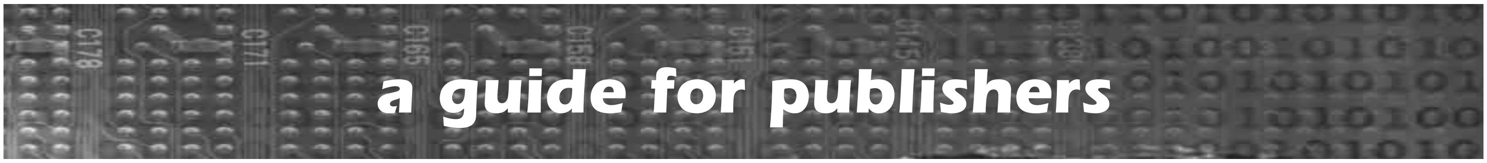

a guide for publishers

Metadata DEMYSTIFIED

Amy Brand, Frank Daly
and Barbara Meyers

NISO Press THE SHERIDAN PRESS

ISBN 1-880124-59-9

Metadata Demystified:

A Guide for Publishers

Table of Contents

Published by: The Sheridan Press & NISO Press

Contributing Editors: Pat Harris, Susan Parente, Kevin Pirkey, Greg Suprock, Mark Witkowski

Authors: Amy Brand, Frank Daly, Barbara Meyers

Printed July 2003

Metadata Demystified: A Guide for Publishers

This guide presents an overview of evolving metadata conventions in publishing, as well as related initiatives designed to standardize how metadata is structured and disseminated online. Focusing on strategic rather than technical considerations in the business of publishing, this guide offers insight into how book and journal publishers can streamline the various metadata-based operations at work in their companies and leverage that metadata for added exposure through digital media such as the Web. This exposure is an additional way of sharing information about content. It benefits not only publishers, but also potential readers who seek access to published products and the resource discovery environment more generally.

Publishers work with metadata on a daily basis. It is in the manuscript tracking process, in internal reports and content management systems, in marketing copy, and in the information transmitted to the supply chain. Whenever publishers complete copyright registration forms or supply promotional and library cataloging information during the editorial/production process, they create metadata. Similarly, whenever authors cite other publications, or libraries record their holdings, they create metadata.

What Metadata Is

The term *metadata* refers to information about information or, equivalently, data about data. In current practice, the term has come to mean structured information that feeds into automated processes, and this is currently the most useful way to think about metadata. This definition holds whether the publication that the metadata describes is in print or electronic form. While metadata in publishing can be classified according to a variety of specific functions, such as *technical metadata* for technical processes, *rights metadata* for rights resolution, and *preservation metadata* for digital archiving, this guide focuses on *descriptive metadata*, or metadata that characterizes the content itself.

Occurrences of metadata vary tremendously in richness; that is, how much or how little of the entity being described is actually captured in the metadata record. The strategic decisions publishers make about metadata often concern how much to expose. The answer to this question depends on the application at hand. In order to enable reference linking across publisher platforms, for instance, the number of metadata elements required is minimal, often less than what occurs in a typical citation. The CrossRef metadata set, which we will look at in section 5, contains only a handful of required elements. For electronic bookselling, where one role of metadata is to approximate the experience of perusing a physical book in a bookstore, the richer the metadata record, the better. Hence, the Online Information Exchange (ONIX) standard for books specifies over 200 elements.

To illustrate what metadata is, let’s look at a simple metadata standard called Dublin Core. The Dublin Core Metadata Initiative (DCMI) got underway in 1995 as a joint effort among professionals from the publishing, library, and academic communities. One outcome of this effort was the Dublin Core Metadata Element Set, which became a NISO standard in 2001 (ANSI/NISO Z39.85-2001) and an international standard (ISO 15836) in 2003.

The DCMI standard includes fifteen optional metadata elements for describing cross-genre, cross-disciplinary information resources. These elements are: *title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage,* and *rights*. Some of these elements relate to the content of the item, some to the item as intellectual property, and others to the particular instantiation, or version of the item.

The Dublin Core website (http://dublincore.org) uses its own metadata scheme to display document information. Table 1 shows a three-element Dublin Core record.

The left-hand column lists element types, and the right-hand column assigns element values for this particular document. Dublin Core has been mapped to several other metadata formats, including the Machine Readable Cataloging (MARC) 21 bibliographic format for representation and exchange of bibliographic information that most library catalogs use today. See http://www.loc.gov/marc for more information.

Metadata in the publishing and communication cycle is not new. What *is* relatively new to the broader publishing community, and crucial for interoperability in the digital age, is standardization. This is the process of building consensus around best practices in the formatting and use of metadata for specific applications, so that machines can interpret and exchange this information efficiently. In recent years, clear standards have emerged to define metadata elements and the record layout for transmitting those elements.

Standards-building is an ongoing, collaborative process in which book and journal publishers should participate. Despite the fact that a much greater proportion of journal content than book content is digitized, publisher-driven standardization initiatives in book publishing are more advanced than in journal publishing. Book publishers have been driven toward standardization in order to capitalize on aggregated bookselling—traditionally via wholesalers and now through the Internet—which has required them to conform to standards for supplying promotional metadata. Even existing standards have a routine review process to incorporate new features, and publishers can take part via organizations such as the National Information Standards Organization (NISO, http://www.niso.org), in order to have input on how both current and new standards take shape.

