Streamlining Book Metadata Workflow

by

Judy Luther

Informed Strategies

102 West Montgomery Ave. #B
Ardmore, PA 19003

June 30, 2009

A White Paper
prepared for the

National Information Standards Organization (NISO)
and
OCLC Online Computer Library Center, Inc.

Published by

NISO
One North Charles Street
Suite 1905
Baltimore, MD 21201

and

OCLC, Inc.
6565 Kilgour Place
Dublin, Ohio 43017-3395

ISBN 978-1-880124-82-6

About NISO White Papers

NISO White Papers are contributed or solicited papers that address an issue that has implications for standards development. White Papers can be viewed as a pre-standardization activity. A NISO White Paper might define and explore some of the questions that come into play before formal standardization work is started. Or, a NISO White Paper might identify areas that are opportunities for standards development and suggest possible approaches. All White Papers are posted on the NISO website (www.niso.org).

Table of Contents

Figures

Introduction

The metadata that enables readers to identify and acquire books online also drives multiple functions for the publisher supply chain and for library workflows both before and after publication. With the amount of digital book content growing rapidly and the number of formats multiplying, the metadata required to support the discovery, sale, and use of content by a global audience is increasing exponentially.

Paradoxically, easy access to a huge amount of metadata on the web has devalued its importance with some believing that any type of data can be used to find books. Yet the ability to efficiently transmit and correctly display metadata that supports decisions to acquire books has made the underlying structure and exchange of metadata even more important to the success of the publisher supply chain and libraries. As stakeholders become aware that metadata transfer is essential for book selection and sales and for smooth business operations, they become more involved in the development and adoption of standards and systems to support metadata exchange.

In the current discovery environment, it is difficult to measure what is *not* found and extremely difficult to quantify the impact and cost of poor, incomplete, or missing metadata on business and collection analysis decisions that ultimately affect consumers. Expectations to view book information online means that metadata from internal publisher systems is transmitted and displayed on the web rather than edited for print marketing materials that are directed to a specific audience. This evolution in marketing and communications requires precise distribution of metadata to partners in the supply chain.

These factors combined with the economic downturn have put exceptional pressure on publishers, wholesalers, booksellers, metadata vendors, and librarians to "do it right" but do so more efficiently. Those working in this space know that it is more easily said than done. The potential for process improvement requires a holistic view of the metadata workflow across the entire system. Both NISO & OCLC share the vision of an environment where metadata is exchanged seamlessly between different stakeholder systems and they are focused on reducing the costs of this exchange for all participants in the supply chain of data and content.

To support discussions among stakeholders, this industry overview defines their roles, examines their use of metadata, and describes the landscape. More than 30 industry representatives were interviewed to discuss the nature and volume of the metadata they manage, changes in its lifecycle, current issues and challenges, and ideas for improvement. Their collective comments and insights form the basis of this report and the Book Metadata Exchange Map which depict the complex environment and provide an opportunity to identify redundant efforts and increase efficiencies within the book industry supply chain.

Stakeholder Perspectives

The perspectives of stakeholders in the book industry supply chain reflect the diverse roles of those who create, enrich, maintain, distribute, and use metadata. The volume of book metadata that is generated and managed by stakeholders varies widely: publishers may be small with only 1 title or very large with 20,000 titles per year; metadata vendors and booksellers manage around 14 million records; the Library of Congress is the world's largest library with 32 million books; and WorldCat contains over 138 million cataloged book records among OCLC's member libraries, with a new record added every 10 seconds.

Books are being digitized rapidly from print library collections, titles in the public domain, and publisher backlists. Not only is the number of titles in the supply chain growing, but multiple records are needed for every title since separate ISBNs are required for each e-book and media format, resulting in a massive increase of 8-10 times the current volume of records that metadata vendors and booksellers must manage.

This volume of metadata makes it critical to clearly present a title and the options for reading or listening to it so that customers can select their desired format. Concern about the increased volume of metadata was a recurring theme among different stakeholders.

Publishers

Publishers are working to digitize their backlists since current print-on-demand technology allows them to offer a larger list of titles without incurring the expense of print inventory and warehouse space. Given the low cost of storing digital files and the ease of printing copies when they are ordered, an "out-of-stock" or "out-of-print" status is no longer meaningful. As publishers convert their titles a few have begun to assign metadata to elements below the title level to provide greater access to book chapters and to prepare for future revenue opportunities.

Approximately 20 very large publishers release thousands of titles each year while hundreds of midsize publishers produce anywhere from a dozen to a few hundred titles annually. Thousands of small publishers comprise the long tail and produce only a few titles each year. ONIX for Books is used by the leading publishers to structure large amounts of descriptive and administrative data that is transmitted in XML to supply chain partners. However, Excel files, proprietary formats, and print catalogs are more common methods of compiling metadata among mid size and small publishers. The challenge of accommodating these myriad and inconsistent formats adds to the difficulty of managing metadata across the spectrum of metadata creators.

Publishers processing large files acknowledge that significant progress has been made in handling metadata in ONIX over the last ten years. ONIX is a flexible but complex format that can be challenging for large publishers and daunting for small and medium publishers. Support for providing metadata in a transferrable format must originate at a senior level in organizations and most publishers do not yet appreciate the impact of metadata on their buyers and supply chain partners in the discovery, sale, and use of their content. Publishers can rely on vendors of publishing software and web based services such as Anko, NetRead, BooksoniX, and Firebrand, to manage the distribution of their metadata in ONIX.

While use of the ONIX standard has improved the amount of data available and facilitated its distribution, both publishers and vendors agree that the use of ONIX as a format does not guarantee that metadata is timely or accurate. The Book Industry Study Group (BISG) in the US, the Book Industry Communication (BIC) in the UK, and Booknet Canada offer publisher certification programs that evaluate ONIX files and recognize the publishers who meet criteria for data quality and timeliness. Unfortunately, participation in these programs has been limited. This may in part be due to the use of legacy business systems or content management systems that would require significant infrastructure expenditures to upgrade, affecting a publisher's ability to implement new or changing standards.

Metadata Vendors

A wide variety of organizations collect, enhance, and redistribute metadata. These institutions include registration authorities (such as Bowker and Nielsen Book), cataloging service agencies (such as BDS in the UK), and member organizations (such as OCLC and CrossRef). Although national libraries and wholesalers also perform these functions, it is not their primary role.

Commercial metadata vendors sit at the nexus of the supply chain. They aggregate data from all sources in various formats (ONIX, Excel, print), ingest it into their own system, enhance it according to specified standards, and then deliver records in MARC, ONIX in XML, and other formats. They collect and disseminate records for publishers, enhance and produce MARC records for libraries, and deliver robust records to booksellers. They are acutely aware of the need for interoperability and they play a leading role in the development and application of standards. For example, Nielsen BookData is part of the review process for publishers applying for certification in the UK with BIC's Product Data Excellence Awards for timely and accurate information.

Of the approximate 300,000 records that Bowker adds each year, 50% are received in ONIX, and 45% in Excel or another electronic format, with only a few still submitted in paper.

