended for further processing by machines rather than directly by human readers.

API access
(application programming
interface access)

Generic description of a method for exposing services and data through an advertised programming interface. Specification of the programming interface through either a proprietary or published standard is required.

Actors

end user

The final consumer in an information retrieval session.

licensee

The institution or individual who has acquired rights to access others' content or services.

licensor

The rights holder granting search and/or access rights to others. Also referred to as *rights owner*.

content provider

The organization providing dissemination of the content (literature or information). May be a publisher, aggregator, OA or institutional repository, or A&I service provider. The same content may be available from multiple content providers.

publisher

A person or organization whose predominant activity is to commission, create, collect, validate, and edit information in printed and/or in electronic form. A publisher may also act as a content provider for its material.

aggregator

The organization that collects information from varied sources and provides consistent search, presentation, and/or access.

Open Access
Repository/Institutional
Repository

The entity that collects and disseminates content created during the research process at academic institutions.

A&I service provider

The organization that provides abstracting and indexing services, including controlled vocabularies, for published content.

License Terms

authenticated access

Access to search, metadata display, or content display via a known method for verifying end user identity or institution affiliation.

subscription

Provision of access to search, metadata, or content through specified license terms for a period of time, usually annually.

mutual subscriber

A collaboration between two service organizations agreeing to provide equivalent access terms to a single individual or institution.

market product

A specific, defined collection of resources made available for license or use by a content provider to a library. This could be a citation database, a set of e-journals or e-books, or other constellation of titles or objects commonly licensed as a single entity.