The remainder of this document is structured as follows: In the next section, we will refine our operational definition of metadata by explaining its relationship to Extensible Markup Language (XML) and to identifiers. Then we will look at the internal and external roles of metadata in today's publishing companies, and why metadata has become a strategic issue. Next, we will turn to metadata practices and trends in book publishing. In the final section, we will discuss evolving standards in journal publishing.

Along the way, we will provide pointers to tools and resources that publishers should be

Table 1. Dublin Core Record

Title	Overview of Documentation for DCMI Metadata Terms
Identifier	http://dublincore.org/usage/documents/overview
Description of Document	This page provides an overview of official documentation of all DCMI metadata terms._

familiar with as they embark on integrating automated metadata processes into their content management, production, and marketing/supply systems. A handful of sample metadata records will be displayed, but these are not intended to replace implementation guidelines for the various standards they illustrate, nor do they reflect the full range of metadata schemes, standards, and initiatives presently in use across the information industry.

What Metadata Isn’t

The term *metadata* has come to refer to standardized, structured information that machines can interpret and use. The boundaries of this definition often overlap, yet are not to be confused with, two related sets of conventions: *XML*, a widely adopted standard for structuring and exchanging data, and *identifiers*, which are standards for uniquely naming a piece of content or intellectual property. In this section we take a brief look at XML and identifiers to explain their relation to metadata.

XML

Although not a programming language per se, XML is a language for expressing rules that give structure to any kind of textual data, including but not limited to metadata. One way to think about XML in this context is as the information “wrapper” or container of choice for your metadata. XML has been widely adopted because it was designed for precisely the kind of data transfer that comprehensive electronic publishing requires. It also provides an application-independent method for sharing data, and because it is free to license, XML can save publishers money through the use of inexpensive, off-the-shelf tools. A large part of its power comes from the nearly universal support it receives from product vendors, standards bodies, academia, and the open source community.

XML syntax. XML uses a simple syntax that both people and machines can easily process. The syntax consists of matching start and end tags, such as <journal> and </journal>, to mark up information elements. These tags can also be associated with attributes, also known as name-value pairs (e.g., type = “print”).

Document Type Definition (DTD). An XML DTD provides a description (actually expressed in Standard Generalized Markup Language, or SGML) of the building blocks of any type of XML document, whether that document is a list, a metadata record, a journal article, or a whole book. It includes what to call different types of elements, how they should be ordered, and how they interrelate. Some DTDs are proprietary—created by a company for their internal use—while others are standardized and freely available. The latter include the metadata formats we will discuss in sections 4 and 5.

XML schema. An XML schema (also called an XSD file) is itself an XML document and is an alternative to the DTD that provides developers with enhanced validation capabilities and more refined tools for structuring their own XML-based formats. Whereas DTDs only allow for relatively simple data types, a schema has a set of powerful, flexible semantics for defining what an XML file can contain.

XML workflow. This is not a technical term, but a way of describing the infrastructure that publishers put in place in order to capture data in XML format as needed and streamline processes for creating, re-purposing, and disseminating that data.

For additional information on XML,
go to http://www.w3.org/XML

Identifiers

Identifiers are names or strings adhering to certain conventions that, if properly employed, ensure uniqueness. While standard identifiers for publications have been in use for decades, unambiguous identification of content entities has become especially important for electronic publishing and e-commerce platforms. Identifiers and metadata are not one and the same, yet identifiers are most useful in association with metadata.

Identifiers for book and journal publishing can be characterized according to six parameters or features:

- Whether the identifier itself is *transparent* (derivable) or *opaque* (not inherently meaningful or interpretable).
- Whether the entity that the identifier points to is a *work* (an abstraction, not tied to any particular physical medium) or a *manifestation* (an exemplar in a particular physical medium of a work, such as the online version of a journal that exists in multiple media).
- Whether or not the identifier is *actionable* in the electronic environment, so that clicking on it takes you directly to the thing being named; for example, the URL identifier.
- If actionable, whether or not the identifier is truly *persistent*: that is, designed to withstand changes in the online location of content identified. URLs are actionable but not persistent.
- Who drives or regulates the identifier registration process (e.g., the author, publisher, or library community).
- Whether descriptive metadata is registered in association with the identifier.

The following identifiers are the most familiar ones for books and journals in terms of these properties.

ISBN. The International Standard Book Number (ISBN/ISO 2108) is a ten-digit numeric string (e.g., ISBN 0-500-27664-1) that uniquely identifies each manifestation of a book or non-serial publication. Although sub-parts of the ISBN identify country (or language area) and publisher, ISBNs strings as a whole are opaque ("dumb numbers"), non-actionable at present, publisher-driven, and not currently associated with a metadata registry. The ISBN standard is now undergoing a revision process that will increase the ISBN string to thirteen digits, in addition, publishers will be encouraged to deposit a core set of metadata as part of the registration process.

For more information on the ISBN, go to http://www.isbn.org/standards/home/index.asp

ISTC. The International Standard Text Code (ISTC/ISO 21047) is a new numbering scheme, mainly but not exclusively for books, under development for unique identification of textual works, as opposed to manifestations. It is intended to be opaque, actionable, and persistent, and potentially to be assigned as soon as a work is conceived by a creator or author. It may potentially be used as an overarching identifier to tie together the various related identifiers registered at the manifestation level. As part of the ISTC registration process, descriptive metadata will be captured by the ISTC registration agency and will include, at minimum, a title, the name of the author or contributor, a unique identifier for the ISTC registrant, registration date, and whether the work identified is a derived or original.