Both Bowker and Nielsen Book standardize series titles and contributors, perform authority control functions, and provide links to publisher records. They may also provide additional content such as: first chapters, media mentions, cover images, tables-of-contents, review citations, bestseller citations, author biographies, book awards, readership levels, user generated reviews, and ratings and recommendations.

Bowker serves as the ISBN agency for the US and Australia and Nielsen Book serves as the ISBN agency for the UK and Ireland. In this capacity they assign prefixes to publishers. In 2008 Bowker, Nielsen Book, CISAC, and IFFRO founded the International Standard Text Code (ISTC) agency which promotes, coordinates, and supervises the system used to identify "textual works" so that different formats of the same work can be collocated.

A commercial cataloging service in the UK, Bibliographic Data Service (BDS), is the vendor to whom the British Library has outsourced their Cataloging in Publication (CIP) program. BDS creates ~75,000 prepublication records each year and has created crosswalks from ONIX data to MARC 21. Their staff is mostly trained in-house and it takes a year for new staff to learn the Dewey Decimal System, Library of Congress (LC) Classification System, LC's Name Authority (NACO), and LC Subject Headings (LCSH).

OCLC is a member organization of libraries in 112 countries. The WorldCat database contains 138 million records, which equals the entire output of the Library of Congress and the British Library combined. Approximately 60,000-120,000 records are contributed by Brodart, Ingram, Baker & Taylor and others through the Vendor Record Contribution Program. OCLC maintains a staff of over 70 metadata specialists and catalogers, in the U.S. and Canada, who create records relating to materials in multiple formats and languages for special collections, new acquisitions, and for publishers and vendors.

OCLC is using the principles outlined in the Functional Requirements for Bibliographic Records (FRBR) to display records for related items together in WorldCat, its global public catalog that is accessible through Google. OCLC's NextGen Pilot involves working with publishers to develop the crosswalks between ONIX and MARC; they are launching a service that will provide enriched ONIX data to publishers and suppliers to enhance their systems.

Book metadata is currently a small but growing component in two organizations that have robust files of journal metadata. CrossRef is a member organization of publishers with 1.6 million book DOIs (Digital Object Identifiers). The Knowledge Base in Serials Solution contains over 1 million e-book records. Originally developed to support journal literature, these services link from citations to the full text.

Wholesalers

The largest book wholesalers, Baker & Taylor and Ingram, have databases that grew by 10% or more last year. Although new books are estimated to be around 200,000 titles per year, book vendors are handling approximately 700,000 new records due to the increasing number of books available in digital form and the growing volume of media. On average each title will have 2-3 records created for its different formats and editions. Records for new titles are likely to be updated 3-7 times, mostly before publication.

Publishers that provide their data in either ONIX or Excel account for 95% of the new records. While CIP data is used by the distributors to enhance records, publisher data takes priority. When the print book is received, all records are reviewed for accuracy, style, and consistency.

Increasingly, libraries expect vendors to deliver MARC formatted bibliographic records with the book and there is a recent surge in demand for materials that arrive processed and "shelf ready." Budget pressures in libraries, combined with a desire to streamline their acquisitions processes, have accelerated this trend.

International vendors such as Casalini, Harrassowitz, Pupil, and others that sell to the US academic market are working with the Program for Cooperative Cataloging (PCC) to train their staff to create high quality MARC records on books from their countries. MARC records are expensive for local libraries to produce, especially for foreign language titles, and libraries look to vendors with language expertise to deliver the MARC records for these books.

Booksellers

Barnes & Noble knows that descriptive metadata increases sales and that the lack of good metadata costs them sales. Each year thousands of orders for titles which total millions of dollars are cancelled because metadata records indicate that the status is "out-of-stock with no due date" or there is no distributor listed. To encourage the use of descriptive metadata and accurate and timely status updates, Barnes & Noble asks publishers to supply 44 data elements—14 more than the 30 core elements required for certification by the Book Industry Study Group (BISG). Publishers that comply appear on an "A" or "B" List recognizing the quality of their metadata. These lists are shared within the industry to stimulate better metadata to increase book sales.

As the point of sale, booksellers must have accurate and timely data on the changing price and availability of titles. While ONIX is considered effective in delivering extensive bibliographic data and is unsurpassed for granularity and definition, EDI (Electronic Data Interchange) is a lightweight standard transmission format that is effective for updating volatile elements such as price or status. Warehousing organizations that are poor at supplying descriptive metadata can provide price and availability with increasing speed, sometimes twice a day or even hourly. Of the 43 million records processed last year by Nielsen Book, 86% were updates on price and status.

E-book vendors use MARC records as part of their service offering to libraries, providing access for selection and ordering purposes and then delivering the records to the library with access to the book. Metadata for e-books is used for multiple applications: to identify specific titles, to export citations, to generate a title list for library systems, to support patron initiated acquisition in libraries, and for resellers to order titles.

Large web booksellers such as Amazon ingest metadata from many different sources so standards and best practices have to be widely adopted to have significant impact. Although the majority of North American books have ONIX, it does not apply to audio or video formats sold by Amazon.

For mass merchandisers, such as Wal-Mart and Target, books comprise a small percent of their overall volume and they do not use ONIX. Preliminary conversations have begun between book industry representatives and GS1 which coordinates global standards for the supply chain.

National Libraries

National Libraries play a key role in their respective countries in gaining consensus on the development of standards related to data exchange. Although the Library of Congress (LC) is not designated or funded as a national library, it effectively functions that way for the US. In 2008, LC released its *Working Group Report on the Future of Bibliographic Control* which presented a vision of metadata creation that is collaborative, decentralized, international, and web-based. A report is due on the next phase, managed by R2 Consulting, which will provide LC with an analysis of the MARC record marketplace in North America.

LC has served in a leadership role internationally developing guidelines and providing training to support the cost effective creation of high quality MARC records through a distributed and collaborative effort. Records enhanced by the participants in the Program for Cooperative Cataloging (PCC) are highly regarded for the use of authorized names and subject headings. The PCC recognizes the shift away from a single standard such as MARC towards interoperability among multiple standards for metadata that is used and recycled by the book industry, rights management, and library and information sectors. Increasingly metadata is exchanged between machines rather than humans which reduces errors and increases efficiency rather than relying on manual distribution.

Both BL and LC manage Cataloging in Publication (CIP) data programs that serve the dual role of providing prepublication data distributed in a MARC format to libraries for ordering and processing and to book trade vendors alerting them to new titles typically 3-6 months before publication. The CIP record appears on the back of the title page of each book and this metadata is distributed electronically as MARC records.

Although copyright laws vary by country, the British Library (BL) and LC receive on "legal deposit" copies of works published in their respective countries. The BL is exploring how to best work within a voluntary framework for digital materials that are deposited in the UK. Currently, it captures the output from multiple publishers: full text e-journal articles in XML are converted to the NLM's DTD with books in PDF, Word, or ePub being left in the original format. The use of XML formats potentially enables publishers to extract coded elements from the item into an ONIX record. The BL is also experimenting with extracting metadata from full text XML and text based PDF files. The aim will be to ultimately use MODS format for descriptive metadata together with a METS "wrapper."

