Open Discovery Initiative: Recommended Core Metadata Elements for Content Providers

Becky Baltich Nelson, MLS

NLM Associate Fellow, 2015-2016

Date: May 28, 2016

Lou Knecht, MLS

Project Sponsor

National Library of Medicine, Bibliographic Services Division

Table of Contents

Abstract

Objective

The goal of this project was to determine whether the MEDLINE journal citation includes all of the core metadata elements recommended by the Open Discovery Initiative and, if necessary, make suggestions as to how to incorporate any missing elements into the MEDLINE DTD. The Open Discovery Initiative is a technical recommendation developed by the National Information Standards Organization (NISO) regarding data exchange and encourages the use of specific metadata elements by publishers, aggregators, and abstracting and indexing service providers.

Methods

The National Library of Medicine MEDLINE DTD, PubMed Journal Article DTD, and Journal Article Tag Suite (JATS) were reviewed to determine whether they included all of the core metadata elements proposed in the Open Discovery Initiative. Advice was solicited from librarians and other professionals throughout NLM to ensure that the assessment was correct and to discuss the feasibility of adding the missing metadata elements to the MEDLINE DTD. Additionally, websites of two other abstracting and indexing service providers were analyzed to ascertain their level of compliance with the Open Discovery Initiative.

Results

Six of the fifteen recommended core metadata elements and one of the three recommended enriched content elements are not currently a part of the MEDLINE DTD. Information gathered from National Library of Medicine (NLM) professionals indicated that three of the missing metadata elements (Full Text Flag, Content Type, Content Format) could be added to the MEDLINE DTD, but that the data necessary for the others is either not collected or not practical for incorporation into the DTD.

Review of two other abstracting and indexing service providers indicated that the NISO ODI recommendations are not presently being fully addressed in practice by content providers similar to NLM.

Conclusions

The Open Discovery Initiative core set of metadata elements for content providers requires data that is more easily supplied by publishers than abstracting and indexing service providers like NLM. It is recommended that adding new fields for the missing core metadata elements is postponed until the development of a new DTD is under way. In the meantime, it is proposed that NLM post the ODI compliance checklist, with explanations for missing elements and the plans for future compliance. This confirms the NLM’s role as a promoter of national standards and facilitates the transparency urged by the ODI.

Introduction

The ubiquity of Google Search, with its simple and user-friendly interface, has led many people to expect a similar experience from any other type of information provider. Libraries are working to meet this user expectation by shifting to a discovery service model based on a central index, rather than federated searching or other metasearch options. A centrally indexed discovery service employs a single index housing standard and comprehensive metadata for each item from content providers; federated searching on the other hand, searches content providers' individual indexes, which are constructed using customized metadata elements. This new model provides better search outcomes, but it also creates new challenges. To address these challenges and promote best practices, the National Information Standards Organization (NISO) released a technical recommendation in 2014 called the Open Discovery Initiative (ODI). The aim of the ODI was to develop and recommend a set of standard practices that libraries, content providers, and the creators of discovery services could incorporate into their workflows to increase the effectiveness of centrally indexed search tools and technologies.

These recommendations are relevant to the National Library of Medicine (NLM) in its roles as both a supporter of national standards and as a content provider, which the ODI defines as an organization that offers "content products or services, primarily intended for access by library patrons or the general public".[1] NLM offers many such products, like PubMed and MedlinePlus. Centrally indexed discovery services must have cooperation from content providers so that content items can be indexed fully and consistently, making them easily discoverable. And in order for libraries to evaluate and choose the discovery service provider that will best meet their needs, there must be transparency regarding the extent to which the library's free and licensed content is indexed and discoverable within each discovery service. To that end, the ODI puts forth general requirements detailing the metadata elements that content providers should supply to discovery service providers and to libraries.

The goal of this spring project was to review the ODI and its recommendations for content providers; survey the NLM MEDLINE DTD, PubMed Journal Article DTD, and the Journal

[1] "Open Discovery Initiative: Promoting Transparency in Discovery," *National Information Standards Organization,* accessed February 10, 2016, http://www.niso.org/workrooms/odi/publications/rp/rp-19-2014

Article Tag Suite (JATS) to determine current compliance with the ODI recommendations; and to make suggestions for incorporating missing core metadata elements into the MEDLINE journal citation via the MEDLINE DTD.