For more information on the ISTC, go to http://www.nlc-bnc.ca/iso/tc46sc9/istc.htm

ISSN. The International Standard Serial Number (ISO 3297) is an opaque eight-digit number (e.g., 1234-1231) for unique identification of journals and other serial resources; the same serial in a different physical medium is assigned a different ISSN, and title changes to serials frequently call for new ISSNs as well. ISSN assignment is a regulated process. ISSNs are assigned by ISSN national centers; publishers should contact their national ISSN center to request an ISSN assignment. The National Serials Data Program (NSDP) at the Library of Congress coordinates the U.S. ISSN program.

Each ISSN assigned to a serial publication is registered in an international database (the ISSN Register), along with a relatively rich metadata record. Among the bibliographic elements in these records are ISSN, key title, abbreviated key title, frequency of publication, language, other forms of the title, place of publication, publisher, former title(s), pointers to other language editions and other media editions, and URLs. ISSN records are available in MARC-compatible format.

For more information on the ISSN,
go to http://www.issn.org

SICI. The Serial Item and Contribution Identifier standard (SICI) (e.g. SICI: 0002-8231(199412)45:10<>1.**1**.TX;2-M) is a NISO standard (ANSI/NISO Z39.56-1996) for the unique identification of a serials issue or article, regardless of the distribution medium. SICI is designed to be dynamically constructed and in that sense is a transparent identifier. It is neither actionable nor associated with a metadata record in its current implementation. Due to their strict, derivable format, SICIs can be created and used by anyone involved in serials management, and automated SICI generators have been created for this purpose.

For more information on the SICI,
go to http://sunsite.berkeley.edu/SICI

DOI. The Digital Object Identifier (DOI) syntax is a more recent open standard (ANSI/NISO Z39.84-2000). The DOI system is a complete system for implementing persistent identifiers, and DOIs themselves are variable-length alphanumeric strings (e.g., doi:10.1101/gr.10.12.1841) assigned by publishers at any level of granularity. Although the DOI is designed to be opaque, the DOI suffix can incorporate other existing identifiers as an option. The DOI is also actionable; one click on a properly implemented DOI gets the user to the location of the content being identified. The DOI is persistent because it is paired with the content object's electronic address, or URL, in an updateable central directory and published in place of the URL; this avoids broken links while allowing the content to move as needed. Although the content that it is linked to may take the form of a manifestation (the electronic version of an article, for instance), the DOI can function as a work-level identifier when it is associated with a rich set of metadata elements describing a work.

Declaration of kernel metadata is in the process of becoming mandatory for all DOIs in the global DOI directory; the requisite metadata follows a carefully designed scheme based on indecs (http://www.indecs.org) to maximize interoperability. This way, the DOI can support a range of applications for electronic content, such as e-commerce, management of rights and permissions, and the creation of learning objects. For example, among the official DOI registration agencies of the International DOI Foundation, Learning Objects Network (http://www.learningobjectsnetwork.com) is applying DOI functionality to SCORM-compliant learning objects. Sharable Content Object Reference Model (SCORM) consists of metadata specifications for a range of e-learning content applications. See http://www.adlnet.org for more information.

For more information on the DOI,
go to http://www.doi.org

This selection of identifier standards currently in use in book and journal publishing indicates a clear trend toward identifiers with the following properties: actionability, persistence, opacity, and association with metadata. Identifiers with these characteristics best meet the demands of the digital medium. While the ISBN, ISSN, and SICI are not currently actionable, they could well be in the future. Actionable, persistent identifiers add value to publications because they enable new functionality and work reliably in the Web environment. Identifiers do not need to be transparent or inherently meaningful if they are associated with descriptive metadata and primarily interpreted by machines. Finally, registration of an identifier along with metadata lays the groundwork for constructing other automated services around the content being identified.

Why Metadata Is Important

Metadata can take many forms, and metadata records can vary tremendously in richness, creating an array of content management and economic models.

What metadata means to the publisher

Publishers benefit in many ways from automating and streamlining their internal metadata practices. In book publishing, it is still common to see employees in different departments re-keying the same descriptive information for different purposes; for instance, when a new contract is logged, when that same manuscript is launched for editing, when its marketing and catalog copy is created, etc. With appropriate back-office tools and procedures in place, a publisher can set up a database of metadata elements compiled from the various departments. This database can feed multiple metadata templates corresponding to the formats required for different purposes, both internal and external. Given such a system, responsibility for validating the data can be easily shared across departments. At the same time, any update to an information element, such as the title or the price, is automatically propagated to all outputs.

While supplying structured metadata according to several formats may seem like a huge task, the web of mappings among common metadata standards continues to grow, and there are many shared elements across the different standards. For example, the data elements currently proposed for the new ISBN kernel were developed as a subset of ONIX and are a subset of Dublin Core. All of the standards a publisher now encounters are likely to be tagged in XML and to function across several formats.