For all items received the BL employs different workflows based on the appropriate level of cataloging for the item, expediting fiction with an abbreviated record, and providing full records for scholarly works. If records are expected to be available from LC, the BL will use those records or look for a basic record to upgrade. Last year professional catalogers created or upgraded records for 80% of the 350,000 titles processed by LC and 55% of the 260,000 titles processed by the BL. One estimate is that 65% of OCLC's WorldCat records are abbreviated and require authority work on the author or series and the addition of notes, summaries, tables-of-contents, and genre headings.

Local Libraries

At the local level, libraries acquire electronic journals and books that are often packaged and sold as a collection, and increasingly they rely on the publisher and/or book vendor to provide MARC records with this bundled content. However, libraries actually need these records at the point when they place an order since the MARC record format is needed in their integrated library systems for acquisitions. Preliminary records can subsequently be replaced en masse or upgraded individually when the item or collection is made available to users.

Library processes have shifted from cataloging individual titles to managing the ingestion of records from other sources and then upgrading these records as needed. As a result the percent of materials handled by catalogers has declined substantially and large research libraries may catalog less than 30% of the 100,000 items they receive in a year. However, much effort is still expended on adapting records received from outside sources to conform to locally defined cataloging practices.

Many libraries are investing in workflow analysis, seeking to streamline their operations and to be more cost effective. They may have one librarian managing acquisitions and one managing cataloging, supported by paraprofessional staff trained on the job, to enhance basic records that have been bought or shared. A parallel trend is that there are fewer catalogers available to replace those who will retire in the next five years.

While studies over the last ten years have shown that MARC records in an online catalog directly affect e-book usage, some librarians question the value of subject headings for a full text database when the content is accessible via Google. If use patterns for e-books are similar to e-journals, then the majority of readers will access the content by linking from a search result, bypassing the online catalog and the front page of the database.

A UK study by Ked Chad for Research Information Network released in 2009 examined metadata flows for print and electronic books and journals and proposed savings from using networked environments to support metadata that would be more openly shared.

Google

Google's efforts to digitize millions of books and to create the Book Rights Registry to handle payments involve them directly in dealing with familiar book metadata issues. The use case for metadata is to identify works and establish relationships between them. Currently, series and multivolume works present challenges but these are being addressed.

Many of the books being digitized have MARC records but lack ISBNs as they were published before the ISBN was developed in the 1970s. For current books, Google Book Search is ingesting both ONIX and MARC records to collect the best possible metadata. Their preference is for good quality MARC over poorly formed ONIX and well formed ONIX over poor quality MARC. While ONIX has missing data elements, MARC data may exist but isn't machine-friendly in terms of understanding the data.

There are a number of librarians working at Google on metadata issues and Google is also working with OCLC. Google supports standards and follows new ones as they develop and are adopted over time. Google is also working to develop algorithms that may solve the problem of distinguishing related works.

Metadata Workflow

Lifecycle

Metadata flow begins with the publisher and ends with the buyer or reader. In between it serves many purposes among multiple trading partners in the supply chain. The Book Metadata Exchange Map (Figure 2 on page 17) reflects the major stakeholders, the flow of metadata, and the points at which external quality controls (e.g., BISG's certification program, Barnes & Noble's grading system, or PCC training) encourage higher quality metadata.

Prepublication data is fluid; ONIX records are typically released 3 to 6 months ahead of publication and changes from this first release to the time of publication can occur to the title, subtitle, or size of the work. Once the book is published, the bibliographic data stabilizes. Upon receipt of a book, wholesalers and libraries conduct a metadata check to verify that the ONIX or MARC records they have received are correct, and then update or enhance their records. The flow of metadata naturally slows at this point and subsequent changes are made primarily to the price and status. The cumulative effect of these changes requires metadata vendors, wholesalers, and booksellers to touch each record on average 3-5 times. (86% of the 43.2 million record changes at Nielsen Book last year were adjusting price and availability.)

CIP data is created prior to publication and must be updated to reflect the number of pages in the published form as well as any changes in descriptive information. Post publication, some libraries may update their MARC records to include author death dates or new subject headings.

Many libraries obtain their MARC records from OCLC which makes their holdings publicly available through WorldCat and supports interlibrary loan. The sale of digital collections accompanied by MARC records is effectively pushing cataloging services upstream from the library to the publisher.

Standards

Over time the publishing supply chain and the library communities developed standards and best practices to address their objectives. Table 1 shows standards in use or under development by each community to address similar issues. It also distinguishes two types of data elements: bibliographic, which describes the work for identification purposes, and administrative, which is focused on transactions important to the respective stakeholders. MARC emphasizes bibliographic metadata, while ONIX as a newer standard contains more descriptive metadata and the administrative data needed by supply chain partners.

While the ISBN standard works in both communities to identify the book, libraries and publishers have chosen very different subject schemes to describe the book's content. For example, libraries have opted for detailed hierarchical subject headings related to classification and searching (LCSH has 300,000 terms) while publishers are using much lighter weight tools (BISAC has only 3000 terms) designed for store placement. Each of these evolved based on the specific application for which it was intended.

Recent work on the ISNI (International Standard Name Identifier) is an example of a collaborative project. The ISNI is considered a bridge identifier that links proprietary data maintained by registration agencies with library authority data maintained by national libraries in the VIAF (Virtual International Authority Files). OCLC has developed a proof of concept to demonstrate a methodology for using rich metadata algorithms to link identities.

Table 1: Different Standards Initiated by Different Communities for Similar Issues

Publishing Community			Library Community	
Standards	**Purpose**	**Type of Data**	**Purpose**	**Standards**
ISBN	specific publication	Bibliographic	formats	ISBN
ONIX	sales information / discovery		discovery for use	MARC
BISAC, BIC	subjects	Bibliographic	subjects	LCSH, SACO, Wilson
ISNI	authors/royalties	Bibliographic	author/contributor	NACO, VIAF
ISTC	related works	Bibliographic	related works	FRBR
ONIX, EDI	price	Administrative	fund codes	Integrated Library System (ILS)
ONIX, EDI	status	Administrative	holdings	ILS or OCLC

ONIX

ONIX was developed in the late 1990s to enable publishers to transmit to wholesalers and retailers rich product metadata including marketing collateral used to promote books. Structurally, ONIX is a schema with a detailed list of code values that uses XML as the message structure. ONIX has sufficient flexibility to adapt to each organization's internal systems which means that those implementing ONIX tweak the import programs for each publisher and this has succeeded in reducing questions on fielded data by 40%.

The typical ONIX data workflow for the publisher is as follows:

- Basic metadata originates in the editorial process and is stored in an Excel file, a press database, or a hosted solution. It can be the by-product of a system used for publication management.
- Marketing and production data is added and associated with a content asset database that contains the full text of the work.
- Galleys are sent to LC and used to create a CIP record that is included on the title verso or back of the title page.
- Sales information and catalog data is sent to wholesalers and metadata vendors in different formats as required: an ONIX feed is generated for print titles; an Excel file is often used for e-books; and metadata in XML is sent to A&I services.
- Large image files are sent separately and are linked to from the ONIX records.
- Post publication reviews are associated with book data, and the price and status are updated in ONIX feeds distributed to metadata vendors, wholesalers, and booksellers.
- Since few publishers can generate MARC from their systems, most work with metadata vendors or consultants to produce MARC records for distribution with their content.