Methodology

The work for this project was broken up into three distinct phases. During the first phase, the ODI documentation was reviewed in its entirety, with a special focus placed on the set of core metadata elements to be used by content providers. Once the review was complete, the NLM MEDLINE DTD, PubMed Journal Article Publisher DTD, and the Journal Article Tagging Suite (JATS) were analyzed to determine whether these metadata elements were already present in their respective lists of allowable elements and attributes. Additional information was elicited via email and in-person meetings from the NLM professionals that created and currently maintain these tools. Biweekly meetings with project sponsor Lou Knecht and project resource person David Anderson were established to discuss the progress of this research and to provide guidance and feedback.

The second phase of the project was aimed at determining the feasibility of adding each of the missing recommended core metadata elements to the MEDLINE DTD. Through emails and in-person meetings, information was collected from NLM employees in the Technical Services Division, Library Operations and the Information Engineering Branch, NCBI to ascertain whether NLM collected and maintained the data necessary for each element and if they believed a process could be designed to incorporate that data into the MEDLINE DTD.

The project's third phase focused on a survey of similar content providers' websites to assess their compliance with the ODI and to use this information, along with the information gathered throughout the project, to develop ODI compliance recommendations for the NLM.

Results

Review of the Open Discovery Initiative

NISO described the ODI as "a technical recommendation for data exchange including formats, method of delivery, usage reporting, frequency of updates and rights of use; a way for libraries to assess content providers' participation in discovery services; [and] a model by which content providers work with discovery service vendors via fair and unbiased indexing and linking."[2] In addition to these topics, the ODI also covers the history of library discovery services, summarizes other initiatives aimed at providing discovery service recommendations, proposes future work and next steps, and provides conformance checklists for both content providers and discovery service providers.

The focus of this project is on the ODI best practices for content providers. The ODI establishes two sets of best practices, both sets outlining the metadata elements that content providers should provide to the other stakeholder groups. The first set includes those metadata elements that should be provided to discovery services. The general requirements are as follows:

1. Content providers should make available to discovery service providers core metadata, and underlying full-text/original content for complete offerings for the purposes of indexing to meet licensed customers' and authenticated end users' needs.
2. To this aim, all content providers should make available to discovery service providers, at a minimum, the core set of metadata elements for each item they submit for indexing.
3. Content providers should provide the content item (full text, transcript, etc.) and additional descriptive content (abstract/description and controlled and/or uncontrolled keywords) for as much of their content as possible.[3]

The core metadata elements mentioned here were developed using the KBART metadata encoding schema as a framework with additional elements to ensure the core set yields a thorough description.[4] The ODI recommends that, at a minimum, this core set of metadata elements be made available for every individual item sent to a discovery service provider. Where possible, the ODI advocates for also providing enriched content as well. The ODI points out that "inclusion of enriched content in indexes and as used for relevancy ranking greatly improves the discovery service for providers; it brings particular benefit to librarians and advanced researchers who are accustomed to controlled vocabularies."[5] These core metadata elements and enriched content metadata elements are defined in the tables below.

[2] "Open Discovery Initiative," *National Information Standards Organization,* accessed February 10, 2016, http://www.niso.org/workrooms/odi/

[3, 4, 5] "Open Discovery Initiative: Promoting Transparency in Discovery," *National Information Standards Organization,* accessed February 10, 2016, http://www.niso.org/workrooms/odi/publications/rp/rp-19-2014

Table 1. Open Discovery Initiative recommended core metadata elements for content providers.[6]