The benefits of structuring and tagging text apply not only to metadata but to full-text content. Full-text mark-up of books and journals allows them to be readily re-purposed for course-packs, in derivative works requiring a subset or re-ordering of the original content, and as input to emerging archival standards. The key to successful metadata usage is to develop the systems and procedures necessary to maintain and disseminate metadata as an integral part of the publication process. Creating structured metadata as a normal part of the production workflow allows a publisher to provide consistent information about products to all the communities using that information.

What metadata means to the reader

Many of the advantages that publishers reap from effective use of metadata turn out to benefit the reader and research communities as well. For example, the online aggregation of book metadata brought about by centralized Internet bookselling was a boon for publishers, who saw an unprecedented surge in sales of the backlist titles that they no longer promoted through established channels; it was equally an advantage for scholars seeking out those obscure backlist titles. Readers, for the first time, had at their disposal an easy way to search across a comprehensive, cross-publisher database of available books *and* complete a purchase. The digital medium has made published information easier to disseminate, search, and sell, and metadata plays a critical role in these advantages.

In the publication of journals, cross-publisher metadata has traditionally been aggregated by intermediaries, or secondary publishers, who create sophisticated tools and services (e.g., citation indexing and resource discovery) around subject-based databases of bibliographic information and journal article abstracts. The process of compiling this metadata has been substantially automated, although there are still some manual components, such as selecting content for inclusion, classification of content, and writing abstracts where they are not already available.

Abstracting and indexing (A&I) services have been a source of income for publishers who sell metadata or have their own secondary divisions. Publishers also earn income from aggregators that license full-text content or link back to publishers and thereby drive article sales and journal subscriptions. These business models are currently in flux. Many publishers now have their own journal websites where they freely provide bibliographic information, abstracts, tables of contents, and other resources that they may previously have considered proprietary.

From the end-user perspective, metadata and its innovative use by publishers have already transformed the research process. One important metadata-driven trend is toward virtual, or distributed, aggregation of information resources. Researchers who have long relied on specialist databases to mine authoritative information resources in their fields now turn to powerful search engines that index, but do not aggregate, those resources. The more robust the metadata that publishers expose for this purpose, the more they will benefit from this trend. Interlinking of resources is another example of distributed integration in e-journal publishing. Both publisher and researcher benefit from initiatives that use metadata and identifier registration to enable cross-publisher linking without aggregation of any proprietary content. (The term *distributed integration* is attributed to Brian Schottlaender; see Schottlaender, B., "Portals for Integration and Collaboration" presented at the AAP/PSP Annual Conference, Washington, DC, February 2003.)

Publishers are now cooperating directly with one another, with some exposing not only their metadata but also their full text for search and navigation purposes. In addition, automated tools for the intelligent classification of content have become more available. As a result of these trends, there will be less of a need for manual aggregation of subject-based resources in the future. As publisher-supplied metadata grows to include more semantic information about a publication, conceptually based research tools will also evolve. As standards emerge for capturing metadata associated with the individual user (e.g., access rights profiles and personal preferences), frameworks will be required for structuring how that kind of metadata interacts with the metadata for information resources. A number of initiatives are currently underway to specify, at a high level, how metadata standards for different domains (publications, individuals, e-commerce applications) should interoperate.

(See, for example, http://www.indecs.org, http://www.cores-eu.net, and http://www.w3.org/RDF).

Metadata is thus both a marketing tool and a way to add functionality to electronic publications. It allows publishers to "open up" their proprietary content for e-commerce and resource discovery applications such as indexing, search, and linking, while maintaining control over their own trading practices.

Book-Oriented Metadata Practices

An explanation of how book wholesalers, retailers, and libraries use metadata will clarify why metadata is becoming critical to the overall success of every publisher. Historically, wholesalers obtained their information about forthcoming titles from visits by publisher sales representatives and/or catalogs. The wholesaler used the information in a publisher's catalog to update their in-house inventory database manually, re-keying the data elements their system needed to track customer demand and order a title.

This inventory database was often the source of the catalogs and selection lists the wholesaler created and mailed to their customers (mainly libraries for research and scholarly titles). As wholesalers expanded the number of book titles they stocked or were willing to obtain for their customers, the cost of this re-keying of data increased. At the same time, the shift in technology from microfiche to CD-ROM and then to the Web increased the amount of information that publishers provided on each title.

These factors led wholesalers to seek an electronic means of obtaining title information from publishers. The earliest standard was the Book Industry Systems Advisory Committee (BISAC) Title Status format. This format was eventually superseded by the BISAC X12 832 transaction. Both of these formats are now obsolete, although the push to adopt ONIX as a standard method of communicating metadata has not entirely replaced them.

Ultimately, wholesalers created web-based applications for their customers that require the detailed range of data included in an ONIX record. At present, wholesalers are using a combination of publisher-provided electronic files and manual keying of data to maintain these applications. ONIX is fast becoming the method wholesalers use to update their web products, and the same wholesalers have been licensing their databases to Internet booksellers for use on bookseller sites. The data that publishers provide to wholesalers, therefore, not only updates their internal inventory file but also feeds the wholesalers' websites and often those of several Internet booksellers. Like wholesalers, booksellers require detailed information about titles to decide on the initial buy. They also require an easy method of placing basic information in their inventory management system, and those with websites require rich metadata for their promotional web pages.