ONIX 3.0, which was recently released, is not backwards compatible with earlier versions, but it includes several key enhancements that improve handling of digital products, allow partial record updates, use the ISTC to connect different products of the same work, include additional marketing collateral, and improve the handling of series. Publishers and vendors may need to maintain both versions for a transition period and this could be challenging in the current economic climate, although there is also a cost to being backwards compatible.

MARC

Created in the 1970s, MARC (Machine Readable Cataloging) was designed as a data exchange format with tagged fields and has since evolved into a family of standards, rules, and guidelines. Cataloging served the dual role of documenting a library's collection and creating reliable access points that would allow readers to find titles. Today MARC 21 is an international standard with crosswalks to national flavors of MARC.

At the National Library Level

MARC records are initiated early in the publications lifecycle when publishers send galleys to a national library (Library of Congress or BDS for the British Library) and promptly receive a Cataloging in Publication (CIP) record for inclusion in the published work. The CIP program was created to provide authorized headings (author, title, and subject) along with classification numbers (Dewey or LC) on books that libraries were likely to acquire. LC and BL have different criteria for excluding categories of materials that are designated as out of scope and not all eligible books are submitted. (See the LC exclusion list at: http://cip.loc.gov/scope.html and the BL list at: http://www.bl.uk/bibliographic/exclude.html.) The volume of CIP records for the BL and for LC is roughly 50,000, which is less than 20% of the titles catalogued at either institution.

CIP workflow at LC is completed within 10 working days:

- The electronic copy of a portion of the work that includes the front matter, copyright, title page, index, and first or last chapter is sent to LC.
- The 10 largest publishers send ONIX records for most of the 55,000 CIP titles though ONIX data is not used in creating CIP.
- LC extracts the summary and table of contents from ONIX and links to it from the MARC record.
- Authority work is performed to uniquely identify the author/contributor; new authority records are created for approximately 1/3 of titles. Appropriate subject headings are applied.
- A nearly complete MARC record is created.
- CIP data is sent to the publisher via e-mail to be included in the book during production but it is not provided in XML or fed into the publisher's ONIX data stream. (A pilot project is being conducted at LC in 2009 to address questions arising from the process of converting ONIX data to MARC.)
- MARC records are distributed via the Cataloging Distribution Service to metadata vendors, booksellers, and wholesalers.
- Upon publication the MARC record created for CIP is updated.

The British Library outsources the creation of CIP to Bibliographic Data Services (BDS) in Scotland which has created crosswalks from ONIX to MARC. LC is supported by the National Agriculture Library (NAL), the National Library of Medicine (NLM), and university libraries in producing CIP records in a timely manner.

The incentive for publishers to obtain CIP is that prepublication announcements are made to the book trade. Libraries benefit by having MARC records available to use in ordering forthcoming books. National libraries also use these records to identify which titles they have not received on legal deposit as specified by their copyright law.

At the Local Library

The MARC data workflow in a library that acquires content is:

- To begin the ordering process the library enters a brief record from the publisher, wholesaler, CIP from LC, or an OCLC record. If no record exists, the library enters a temporary record.

- Once the book is delivered the initial record is upgraded and local holdings data are added.
- The library may subscribe to additional descriptive publisher data from a metadata provider.
- A few libraries use new online catalogs that collect subject tags contributed by users.

ONIX often includes additional descriptive information (table of contents, author bio) that is needed by those making book selection decisions online. A growing number of librarians acquire this rich descriptive data which is available through metadata vendors such as Bowker, Nielsen Book, and BDS.

Metadata Silos

The need to create crosswalks between ONIX and MARC reflects how these communities have addressed their concerns separately. The online environment, however, requires interoperability and suggests a big picture view that involves both the library and the publisher supply chain in conversations about the potential for more effective and efficient approaches to managing metadata. Figure 1 illustrates the metadata silos among the stakeholders.

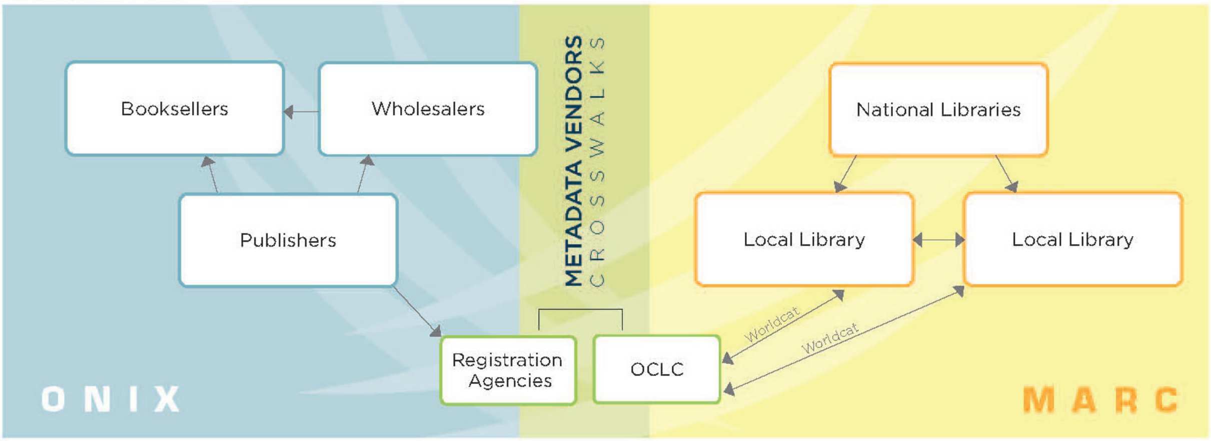

Figure 1: Metadata Silos

Quality

The quality of metadata is best measured by its effectiveness in supporting the goals of the stakeholders to ensure book discovery for purchase and/or use. The data elements required to achieve this goal are expanding to accommodate the online environment where descriptive metadata (author bios, cover images, book summaries, tables of contents) plays a key role in the book selection process online.

The quality of the process of metadata exchange is determined by its delivery and usefulness to other stakeholders in the supply chain. The Global Data Synchronization Network (GDSN) has defined characteristics of good quality metadata as:

- consistent (from record to record and file to file),
- core elements complete,
- according to an agreed upon standard, and
- distributed in a timely manner.

GDSN supports supply and demand chains internationally by enabling partners to have consistent data in their systems at the same time since there is one "owner" of the metadata and they control what is distributed as an update. "By improving the quality of data, trading partners reduce costs, improve productivity, and accelerate speed to market. Good quality data is foundational to collaborative commerce and global data synchronization."