Field Name	Definitions
Title	The main title of the item.
Authors	The author(s) if the item. Individual authors should be listed in lastname, firstname order.
Publisher Name	The name of the publisher of this item.
Volume	Volume number of the resource, where applicable.
Issue	Issue number of the resource, where applicable.
Page(s)	Page numbers of the resource, where applicable.
Date/Date Range	The date of publication. For a serial run, coverage dates included for the serial.
Item Identifier	One or more standard identifiers for the print or online version of the item (e.g. ISSN, OCLC number, ISBN, DOI, etc.). The identifier should be preceded by a label indicating the type of identifier.
Component of Title	Describes the publication or serial of which the individual item is a part (e.g., for journal articles, the serial title; for track on a CD, the album title; etc.).
Component of Title Identifier	Provides a standard identifier for the component title defined above (e.g., ISSN, OCLC number ISBN, DOI, etc.). The identifier should be preceded by a label indicating the type of identifier.
Item URL	Either an OpenURL or direct link for the specific item's full text.
Open Access Designation	To comply with the NISO Open Access Metadata and Indicators (OAMI) group's recommendations, if an item is open access, this status should be indicated with "free_to_read" and otherwise left blank. See www.niso.org/workrooms/oami/.
Full Text Flag	A yes/no statement describing whether the content provider makes this item available in full text (or for non-print media, a full-length or high-resolution version) o the DSP for the purpose of indexing. It is expected that this will be disclosed by DSPs to libraries in future when describing indexing coverage for a title or collection.
Content Type*	Intended to be used to identify whether the content being described is textual, a visual recording, a s recording, etc. The textual descriptors from the controlled list established in the MARC 21 Type of Re position (06) of the Leader field is recommended to be used for this field's content.
Content Format*	Intended to be used to identify whether the nature of the content being described is monographic, serial, a component part, collection, etc. The textual descriptors from the controlled list established in the MARC 21 Type of Record position (07) of the Leader field is recommended to be used for this field's content.

*It is recognized that many content providers merge Content Type and Content Format in their systems. Providing separate fields for this data is preferred, but the current practice of a single field may continue if separating the data is too burdensome.

Table 2. Open Discovery Initiative recommended enriched content metadata elements for content providers.[7]

Field Name	Definitions
Indexing Data	One or more keywords (from controlled or uncontrolled vocabularies) to describe the content of the item.
Full Text/ Transcript	For text items, the entirety of the document. For audio or video content, a full transcript of the spoken content of the material. May not be relevant for all indexed content.
Abstract/ Description	Either a text summary on the content or (for non-text materials) a description of the item.

Review of the MEDLINE DTD, PubMed Journal Article DTD, and Journal Article Tag Suite

The MEDLINE DTD, PubMed Journal Article DTD, and Journal Article Tag Suite (JATS) were analyzed to determine if the ODI core metadata elements or enriched content metadata elements were already present. These three DTDs serve different purposes: the MEDLINE DTD is used to export citation data to NLM licensees; the PubMed Journal Article DTD is used to accept citation data from publishers or their aggregators; and the Journal Article Tag Suite (JATS) is used to accept full text articles for deposit into PubMed Central (PMC). During the analysis, Jeff Beck, a JATS expert from the Information Engineering Branch, attended one of our meetings to provide direction and ensure accuracy. The research showed that many elements were indeed already included, albeit with different naming conventions. The chart below displays the results of the analysis, including the name for the metadata element in the markup if it is present and an X to signal a missing metadata element.

6, 7 "Open Discovery Initiative: Promoting Transparency in Discovery," *National Information Standards Organization,* accessed February 10, 2016, http://www.niso.org/workrooms/odi/

Table 3. Results of the analysis of NLM DTDs regarding the presence of ODI recommended core metadata elements.