Many library suppliers have developed web-based search and order applications that resemble an Internet retailer's site, with jacket image, table of contents, first chapter, and so on. These sites allow librarians to access considerably more information about a title than could be provided in a catalog. Librarians are also licensing wholesaler and bibliographic databases such as Bowker's *Books in Print* for use on their internal acquisitions systems and Online Public Access Catalogs (OPACs), and beginning to use portions of publishers' ONIX records to enhance their MARC record data.

ONIX

The ONIX initiative got underway in 1999, with the American Association of Publishers (AAP) bringing together the major publishers, wholesalers, online retailers, and book information services personnel to create a universal, international format in which all trading partners, regardless of their size, could exchange information about books. The working group released ONIX 1.0 in January 2000. Release 2.1 of ONIX is currently in development.

ONIX is now published and maintained by EDItEUR in association with the Book Industry Study Group (BISG, http://www.bisg.org) in the U.S. and the Book Industry Communication (BIC) in the U.K, and has become the international standard for book-trade metadata. In addition to the United States and United Kingdom, France, Germany, and Korea have set up national implementation groups; the ONIX DTD has been extended to accommodate the trading practices in these countries.

ONIX comprises both a content specification and an XML DTD. The content specification includes a comprehensive set of carefully defined data elements, code lists and XML tags, that can be either short codes (e.g. <b012>) or text labels (e.g. <ProductForm>). XML schemas have also been defined for trial purposes.

Originally designed for books and other non-serial materials such as audio and point of sale materials produced by book publishers, the scope of ONIX has now grown to cover serials (see below) and a version of ONIX has been developed for the video/DVD sector.

ONIX data elements include structured tables of contents, text items (e.g. descriptions, reviews, extracts, author biographies), images (e.g. jackets, author pictures, double page spreads), links to video, audio or websites, territorial rights information, price and availability in different markets, and promotional information, as well as comprehensive bibliographic information.

The following examples show part of the same ONIX sample record, in the first box using plain text “reference names” in XML, and in the second using short tags:

```
<ProductIdentifier>
   <ProductIDType>02</ProductIDType>
   <IDValue>0816016356</IDValue>
</ProductIdentifier>
<ProductForm>BB</ProductForm>
<Title>
   <TitleType>01</TitleType>
   <TitleText textcase = “02”>British English, A to
   Zed</TitleText>
</Title>
<Contributor>
   <SequenceNumber>1</SequenceNumber>
   <ContributorRole>A01</ContributorRole>
   <PersonNameInverted>Schur, Norman
   W</PersonNameInverted>
   <BiographicalNote>A Harvard graduate in Latin and
   Italian literature, Norman Schur attended the
   University of Rome and the Sorbonne before returning
   to the United States to study law at Harvard and
   Columbia Law Schools.  Now retired from legal
   practise, Mr. Schur is a fluent speaker and writer of
   both British and American
   English</BiographicalNote>
</Contributor>
```

```
<productidentifier>
   <b221>02</b221>
   <b244>0816016356</b244>
</productidentifier>
<b012>BB</b012>
<title>
   <b202>01</b202>
   <b203 textcase = “02”>British English, A to
   Zed</b203>
</title>
<contributor>
   <b035>A01</b035>
   <b037>Schur, Norman W</b037>
   <b044>A Harvard graduate in Latin and Italian
   literature, Norman Schur attended the University of
   Rome and the Sorbonne before returning to the United
   States to study law at Harvard and Columbia Law
   Schools.  Now retired from legal practise, Mr. Schur is
   a fluent speaker and writer of both British and
   American English </b044>
</contributor>
```

Creating an ONIX message involves two basic steps: organizing the data into ONIX-specified fields and storing it in a database; and using an XML software application and the ONIX DTD to organize and tag that data. A single ONIX message may contain data about multiple titles. An ONIX message is transmitted across networks and the Internet the same way that other data is transferred; for instance, as an email attachment or via FTP. Once an online retailer receives an ONIX message, the same tools (an XML software application and the ONIX DTD) are used to validate the data. From that point, the retailer translates the delivered data into what is seen on a web page.

ONIX differs from other metadata standards in that it is a very rich record with over 200 data elements, some optional and some required. For example, ISBN, author name, and title are required elements; book reviews and cover image remain optional. In contrast, DCMI uses only fifteen repeatable, optional elements. A full ONIX record loaded onto a website provides a searching experience similar to that of browsing the physical book. Just as book retailers and wholesalers came to require an ISBN and a bar code, they will soon require an ONIX record for every new title. Several publishers are already delivering ONIX data feeds to their trading partners.

For more information on ONIX,
go to http://www.editeur.org/onix.html

Journal-Oriented Metadata Practices

Journal publishers have been slower to converge on their own metadata standards than book publishers, in part due to a business environment in which metadata was largely the purview of other parties, such as subscription agents, aggregators, and libraries. Although electronic publishing has taken a firm hold in journals publishing, most publishers have focused their energies on their own proprietary journal platforms and formats. This approach is changing as libraries, publishers, and third parties exchange an increasing amount of catalog information, serials subscription data, and other structured data at multiple bibliographic levels (journal, volume, issue, article). It is in this environment that the developers of ONIX have undertaken efforts to extend ONIX to serials.