Large publishers distribute files weekly that contain thousands of records and updates which can take a weekend to process. The growing volume of metadata on each title requires machine to machine interfaces with minimal manual review or intervention. Issues that arise are likely to be problems of either definition (i.e. clarity on what a "series" is) or implementation (whether the series appears in the ONIX record and where).

Elements of Quality

One approach to identifying data elements is to consider whether they are intrinsic (describing the work itself) or extrinsic (associated with the work).

- **Intrinsic** data elements are those that are descriptive of the work—such as the title, author, publisher, date of publication, and tables of contents—and are used to identify it, such as series, subjects, classification numbers and authority work for authors in MARC records. These elements form the basis of bibliographic metadata.

- **Extrinsic** data elements are those which are used to support the marketing and administrative activities of various stakeholders. For publishers and booksellers these include such data as price, release date, and book reviews; for libraries these include such items as fund codes and holdings.

When digital products are created in XML, much of the intrinsic metadata can be derived from the content since it can be lifted directly from the tagged work rather than assigned or attached to it by subsequent analysis. Metadata then becomes a surrogate for the content which is in XML.

Programs to Encourage Quality

Both publishers and librarians have developed vehicles to encourage the adoption of standards and best practices.

Publisher programs

- BISG in the US offers the Publisher Certification Data Program (PCDP) that mandates the use of 30 of the 230 data elements that publishers can send in either XML or Excel.
- BIC in the UK offers Product Data Excellence Awards that provide publishers with feedback on the quality and consistency of their data. These award guidelines can vary based on different book trade practices in different countries.
- Booknet Canada offers a bronze, silver, and gold level of certification for publishers and vendors who submit samples to auditors for evaluation.
- Barnes & Noble grades publishers on 16 additional data fields in addition to the 30 BISG fields which include data such as cover images that impacts books sales. B&N's goal is for their top 1000 publishers to achieve an A+ rating.

Library programs

- The Program for Cooperative Cataloging (PCC) administered by the Library of Congress provides training to catalogers who are then authorized to submit records according to standards (BIBCO for books and CONSER for serials) with the appropriate authority controls for authors (NACO) and subjects (SACO).
- Recently the PCC training program has been extended to international book vendors to enable them to provide high quality MARC records for use by many libraries, sparing the libraries

redundant efforts at the local level. These include Casalini Libri in Italy, Harrassowitz in Germany, Garcia Cambeiro in Argentina, Puvill Libros in Spain & Mexico, and East Asian vendors in China and Korea.

- A recent project from LC—with participation by the national libraries of Germany and France and support by OCLC—is the Virtual International Authority Files (VIAF), which has a goal of creating a cross-referenced file of 7.8 million source authority name records that combine data on individuals, link related records, and display the name in the local language.

Use of these programs is growing slowly yet is having a positive impact on the quality of metadata available to all stakeholders in the supply chain. The efforts of different stakeholders to improve the quality of metadata in the supply chain are indicated on the Book Metadata Exchange Map (Figure 2 on page 17).

Opportunities

During the development of this report, OCLC hosted a Symposium for Publishers and Librarians to explore metadata needs and practices. Ideas and questions raised in conversations at the Symposium and during conversations with Stakeholders in preparing this document reflect their current thinking on the state of the metadata supply chain and ways to streamline it to improve its effectiveness and reduce costs. Current initiatives and ideas are organized into three broad topics: Identifiers, Organizational Schemes, and Best Practices.

The economics of metadata could suggest a model with a centralized bibliographic record that would be enriched over time with decentralized administrative records exposed to the respective communities. Another approach would be to have a record with fields that contain fixed data elements and then link to variable data elements. Diverse stakeholders expressed concern about a model that relied on a single source of metadata. It should be possible to homogenize the way data is shared rather than attempting to homogenize the data.

As publishers increasingly recognize the importance of robust, timely, and accurate metadata in the supply chain, libraries are increasingly aware of the redundant costs of creating metadata records locally, and look to publishers and providers to supply them with MARC records. Digital works naturally support metadata creation and a holistic view of the system can help streamline the subsequent enhancement process.

Identifiers

Identifiers are used both to disambiguate items and to present related items together; they determine if authors and works are treated separately or alike.

Authors

The ISNI, the International Standard Name Identifier (ISNI), was referenced most often as the potential solution for author identity and it makes use of existing files among diverse stakeholders including the authority files of national libraries. A third of new works in 2008 had authority records created by libraries indicating that these were new authors. While the birth/death dates used in authority records may be useful to distinguish authors, publishers expressed concern about including these dates in the ONIX record due to the risk of inconsistent display across different bookseller systems.

Multiple solutions to correctly identify authors are being developed throughout the supply chain and requirements for data elements vary across publishers and stakeholders. Reproduction Rights Organizations (RROs) and some publishers of fiction require a system that deals with aliases in order to effectively manage royalties for a single author writing under different names. Google's Book Rights Registry requires clear identification of authors for accurate payments. Science publishers need to distinguish among a global pool of authors and need to retain honorifics with fields for the prefix and suffix. As author identifier systems emerge, standards will be needed to drive the clean-up of information on contributors, especially with so many older works being digitized.

Individual Works

The proliferation of ISBNs for different formats of e-books is encountering resistance from some publishers who want one ISBN for any e-book format. Content below the chapter level is being identified with an actionable ISBN or a DOI as publishers experiment with different models of providing access to book content below the title level.

The ISBN is an established standard, however a few cautionary notes were offered on current applications. As new ISBNs are assigned to older works when they are digitized, libraries must be alert to avoid duplicating a title they already have. Some titles that were assigned ISBNs were never published because the posted title was used to test for market demand. Questions have been raised about whether the ISBN is appropriate for digital scanning projects where items may not be traded.

Titles physically handled by booksellers and mass merchandisers are identified with an EAN or UPC which are both GTIN (Global Trade Item Numbers) approved by the GS1 and used to track the movement of inventory in the supply chain. Both the EAN and the UPC display in the form of a barcode.

Series

Series were frequently mentioned as problematic because the designation of a series in trade publishing is driven by marketing factors rather than a library's structured definition of what defines a set or series. In some cases the series title is more important than the book title or the publisher wants individual titles grouped together and places the series in the title field or makes it the subtitle. The publisher may also be limited in the way that titles were set up in legacy systems which means that the fields for series data are often inaccurate or incomplete. ONIX Version 3.0 addresses series and there is hope that in the online environment where publishers connect directly with readers, they can shift to a more universal view.

Related Works

The ISTC, International Standard Text Code, is designed to create associations among manifestations of the same work and allows users to find all the expressions of a work. The ISTC Agency is now a legal entity with a published schema. A pilot project due to conclude in 2009 enables linking of backlists to optimize discovery and to develop workflow with publishers so that an ISTC is incorporated with the current bibliographic metadata. To expand adoption, perhaps OCLC and/or the national libraries could assign ISTC's retrospectively for existing works.

Subject Schemes

Publishers have used subject schemes to organize print books for browsing in physical stores, while libraries have used subject schemes as search points in online catalogs. With the number of online sales growing rapidly, some publishers expressed the belief that more detailed subject terms would be useful. For example, a title coded "French Cooking" instead of just "Cooking" will appear in online results sets for either search.