ODI Core Metadata Elements	**MEDLINE DTD**	**PubMed Journal Article DTD**	**Journal Article Tag Suite (JATS)**
Title	<ArticleTitle> <VernacularTitle>	<ArticleTitle> <VernacularTitle>	<article-title> <trans-title> <subtitle>
Authors	<AuthorList> (and following tags)	<AuthorList> (and following tags)	<contrib>
Publisher Name	X	<PublisherName>	<publisher-name>
Volume	<Volume>	<Volume>	<volume>
Issue	<Issue>	<Issue>	<issue>
Page(s)	<Pagination>	<FirstPage> <LastPage>	<fpage> <lpage> <elocation-id>
Date/Date Range	<PubDate> <ArticleDate>	<PubDate> (and following tags)	<pub-date>
Item Identifier	<ELocationID>	<ELocationID> <ArticleId>	<article-id>
Component of Title	<Title> <ISOAbbreviation> <MedlineTA>	<JournalTitle>	<journal-title> <abbrev-journal-title>
Component of Title Identifier	<ISSN> <NlmUniqueID>	<ISSN>	<issn> <isbn>
Item URL	X	X	<self-uri> <uri> <ext-link>
Open Access Designation	X	X	<ali: free_to_read>
Full Text Flag	X	X	X
Content Type	X	X	X
Content Format	X	X	X

Table 4. Results of the analysis of NLM DTDs regarding the presence of ODI recommended enriched content metadata elements.

Enriched Content	MEDLINE DTD	PubMed Journal Article DTD	Journal Article Tag Suite (JATS)
Indexing Data	<SupplMeshList> <MeshHeadingList> <KeywordList> <ChemList> <PubTypeList>	X	<kwd> <kwd-group>
Full Text/Transcript	X	X	<article>
Abstract/Description	<Abstract> <AbstractText> <OtherAbstract>	<Abstract> <AbstractText> <OtherAbstract>	<abstract> <trans-abstract>

Implementation Feasibility

The MEDLINE DTD is missing six of the fifteen core metadata elements and one of the three enriched content metadata elements recommended by the ODI. The six missing elements are: Publisher Name, Item URL, Open Access Designation, Full Text Flag, Content Type, and Content Format; the one enriched content metadata element is Full Text/Transcript. To find out if it would be feasible to incorporate these elements into the MEDLINE DTD, it had to be determined whether the data necessary for each element was being collected somewhere at NLM and if it is possible to develop processes to transfer that data to the MEDLINE DTD, and ultimately, the MEDLINE journal citation.

Publisher Name

Publisher name is missing from the MEDLINE DTD, but is present in the PubMed Journal Article DTD and in the NLM Catalog. In an email exchange with Sarah Weis from NCBI, it was learned that while this metadata element is required in the PubMed Journal Article DTD, the information entered by the publisher is not validated and never updated. Similarly, Diane Boehr from Technical Services explained that the catalogers look at the item and manually enter the publisher information for the NLM Catalog, but it is only ever updated if it comes to their attention that a publisher change has occurred. However, it would technically be feasible to design a process in which Publisher Name data from either the PubMed Journal Article DTD or the NLM Catalog could be accessed and incorporated into the MEDLINE DTD.

Item URL

NLM indexes and houses abstracts of journal articles but does not itself own the full text of the items. To incorporate an Item URL into the MEDLINE DTD, the URL would have to be provided by the publisher. This presents two challenges: obtaining the URL and maintaining good links. Publishers change links to articles frequently, so this would require not only submission of the URL to NLM, but also updating NLM as changes are made. Publishers are resistant to adding more work to their MEDLINE submission processes and NLM does not currently have the manpower or mechanisms necessary to keep this information current.

There is a function within PubMed called LinkOut, which does indeed give the user an option to click and be redirected to the article at the publisher's website. But in most cases, payment is mandatory in order to access the article. It is not a link to a freely available full-text version. The database that houses LinkOut data is separate from the MEDLINE database, and daily updating is required to fix broken links and add new data. This means that if the Item URL metadata element was added to the MEDLINE DTD, the journal citations would need to be updated with the same regularity as the LinkOut database in order to avoid obsolete data.

Open Access Designation

The ODI requests that this metadata element be left blank if the item is not available via open access, or to add the phrase "free_to_read" if it is. While LinkOut houses the links to each article, it does not flag whether or not the item is free. As such, LinkOut would not be the correct avenue for obtaining this information. This data is not gathered elsewhere at NLM for all MEDLINE journal articles. Another consideration is the fact that the open access designation can change for new articles during the embargo period. If there were a way to incorporate this data into the MEDLINE DTD, it would have to be updated as the embargo period for each article came to an end.