ONIX for serials

There are three new ONIX records specific to serials that are currently under review: the Serial Title Record, the Serial Item Record, and the Subscription Package Record. The Serial Title Record is the proposed ONIX format for exchanges of rich catalog information. It provides a readily extensible framework for the description of a journal as a bibliographic item, including such details as the cost of an individual subscription item. The Serial Item Record is the ONIX format for alerting, shipping, library check-in functions, and structured multilevel bibliographic description of serial parts. The Subscription Package Record is the ONIX format for communicating a publisher's or agent's product catalog information about subscription packages, along with the Serial Title Record, which carries product catalog information about individual serials.

A Serial Title Record file is linkable to an accompanying Serial Item Record file when more complex price information is required, such as the ability to specify "off-the-shelf" or tailored subscription packages of the kind increasingly being offered by academic journal publishers. This linkage could prove invaluable for sales of journals to consortia.

JWP on the exchange of serials subscription information

Taking ONIX for serials as a starting point, NISO and EDItEUR have recently launched a

Joint Working Party (JWP) to explore the creation of standard formats for the exchange of serials subscription information. At the present time, most such exchanges make use of variable, proprietary formats, except where formats appropriate to a given exchange already exist, such as use of the MARC 21 bibliographic format for library holdings data. In the future, there will probably be more pressure on publishers and others to exchange this information in an accurate, efficient, and secure manner. Development of these guidelines also requires standard identifiers for the key elements in the exchange, including parties to the exchange, aggregations, subscription packages, and the journals themselves.

The JWP is currently functioning as three subgroups: one on identifiers, another on publisher-to-library exchanges, and a third on PAMS (Publication Access Management Service)-to-Library exchanges. The immediate goals of the JWP are to implement pilot programs in these three areas during 2003 and ultimately recommend specific enhancements to the ONIX for serials schema.

For more information on the JWP, go to http://www.niso.org/news/SerialsExchange.html

CrossRef

CrossRef is a DOI-based system for the persistent identification of scholarly content and cross-publisher reference linking to the full text of a journal. CrossRef DOIs link to publisher response pages, which include the full bibliographic citation and abstract, as well as providing full-text access as determined by the publisher. The publisher response page often includes other linking options, such as pay-per-view access, journal table of contents and homepage, and associated resources. CrossRef has recently begun adding books and conference proceedings to its linking network.

Publisher members of CrossRef initially deposit a record for a content item that consists of minimal bibliographic metadata: journal title, ISSN, first author, year, volume, issue, page number, DOI and URL. Depositing metadata with CrossRef involves creating a file formatted according to an XML schema. The following example illustrates an abbreviated metadata record containing both journal-level and article-level elements:

```
<journal>
<full_title>Applied Physics Letters</full_title>
<abbrev_title>Appl. Phys. Lett.</abbrev_title>
<issn media_type="print">00036951</issn>
<issn media_type="electronic">10773118</issn>
<doi_data>
<doi>10.1063/aplo</doi>
<resource> http://ojps.aip.org/aplo/ </resource>
</doi_data>
</journal_metadata>…
<contributors>
<person_name sequence="first" contributor_role="author">
<given_name>Ann P.</given_name>
<surname>Shirakawa</surname>
</person_name>
</contributors>
<publication_date media_type="print">
<year>1999</year>
</publication_date>
<pages>
<first_page>2268</first_page>
</pages>
<doi_data>
<doi>10.1063/1.123820</doi>
<timestamp>19990628123304</timestamp>
<resource>http://ojps.aip.org/link/?apl/74/2268/ab</resource>
</doi_data>
</journal_article>
```

After a publisher deposits a record, CrossRef registers the DOI-URL pair in the central DOI directory and maintains the full metadata set in its metadata database (MDDB). In a separate process, the publisher submits the citations contained in each deposited article to the Reference Resolver, the front-end component of the MDDB that allows for the retrieval of DOIs. By using this method, the publisher can, as part of its electronic production process, add outbound hyperlinks to any of an article's citations that point to content already registered in the CrossRef system.

If the identified content migrates from one production system to another (e.g., pre-print to post-print), or moves from one publisher to another if a journal—or the publisher itself—changes ownership, the publisher need only update the URL in one place in order for the DOI to persist. In all these cases the DOI never changes, which means that all the links to that content that have already been made will still function.

The CrossRef Reference Resolver accepts bibliographic metadata and returns the corresponding DOI. Queries are formatted in a pipe-delimited format containing ten fields for queries against journal holdings and twelve fields for queries against books and conference proceeding holdings. These queries are submitted interactively through a Web browser interface or programmatically via the system's HTTP interface. The resolver will also accept a DOI as input and return the associated metadata. When a query result is returned, the metadata can be presented in either the same pipe-delimited format or as XML.

For more information on CrossRef,
go to http://www.crossref.org

OpenURL and CrossRef. The OpenURL is a mechanism for transporting metadata and identifiers describing a publication for the purpose of context-sensitive linking. The OpenURL is currently on the path toward NISO approval.

A *link resolver* is a system for linking within an institutional context that can interpret incoming OpenURLs, take the local holdings and access privileges of that institution (usually a library) into account, and display links to appropriate resources. A link resolver allows the library to provide a range of library-configured links and services, including links to the full text, a local catalog to check print holdings, document delivery or interlibrary loan (ILL) services, databases, search engines, etc. For the user working in an institutional context, it is often useful to be directed to resources outside the publisher's site. For example, the institution may not subscribe to the e-journal itself but may still be able to offer the user access to the desired article through an aggregated database or print holdings. In addition, the library may wish to provide a range of linking options beyond what is available at the publisher's website.