Publishers employ specific fields in ONIX that do not appear in the MARC record such as the audience code and age range of the target audience. When coding books to be released, what may appear to be contradictory categories—for example, when a book is tagged both adult and juvenile, or both fiction and history, may be done deliberately by publishers for product placement for different market segments. Making use of this data in libraries would require an understanding of the motivation in creating it.

Publisher Schemes – BISAC & BIC

BISAC is a relatively closed list with over 3000 categories and 50 high level categories that are reviewed and revised annually, based on an analysis of titles published and user needs. Designed for English language books, the US BISAC and the UK BIC codes are being aligned with each revision. OCLC has begun mapping BISAC Subject Headings to the Dewey Decimal Classification (DDC) system.

Originally intended for store placement, these codes did not have widespread adoption among trade publishers until they dealt with online search. Application of BISAC or BIC codes to each title is now required by many major supply chain players. The codes have come to drive much more than physical in-store placement and are integrated into web search design, have become essential to mechanisms for tracking point-of-sale statistics by category, are used for building wholesaler selection lists for retail and library markets, and are now integral to multiple proprietary business intelligence applications.

Library Schemes – LCSH, Sears, MeSH

Since they have been considered key to the discovery of books, subject terms have been treated by libraries with the same degree of authority control as the other primary access points, such as author. The Library of Congress Subject Headings (LCSH) has more than 300,000 terms, which is 100 times the size of BISAC at 3,000. Both wholesalers such as YBP (part of Baker & Taylor) and e-book vendors such as ebrary use LCSH and call numbers in working with libraries to establish profiles for selection purposes and to support online browsing.

Ideas

- Different subject systems such as BISAC, BIC, and LCSH codes are used in the US and other codes are used in the German Library system. How can the overlap be maximized to rationalize the duplication? Are crosswalks feasible? There is value within each community for its respective code. How can the use be broadened and the value expanded?
- How can user generated meta-tags be made portable and moved from a current online catalog system such as Primo into future online public catalogs that enable readers to provide feedback?
- With keyword searching of digital content in widespread use, is a controlled list still of more use to the searcher than a keyword? Research is needed to explore this question further.
- Would additional data from ONIX be of use to libraries if it could be used as a filter in the search process?

Best Practice Processes

Ideas to improve processes range from identifying larger issues to suggesting actionable ideas. Since the time required to develop a best practice and achieve widespread adoption typically takes several years, it is useful to assess the pace of industry trends when considering changes that could be overtaken by innovations.

The evolution in the use of book metadata and the development of ONIX to distribute it provides a flow from the inception of the work in the publisher's internal management system to participants in the book supply chain. Increased access to this data for libraries through the use of additional ONIX and MARC crosswalks and the exchange of potentially useful data for both libraries and publishers furthers the dialog from the broader perspective.

Obtaining a MARC record early in the process when the order is placed would enable libraries to have a basic bibliographic record that can be enhanced later if the time and resources are available. Publishers and booksellers are working on best practices for data centers and metadata aggregators to support more rapid updates from the publisher into the booksellers system.

Ideas

- Use crosswalks between ONIX and MARC to facilitate the creation of CIP and to provide publishers with an XML feed of MARC data.
- Work to enhance the CIP record post publication could be shared with OCLC member libraries if database authorizations were expanded.

- Expand training on MARC records with more international vendors to ensure broad conformance to standards and to eliminate duplication of effort by libraries.
- Reviewing of OCLC records could be automated if attributes were exposed so that a manual review could be skipped if it was determined that the contributor is a trusted source.
- Explore the value to publishers of incorporating in their systems the unique data elements added by catalogers (authorized names, subjects, series, and classification numbers).
- For large data files received via FTP as part of an electronic feed, a "manifest" is needed to identify the contents to save staff the time of having to open the file to learn what is in it.
- Descriptive metadata such as the series or table of contents could be synthesized and used to support more refined search results that would also allow better navigation from the collection level, to title level, to the article or image level.
- Science publishers that are exploring tagging content at the chapter level need guidelines and expertise in-house to create and maintain good quality metadata below the product level.
- The long tail of publishers would benefit from Best Practices and simple guidance for options on how to present their data in a consistent manner.
- Economic conditions necessitate the need to simplify ingest mechanisms. A best practice could define quality control expectations.
- There is a need for Collection Identifiers that represent the packages of individual titles that libraries acquire from both publishers and vendors.
- The NISO Thought Leader Meeting on Digital Libraries & Digital Collections recommended the development of a tool that would enable publishers to self test the quality of their metadata.
- Establish best practices for exchange, frequency of updating, feedback mechanisms, and reuse—making the supply chain more multi-directional (not just from publishers to community or from vendor to library).
- Explore methods for integrating the recently published International Standard Text Code (ISTC) and the forthcoming International Standard Name Identifier (ISNI) standards into the existing workflow and promoting their adoption. The ISTC can be used to create associations among works and ISNI could provide authority control for authors.

Outlook

The current economic climate combined with the need to manage rapid growth of content available in multiple formats is driving stakeholders to evaluate how they use metadata and the associated costs. Collaborations that improve results and increase efficiency can streamline the flow of metadata for discovery, sale, and use of content.

One of the challenges for the community is to develop a big-picture view of metadata flow among the stakeholders to recognize how value is added in the supply chain. Following the model of the Symposium, a series of meetings that expand the conversation among all stakeholders will provide a lens for viewing metadata by the broader community and a vehicle to explore how each community can take advantage of what the others have to offer.

In the near term, crosswalks between separate systems will allow sharing of metadata among stakeholders. Long term best practices that enable translation between standards will serve all stakeholders in the industry.