Full Text Flag

The requirement for this data element is a yes/no statement simply indicating whether the content provider provides the full text of the item to discovery service providers. MEDLINE houses citations, not full text items; therefore, a no statement would be appropriate for each item provided to discovery service providers. As this data would not become outdated and is identical for each element, it would be possible to create a new field in the MEDLINE DTD or to add the information within the DTD documentation.

Content Type and Content Format

The ODI Content Type and Content Format metadata elements are descriptors taken directly from the MARC 21 Format for Bibliographic Data, the former being the Type of record (06) and the latter the Bibliographic level (07) of the Leader. MEDLINE only houses article citations, so the Content Type and Content Format would remain consistent among each item in the collection. Content Type will always be *a - Language Material* and Content Format will always be *b – Serial component part* (there is one outlying exception, and that is a video journal to which NLM subscribes). Because the data for these elements is consistent across the collection and will not require updating, it would be feasible to add new fields to accommodate them into the DTD or add the information within the documentation.

Full Text/Transcript

NLM is providing an abstracting and indexing service, and does not have the full text to submit to discovery service providers. It will be impossible to enter this enriched content metadata element to the MEDLINE DTD.

ODI Compliance among Similar Abstracting & Indexing Services

The American Psychological Association (APA) and the National Agricultural Library (NAL) provide abstracting and indexing services for literature pertaining to their areas of specialization. APA has developed PsychInfo, a comprehensive bibliographic database that houses abstracts and index records of materials relevant to psychology and related fields dating back to the 1880s.[8] Similarly, NAL's Agricola is a database housing bibliographic records and abstracts of materials from the wide array of fields that pertain to agriculture.[9] As both provide content in a format comparable to MEDLINE, their websites were examined to determine if the recommended core metadata elements are present in their citations and to find out if they have stated a commitment to ODI compliance or have posted the ODI compliance checklist. Neither of them made mention of ODI or had posted the compliance checklist. The results of this research are outlined in Table 5.

Table 5. Missing ODI core and enriched content metadata elements from the abstracting and indexing services PsychInfo and AGRICOLA.

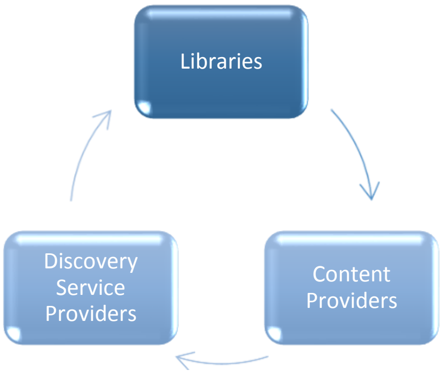
APA PsychInfo	NAL AGRICOLA
Open Access Designation Full Text Flag Content Type Content Format Full Text/Transcript	Open Access Designation Full Text Flag Full Text/Transcript

Discussion

All parties would benefit from adherence to the best practices set forth by the Open Discovery Initiative. If content providers can submit sufficient information in a consistent format for indexing, discovery service providers can be transparent with libraries regarding the depth of their coverage. This will allow libraries to choose the discovery service that will best represent their collection, making their items visible and directing users back to the content providers.

[8] "PsychInfo Highlights," *American Psychological Association,* last modified May 2016, http://www.apa.org/pubs/databases/psycinfo/index.aspx?tab=3

[9] "About the NAL Catalog (AGRICOLA)," *United States Department of Agriculture*, last modified November 3, 2006, http://agricola.nal.usda.gov/help/aboutagricola.html



Traditional content providers such as Gale, IEEE, and Sage have posted the ODI compliance checklist on their website, demonstrating their commitment to discoverability. Abstracting & indexing (A&I) content providers offer a different type of content, a type that does not align as well with the core set of metadata elements. A&I content providers similar to NLM, the American Psychological Association and the National Agricultural Library, have not posted the ODI compliance checklist, nor is ODI mentioned on their websites. It is possible that these types of providers, who cannot be fully compliant with the recommendation, choose not to advertise the standard on their sites as the traditional content providers do.