Information providers are beginning to implement the OpenURL to enable optimal integration with library linking systems. This has caused some confusion among primary and secondary publishers who use the CrossRef/DOI system for cross-publisher links to full text, because of the mistaken perception that the OpenURL and the DOI are competing standards; they are not. CrossRef and the DOI provide persistent identification of scholarly content and centralized linking to the full text and other resources designated by the publisher. The OpenURL is designed for localized linking and enables library-controlled links to a multiplicity of resources related to a citation.

The OpenURL and DOI work together in several ways. First, the DOI directory itself—where link resolution occurs in the CrossRef system—is OpenURL-enabled. This means that it can recognize a user with access to a local resolver. When such a user clicks on a DOI, the CrossRef system redirects that DOI back to the user's local resolver and at the same time allows the DOI to be used as a key to pull metadata out of the CrossRef database — metadata that is needed to create the OpenURL that targets the local link resolver. As a result, the institutional user clicking on a DOI is directed to appropriate resources.

By using the CrossRef/DOI system to identify their content, publishers can make their products OpenURL-aware. Since DOIs can streamline linking and data management processes for publishers, many publishers are beginning to require that the DOI be used as the primary mechanism for linking to full text; link resolvers can then use the CrossRef system to retrieve the DOI if the DOI is not already available from the source, or citing document.

For more information on the OpenURL,
go to http://library.caltech.edu/openurl

The Open Archives Initiative

Although the Open Archives Initiative (OAI) got underway as a means of supporting distributed e-print archives with tools for interoperability, a growing number of publishers now recognize its value as a tool for disseminating publisher metadata. The OAI framework for exposing metadata through the OAI Protocol for Metadata Harvesting (OAI-PMH) is entirely independent of the type of underlying content and the economic models surrounding that content.

OAI-PMH defines an easy-to-implement tool for harvesting XML-formatted metadata from content repositories, or servers. Participation can take one of two forms: data providers use OAI-PMH to expose metadata, while service providers use metadata harvested via the OAI-PMH to build new services. To quote Clifford Lynch, Executive Director of the Coalition for Networked Information (CNI), OAI-PMH is "simply an interface that a networked server (not necessarily an e-print server) can employ to make metadata describing objects housed at that server available to external applications that wish to collect this metadata." (See Lynch, C., *ARL Bimonthly Report* 217 titled "Metadata Harvesting and the Open Archives Initiative" available at http://www.arl.org/newsltr/217/mhp.html.)

For more information,
go to http://www.openarchives.org

Conclusion

Metadata has become an essential part of the publication process. Whether an information resource is published in book or journal form, in print or electronic format, metadata is how the content creator or producer advertises its existence. The richer the metadata record, the greater the possibilities.

As the sea of information grows, being able to locate, discover, link to, search on, re-purpose, integrate, track, exchange, or sell a given information resource all tend to become more complex processes. Good metadata practices reduce some of this complexity and help publishers harness the new opportunities that new technologies will bring.

Where To Go From Here

Without recommending specific products or vendors, the following list provides some information resources on electronic publishing that serve as good starting points:

Since 1997, Sheridan Press has published a series of white papers on information technology and publishing, available at http://www.sheridanpress.com/whitepapers.htm.

NISO standards and guides are available to the public without charge from the NISO website: http://www.niso.org. NISO offers workshops and programs throughout the year focusing on standards and good publishing practices.

Both the Society for Scholarly Publishing (http://www.sspnet.org) and the Council of Science Editors (http://www.councilofscienceeditors.org) offer tutorials on electronic publishing topics.

The Columbia Guide to Digital Publishing, edited by William Kasdorf, is available for online browsing at http://www.digitalpublishingguide.com, and is an excellent, up-to-date resource on XML, content management, and related workflow issues.

Data Conversion Labs publishes a newsletter called DCLNews at http://www.dclab.com/DCLNews.asp that is available via free subscription. It offers in-depth reports, news briefs, and other information about current educational opportunities and resources in electronic publishing.

The NYU Center for Publishing, part of the School of Continuing and Professional Studies (http://www.scps.nyu.edu/departments/index.jsp) offers classes on ONIX and technology and publishing.

Compendium of Cited Web Resources	
Book Industry Study Group (**BISG**)	http://www.bisg.org
Coalition for Networked Information (**CNI**)	http://www.cni.org
Columbia Guide to Digital Publishing	http://www.digitalpublishingguide.com
CORES Forum on Shared Metadata Vocabularies	http://www.cores-eu.net
Council of Science Editors (**CSE**)	http://www.councilofscienceeditors.org
CrossRef	http://www.crossref.org
DCLNews	http://www.dclab.com/DCLNews.asp
Digital Object Identifier (**DOI**)	http://www.doi.org
Dublin Core Metadata Initiative (**DCMI**)	http://dublincore.org
Extensible Markup Language (**XML**)	http://www.w3/org/XML
International Standard Book Number (**ISBN**)	http://www.isbn.org/standards/home/index.asp
International Standard Serial Number (**ISSN**)	http://www.issn.org
International Standard Text Code (**ISTC**)	http://www.nlc-bnc.ca/iso/tc46sc9/istc.htm
Interoperability of Data in E-Commerce Systems (**INDECS**)	http://www.indecs.org
Learning Objects Network (**LON**)	http://www.learningobjectsnetwork.com
Machine Readable Catalog (**MARC**)	http://www.loc.gov/marc
National Information Standards Organization (**NISO**)	http://www.niso.org
NISO-EDItEUR Joint Working Party on the Exchange of Serials Subscription Information	http://www.niso.org/news/SerialsExchange.html
NYU Center for Publishing	http://www.scps.nyu.edu/departments/index.jsp
Online Information Exchange (**ONIX**)	http://www.editeur.org/onix.html
Open Archives Initiative (**OAI**)	http://www.openarchives.org
OpenURL	http://library.caltech.edu/openurl
Serial Item and Contribution Identifier Standard (**SICI**)	http://sunsite.berkeley.edu/SICI
Seybold Reports	http://www.seyboldreports.com
Sharable Content Object Reference Model (**SCORM**)	http://www.adlnet.org
Sheridan Press White Papers	http://www.sheridanpress.com/whitepapers.htm
Society for Scholarly Publishing (SSP)	http://www.sspnet.org