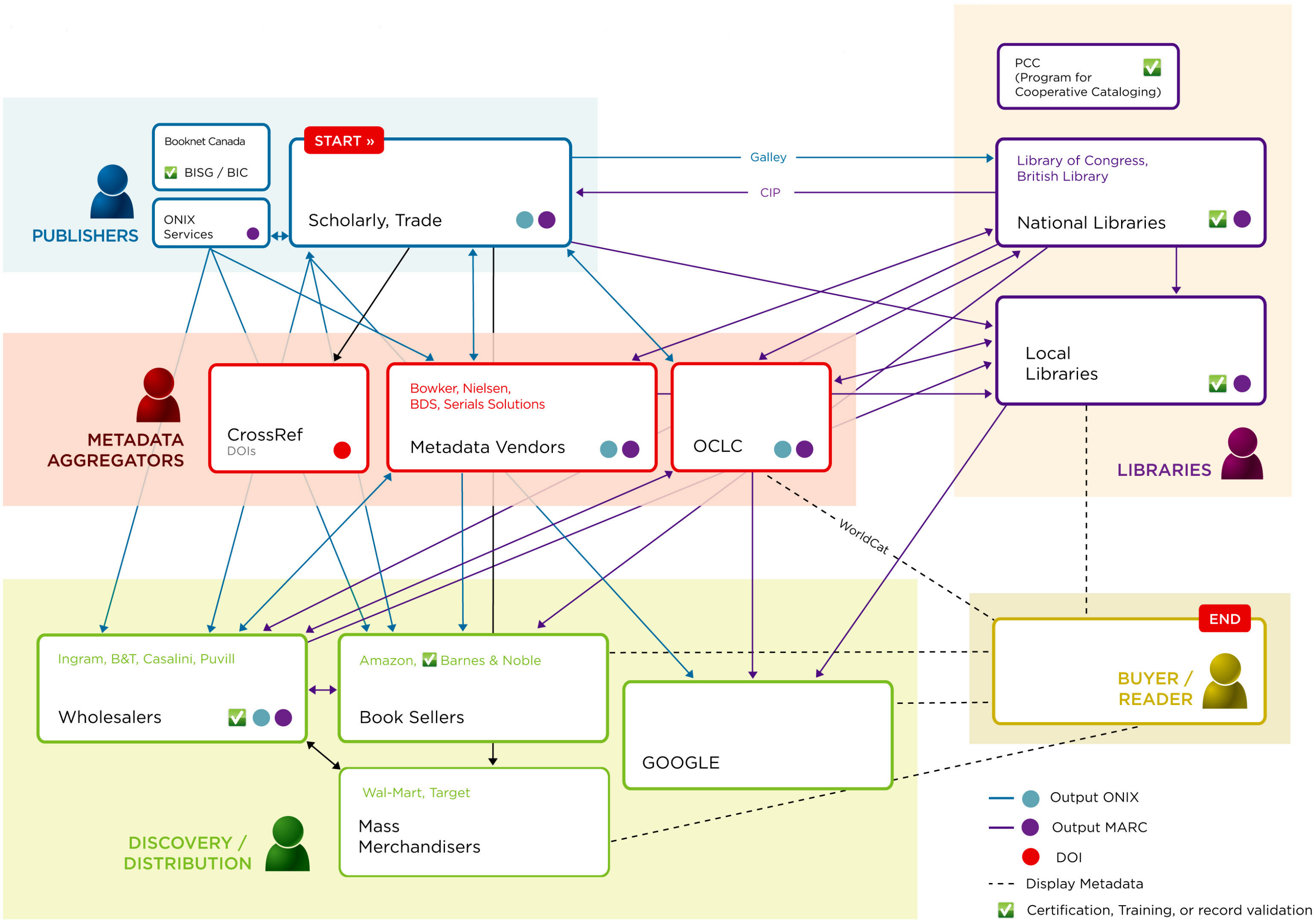

Figure 2: Book Metadata Exchange Map

Glossary of Acronyms with Links

BIBCO – Monographic Bibliographic Record Program
http://www.loc.gov/catdir/pcc/bibco/
See PCC

BIC – Book Industry Communication
http://www.bic.org.uk/
BIC is an independent organization in the UK set up and sponsored by the Publishers Association, Booksellers Association, the Chartered Institute of Library and Information Professionals, and the British Library to promote supply chain efficiency in all sectors of the book world through e-commerce and the application of standard processes and procedures.

BISAC – Book Industry Standards & Communication
http://www.bisg.org/bisac/subjectcodes/index.html
BISAC Subject Codes Committee maintains the Subject Heading and Merchandising Themes lists and provides guidance on the implementation and use of the lists by publishers, retailers, and others. The BISAC Metadata Committee works on ONIX and the Identifiers Committee works on ISBN-13.

BISG – Book Industry Study Group
http://www.bisg.org
BISG is the US book industry's trade association for policy, standards, and research. Membership consists of publishers, manufacturers, suppliers, wholesalers, retailers, librarians, and others engaged in the business of print and electronic media. For over 30 years, BISG has provided a forum for all industry professionals to come together and efficiently address issues and concerns to advance the book community.

Booknet Canada
http://www.booknetcanada.ca/mambo/index.php
Booknet Canada is a not-for-profit agency dedicated to innovation in the Canadian book supply chain. They have their own certification program that uses an audit process.

CIP – Cataloging in Publication by Library of Congress
http://cip.loc.gov/
A Cataloging in Publication record (aka CIP data) is a bibliographic record prepared by the Library of Congress for a book that has not yet been published. When the book is published, the publisher includes the CIP data on the copyright page thereby facilitating book processing for libraries and book dealers.

CDS – Cataloging Distribution Service
http://www.loc.gov/cds/index.html
CDS offers bibliographic products and services to libraries and book vendors in the form of MARC records and the tools used by catalogers.

CISAC - International Confederation of Societies of Authors and Composers
http://www.cisac.org
Founded in 1926, CISAC works towards increased recognition and protection of creators' rights by working with the international network of copyright societies.

CONSER – Cooperative Online Serials
http://www.loc.gov/acq/conser/
See PCC

CrossRef
http://www.crossref.org/
CrossRef is a member association of publishers and serves as the official DOI® link registration agency for scholarly and professional publications. It operates a cross-publisher citation linking system that allows a researcher to click on a reference citation on one publisher's platform and link directly to the cited content on another publisher's platform, subject to the target publisher's access control practices.

DDC – Dewey Decimal Classification
http://www.oclc.org/dewey/
Devised by Melvil Dewey in the 1870s and owned by OCLC since 1988, the DDC provides a dynamic structure for the organization of library collections and is the world's most widely used library classification system.

DCMI – Dublin Core Metadata Initiative
http://dublincore.org/dcmirdataskgroup/
The development and maintenance of a core set of metadata terms is part of the origins and history of DCMI which provides simple standards to facilitate the finding, sharing, and management of information.

EAN – European Article Number by GS1
http://www.gs1.org/ecom/overview.html
The EAN links to the full ONIX record and appears as a bar code.

EDI – Electronic Data Interchange
http://www.x12.org/
EDI is the transfer of structured data, by agreed message standards, from one computer system to another without human intervention. It is the data format used by the vast majority of electronic commerce transactions in the world. In the US, EDI standards are developed by The Accredited Standards Committee (ASC) X12, an accredited ANSI standards developer.

EDItEUR
http://www.editeur.org/
EDItEUR has 90 members in 17 countries participating in the development of the standards infrastructure for electronic commerce in the book and serials industries. Much of EDItEUR focus is on EDI, bibliographic information and related standards, digital publishing, RFID tags, and rights management. EDItEUR provides management services for the International ISBN Agency.

FRBR – Functional Requirements of Bibliographic Records by IFLA
http://www.ifla.org/VII/s13/frbr/
FRBR is a conceptual entity-relationship model developed by the International Federation of Library Associations and Institutions (IFLA) that defines user tasks: find, identify, select, and obtain any of the entities or relationships.

GDSN – Global Data Synchronization Network
http://www.gs1.org/productssolutions/gdsn
The GDSN is built around the GS1 Global Registry®, GDSN-certified data pools, the GS1 Data Quality Framework, and GS1 Global Product Classification, which when combined provide a powerful environment for secure and continuous synchronization of accurate data.