While the National Library of Medicine does promote the use of standards, traditional content providers are in a much better position to offer the full set of core and enriched content metadata elements recommended by the ODI than are abstracting and indexing service content providers. Upon review of the ODI documentation – the MEDLINE DTD, PubMed Journal Article DTD, and JATS – and talking with NLM experts, it does not seem practical and/or possible to incorporate Publisher Name, Item URL, Open Access Designation, or the Full Text/Transcript elements to the MEDLINE DTD. However, it would be possible to add the Full Text Flag, Content Type, and Content Provider elements to the MEDLINE DTD, as the data (no, a - Language Material, b – Serial component part, respectively) is the same for each item in the collection and would remain static over time. This would require an implementation of new fields within the DTD itself or notes within the DTD documentation.

While the addition of these three metadata elements is achievable, NLM should consider whether moving forward with the addition is worthwhile. NLM is already in partial compliance with the ODI recommendations, but it will not be able to comply with them fully. Is it beneficial to be as compliant as possible? There are plans to migrate to a single DTD that would replace the MEDLINE DTD and the PubMed Journal DTD, but the implementation of the new DTD will not take place for a couple of years. With that in mind, a decision will have to be made whether it would it be productive to add the three metadata elements now, or if it would make more sense to wait until the new DTD is in development.

Given that MEDLINE is a database of journal article citations, the data that would be provided within these three metadata elements should be implicit. Adding fields to indicate that its content items are language materials and components of a serial, as well as a note that full-text is not available, may not provide enough additional value to the citation to warrant the work necessary to add it to the DTD at this point in time. Instead, NLM may want to consider posting the ODI compliance checklist documenting its current capacity for making available these metadata elements and plans for the future. NLM can wait to incorporate Full Text Flag, Content Type, and Content Format into the new DTD once it is under way (and any of the other missing metadata elements, should it have become feasible by then). This confirms the NLM’s role as a promoter of national standards and facilitates the transparency required by the Open Discovery Initiative.

Acknowledgements

Thank you to Lou Knecht for proposing this project and for helping me to learn a lot about both NLM and national standards (and for taking me on a drive through a cherry blossom wonderland). Thank you as well to David Anderson for sharing his knowledge and making time for our meetings. A special thank you to Jeff Beck, Diane Boehr, Kathy Kwan, Iris Lee, and Sarah Weis for providing so much helpful information. And lastly, thank you to Kathel Dunn, Loan Nguyen, and Tyler Nix.

Appendix I

Open Discovery Initiative Conformance Checklist for Content Providers

The NISO encourages Content Providers to comply with the recommendations set forth in the Open Discovery Initiative (NISO RP-19-2014) and to report on their level of compliance to increase transparency between Content Providers, Discovery Service Providers, and libraries.

The checklist below pertains to the National Library of Medicine MEDLINE database of journal citations.

Y/P/N	Recommendation	Reference	Comment
P	Content Provider makes available to Discovery Service Providers core metadata and underlying full-text/original content for complete offerings.	3.2.1.1 (1) (p. 15)	Under review.
P	Content Provider makes available to Discovery Service Providers the core set of metadata elements (see 3.2.1.2) for each item submitted for indexing.	3.2.1.1 (2) (p. 15)	Under review.
Y	Content Provider provides the content item and additional descriptive content for as much of their content as possible.	3.2.1.1 (3) (p. 15)	In this instance, “content item” refers to the citation itself.
Y	Content Provider provides libraries, on request, with a statement of participation in the discovery services, including disclosure of coverage depth and content depth.	3.2.2 (p. 22)	
Y	Content Provider agreement with Discovery Services Providers do not include any non-disclosure agreements.	3.2.3 (p. 22)	
N	The transfer of Content Provider’s data to Discovery Service Providers makes use of existing standards where applicable and uses on of the metadata encoding schemes listed in 3.3.3.	3.2.4 (p. 22)	NLM uses its own XML DTD for MEDLINE journal citations.

Yes indicates compliance with the indicated paragraph of this Recommended Practice.
Partial indicates partial compliance with the indicated paragraph of this Recommended Practice.
No indicates non-compliance with the indicated paragraph of this Recommended Practice.