About the Authors and Publishers

Amy Brand joined CrossRef as Director of Business Development in April 2001. Her career spans electronic publishing, book publishing, and academia. She has previously held positions at Ingenta, LEA Inc., the University of Pennsylvania, and The MIT Press where she was an executive editor from 1994-2000. She received her doctorate in cognitive science from MIT in 1989. **Contact Information:** CrossRef, 40 Salem Street, Lynnfield, MA 01940. V: 781-295-0072; F: 781-295-0077; E: abrand@crossref.org.

Frank Daly, until recently, was Executive Director of the Book Industry Study Group. For more than twenty years, Frank was with Baker & Taylor, Inc. During that time he served in a variety of roles, including Director of Marketing, Public & School Libraries, and Vice President, Business Development. Frank is on the advisory boards of Clarion University, NYU's Center for Publishing, and KnowledgeMax, a corporate intranet provider. He is past President of The American Wholesale Booksellers Association. Frank received his MBA from Fordham University and his BBA from the University of Massachusetts. **Contact Information:** 30 Tiberon Drive, Holmdel, NJ 07733. V: 732-817-1774; F: 732-817-1774; E: frankdaly30@hotmail.com.

Barbara Meyers, president of Meyers Consulting Services (est. 1983), provides expert advice and experienced operational support to professional societies, scholarly publishers, and their supplier communities in the areas of management, marketing, planning, and research. One of the founders of the Society for Scholarly Publishers (SSP), Barbara currently serves on the SSP Board of Directors and is a past president of the Council of Science Editors (CSE). She holds two degrees from George Washington University: her bachelor's in science journalism and her master's in science, technology, and public policy with a specialization in technology assessment. For additional information on Barbara's background, services, and client list, please visit the MCS web site at http://www.MCSone.com. **Contact Information**: Meyers Consulting Services, 1836 Metzerott Road, Suite 1003, Adelphi, MD 20783-3448. V: 301-434-6249; F: 301-434-0126; E: mcsone@erols.com.

NISO Press is the publishing program of the National Information Standards Organization (NISO). NISO, a nonprofit association accredited by the American National Standards Institute (ANSI), identifies, develops, maintains, and publishes technical standards to manage information in our changing and ever-more digital environment. NISO standards apply both traditional and new technologies to the full range of information-related needs, including retrieval, re-purposing, storage, metadata, and preservation. **Contact Information:** NISO, 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814. V: 301-654-2512; F: 301-654-1721; E: nisohq@niso.org.
Website: http://www.niso.org.

The Sheridan Press provides a full range of printing and publishing services and technology innovations to associations, publishers, and university presses within the scientific, technical, and medical journal markets. **Contact Information:** The Sheridan Press, 450 Fame Avenue, Hanover, PA 17331. V: 717-632-3535; F: 717-633-8900.
Website: http://www.sheridanpress.com.

THE SHERIDAN PRESS

Printing and Publishing Services
450 Fame Avenue
Hanover, Pennsylvania 17331

For more information about The Sheridan Press or to request additional copies of the Metadata Demystified White Paper, call Prudi Showers at 717-632-3535 or contact her by e-mail at pshowers@tsp.sheridan.com or fax this form to 717-633-8900.

Name ______________________ **Title** ______________________

Company ______________________

Address ______________________

____

____

Phone Number ______________________ **Fax Number** ______________________

E-Mail Address ______________________

The Sheridan Press Publications and Literature

I am interested in additional copies of the following White Papers:

_____ Metadata Demystified (in collaboration with NISO Press) (7/03)

_____ Member Recruitment (4/03)

_____ Digital Art (5/02)

_____ Implementing Information Technology Systems (1/02)

_____ Marketing Reprints (10/01)

_____ Marketing Scholarly Journals (5/01)

_____ Digital Archiving in the New Millennium: Developing an Infrastructure (11/00)

_____ Improving Journal Quality with Process Improvement Methods (5/00)

_____ Digital Workflow: Managing the Process Electronically (3/00)

_____ How to Make the Most of Reprints (5/99)

_____ The Future of the Print Journal (2/99)

_____ Outsourcing (6/98)

_____ Archiving (9/97)

I am interested in more information about:

_____ The Sheridan Press

_____ Sheridan Reprints

_____ The Sheridan Group

ISBN 1-880124-59-9