GS1 – The global language of business (tagline)
http://www.gs1.org/
GS1 is a global organization dedicated to the design and implementation of standards and solutions to improve efficiency and visibility in supply and demand chains across sectors around the world. The GS1 system has four product areas: barcodes, eCom, GDSN and RFID. They have active programs in Healthcare, Defense, and Transport & Logistics.

IFFRO – International Federation of Reproduction Rights Organizations
http://www.ifrro.org
IFRRO works to increase the lawful use of text- and image-based copyrighted works and to eliminate unauthorized copying by promoting efficient collective management of rights through RROs to complement creators' and publishers' own activities.

IFLA – International Federation of Library Associations and Institutions
http://www.ifla.org/
Founded in 1927, IFLA is an organization of societies and individuals that offer a global forum for the library and information profession.

ISBN – International Standard Book Number by ISO
http://www.isbn-international.org
The ISBN (ISO 2108) is a unique international identification system for each product form or edition of a monographic publication published or produced by a specific publisher. ISO 2108 specifies the construction of an ISBN, the rules for its assignment and use, the metadata to be associated with the ISBN allocation, and the administration of the ISBN system.

ISNI – International Standard Name Identifier by ISO
http://www.isni.org/
The proposed ISNI standard (ISO/DIS 27729) is a method for uniquely identifying the public identities of contributors to media content such as books, TV programs, and newspaper articles. It will provide a tool for disambiguating names that might otherwise be confused, and will link the data about names that is collected and used in all sectors of the media industries.

ISTC – International Standard Text Code by ISO
http://www.istc-international.org/
The ISTC standard (ISO 21047) defines a global identification system for textual works and is primarily intended for use by publishers, bibliographic services, retailers, libraries, and rights management agencies. Each ISTC is a unique "number" assigned by a central registration system to a textual work and is used to identify the same content even when it is being published by a different publisher or format.

LCSH – Library of Congress Subject Headings
http://www.loc.gov/cds/lcsh.html
Known as the "red books" the LCSH, with 300,000 terms, provides the most comprehensive list of subject headings; they are used in libraries throughout the world.

Library of Congress Working Group on the Future of Bibliographic Control
http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf
This controversial and forward looking report recommends collaborated efforts for more efficient production of bibliographic records.

MARC – MAchine Readable Cataloging
http://en.wikipedia.org/wiki/MARC_standards or http://www.loc.gov/marc/umb/
MARC records were developed in the 1960s through a Library of Congress initiative to support the exchange of standard bibliographic data consisting of: 1) description of the work, 2) author's name

verified, 3) standards based subject headings, and 4) classification number (Dewey and/or Library of Congress). A MARC record involves three elements: the record structure, the content designation, and the data content of the record. In 1997, the US and Canadian versions of MARC were harmonized into "MARC 21," the current version.

MeSH – Medical Subject Headings
http://www.nlm.nih.gov/mesh/
MeSH is the National Library of Medicine's controlled vocabulary thesaurus; descriptors are arranged in both an alphabetic and a hierarchical structure.

METS – Metadata Encoding and Transmission Standard
http://www.loc.gov/standards/mets/
The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library and expressing it with the XML schema language of the World Wide Web Consortium.

MODS – Metadata Object Description Schema
http://www.loc.gov/standards/mods/
Metadata Object Description Schema (MODS) is a schema for a bibliographic element set that may be used for a variety of purposes, but were particularly designed for library applications.

NACO – Name Authority Cooperative Program
http://www.loc.gov/catdir/pcc/naco/naco.html
See PCC

NISO – National Information Standards Organization
http://www.niso.org
Bringing together publishers, libraries, and systems providers, NISO develops voluntary consensus standards impacting all areas of the information supply chain, from paper permanence and performance measure and search functionality to identifiers, metadata formats, and preservation. NISO is appointed by ANSI as the US Technical Advisory Group (TAG) administrator for ISO Technical Committee 46 on Information and Documentation. NISO also serves as the secretariat for ISO Technical Committee 46, subcommittee 9 on Identification and Description.

OCLC – Online Computer Library Center
www.oclc.org
A member organization comprised of 71,000 libraries in 112 countries, OCLC is dedicated to furthering access to the world's information and reducing the rate of rise of library costs to locate, acquire, catalog, lend, and preserve library materials. OCLC's WorldCat database provides online access to the bibliographic records of the world's largest network of library content and services.

ONIX for Books – Online Information Exchange for Books by EDItEUR
http://www.bisg.org/documents/onix.html
ONIX is a standard format that publishers can use to distribute electronic information about their books to wholesale, e-tail and retail booksellers, other publishers, and anyone else involved in the sale of books. (Note: The ONIX name is also used for other non-book standards such as ONIX for Serials which includes Subscription Products. When used in this report, ONIX refers to ONIX for Books.)

PCC – Program for Cooperative Cataloging (coordinated by Library of Congress & PCC)
http://www.loc.gov/catdir/pcc/
Initiated in 1992, PCC is an international cooperative effort aimed at expanding access to library collections by providing useful, timely, and cost-effective cataloging that meets mutually-accepted standards of libraries around the world. Programs of the PCC are:

- NACO - the name authority program
- SACO - the subject authority program
- BIBCO - the monographic bibliographic record program
- CONSER - the cooperative online serials program

PCN – Preassigned Control Number Program by Library of Congress
http://pcn.loc.gov/
Library of Congress assigns a unique identification number to catalog cards so that they can be ordered. These are basic catalog records that may be created before the book is published. The number is referred to as the Library of Congress Card Number.

PDCP – Product Data Certification Program by BISG
http://www.bisg.org/documents/certification_productdata.html
PDCP is a voluntary program that enables publishers (and certain other organizations) to submit files of product information to the Book Industry Study Group and to have those files evaluated objectively and against clearly defined criteria.

RDA – Resource Description and Access by the Joint Steering Committee for Development of RDA
http://www.rda-jsc.org/
Built on the foundations established by AACR2, RDA is intended to provide a comprehensive set of guidelines and instructions on resource description and access covering all types of content and media. Still in development, the final version is scheduled to be released at the end of November 2009.

RIN – **Research Information Network**
http://www.rin.ac.uk/
RIN was set up by the four Higher Education funding bodies in the UK, the three National Libraries, and the seven Research Councils. It focuses on understanding and promoting the information needs of researchers.

SACO – Subject Authority Cooperative Program
http://www.loc.gov/catdir/pcc/saco/
See PCC

UKSG – United Kingdom Serials Group
http://www.uksg.org
Member organization comprised of libraries, publishers, intermediaries, and technology vendors focused on the exchange of ideas on print and electronic publishing and scholarly communication.

VIAF – Virtual International Authority Files
http://www.oclc.org/research/projects/viaf/
This joint project began initially with the Library of Congress, the Deutsche Nationalbibliothek, and the Bibliothèque Nationale de France and has since expanded to include the national libraries of other countries. VIAF explores the process of virtually combining the name authority files of all three institutions into a single name authority service that reflects the authors' name in the local language. The prototype was developed at OCLC and is available at http://viaf.org/

XML – Extensible Markup Language by W3C
http://www.w3.org/XML/
XML is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere.