Reflections on Metadata

Recognizing Change Is Inevitable

What is the correct response when the ground is shifting beneath your feet? That is where many publishers are finding themselves—in an era of tectonic shifts within the industry, when print sales in most sectors are declining over time, but the promise of new e-product opportunities has yet to be realized. The correct response, I have argued, is to be nimble, to be ready to respond quickly and with flexibility to whatever the market demands. Key to this strategy is the protection and preparation of content.

Market changes—especially where technology is involved—can often be sudden and seismic in nature. We saw it with Napster and the music industry. And it wasn't long after the introduction of Netflix that Blockbuster was rendered lackluster. In the case of publishing, developing technology has had an impact, but the change has been more gradual. The tremors have been felt for many years and sometimes even ignored by publishers. But, make no mistake, the foundations underlying the industry have been drifting irrevocably in new directions.

Because print publishing has not undergone the colossal make-over of its sister media does not mean that technology changes in our industry are about to be reversed. What's not clear is the full extent to which digital will change our publishing landscape.

Most sectors are seeing declining revenue from print, with digital replacing only a portion of those losses. (In recent times, trade publishing has experienced a modest uptick in print revenues, but if we take the long view over the past six years, print unit sales are markedly down in this sector as well.) Should we expect digital to rise as print falls? To a certain extent, yes, but demographics also play a role. We are not the same reading populace we were 20 years ago. So, is it realistic to expect digital publishing revenues to replace print revenues entirely? Most certainly not! But as I have argued in a previous paper, we need to be poised to take advantage of whatever market presents itself.

We've seen how fast new technologies amass new audiences—witness FaceBook, Twitter, and Instagram. Do these technologies represent new markets for publishing? They have certainly changed the way we market our products. And that change has come about rapidly. Making sure that content is in a ready state to be promoted, or even published to new digital audiences is critical to a publisher's success in today's economy. Old can become new in this market where the ability to promote and publicize content (or make it discoverable) has never been more potent.

"Archival" content was not always given high importance. A finished goods repository was once the equivalent of a dusty file cabinet or even a waste bin. With the explosion of digital product, however, the status of that once "archival" content has radically transformed to become a publisher's most valued asset. What was once a tome from the past is now a present and future source of revenue. Publishers who treat this asset casually do so at their own peril. Content now needs to be consistently structured, easily accessible, and actively managed.

Building on that prescription, this paper provides an executive overview of metadata and its importance to the viability of any commercial publishing endeavor. We will examine the increasing complexity of metadata, its exponential growth, and the critical role it plays in expanding the publisher's audience and reach. There are a number of common metadata standards, each with a particular role, serving different functions, markets, and users. Ultimately, like the content it describes or serves, metadata is far too valuable not to be consolidated into a proper data management system for delivery to present and future customers.

Note that the intended audience here is not the metadata technician but the publishing business executive. The topic of metadata may often come across as technological minutia, but the consequences of getting it wrong are anything but that. The purpose here is not to address the fine detail of a specific metadata function but to provide an overview and suggest an overall strategy for dealing with it and avoiding common pitfalls. Descriptions are kept purposefully basic.

Let's start with a basic description: What is metadata? It is essentially information about the product or product content. In a somewhat reflexive definition, it is data about data. It describes the product and governs transactions associated with that product. For the publisher, it enables the increasingly complex business rules for what, when, where, and how content is used.

The Exponential Growth of Metadata

"No Room! No Room!" --The Mad Hatter, et al.

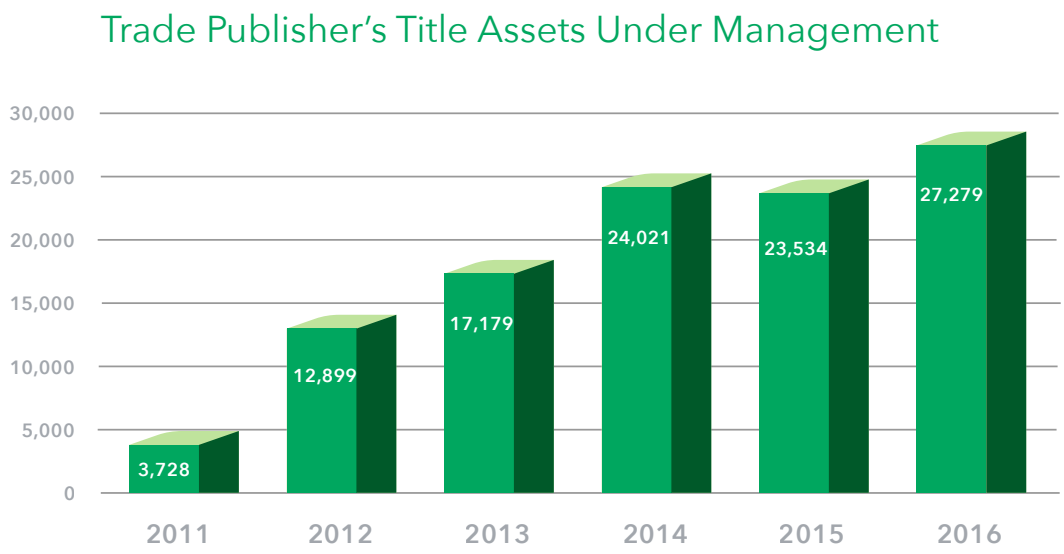
In the past, metadata occupied a humbler and simpler role. With print-only product, and prior to the emergence of large online retailers, metadata was primarily a tool for tracking a single version of each title, providing information for print sales, and basic cataloging. It was often stored (and sometimes still is) in the publisher's title tracking or business system.

With the advent of e-products, online retailers, and the explosive growth of web business, metadata has grown exponentially and has increased both in complexity and importance. A single title that originally carried a single print record suddenly became available in multiple formats, each with its own associated metadata. The number of metadata fields requiring management was now begging for computer assistance.

Typically, there are 30-40 fields of metadata required for every digital product in active distribution. Even for smaller publishes with inventories of less than 2,000 titles, the exposure can run 50,000 to 100,000 data points. See Charts 1 and 2 for one example from a trade publisher.

Chart 1

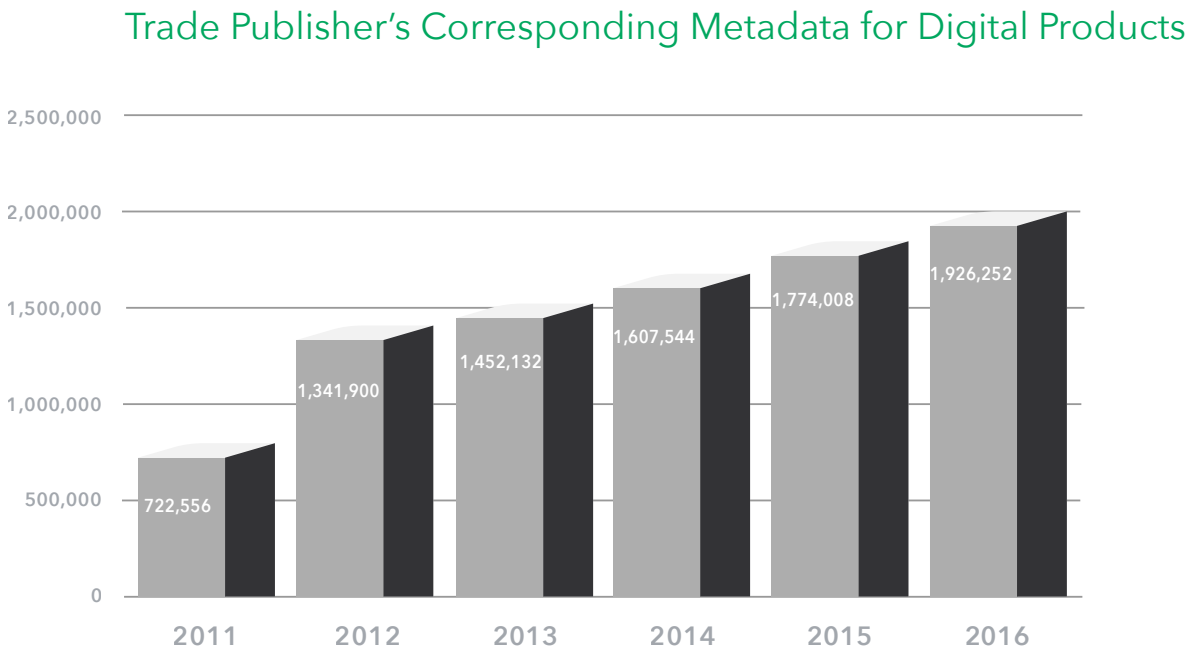
Trade Publisher's Volume of Title Assets (Finished Goods) Under Management



Represented here are titles under management stored by a relatively large trade house over time. This particular publisher organized front list and backlist title assets (finished digital produts as well as works in progress) with the help of a third-party product data repository (PDR).

Chart 2

Trade Publisher's Metadata Under Management in Support of Digital Products



Here the volume of metadata stored in the PDR is associated with the titles assets in Chart 1. Note that this is metadata in support of the publisher's digital products only.

In addition, online markets began demanding a variety of descriptive and transactional metadata. It is metadata that then controls and enforces the rights, permissions, and distribution rules set by the publisher. Furthermore, with the advent of search engines, metadata became the key to discoverability. Precision in writing descriptive metadata became critical.

Publishers saw the amount of metadata doubling, tripling, etc., with each new format added. Many of the title tracking/business systems used by publishers were originally built assuming a print-only architecture. Such systems could simply not accommodate the crush of additional data.

In addition, not all title tracking systems were designed to ensure a consistent format for each metadata element. Multiple users with input rights tend to express the same metadata element in different ways. Consider what can happen with a free text field that allows the same information to be expressed according to the whim of each user within the business. For example, the same trim size might be input as 8-1/2 X 11 in one instance, 8.5" by 11" in another, etc. At that point, a publisher's metadata has devolved into mere text. The inconsistencies come back to bite as soon as the publisher attempts to sort by trim field. Most importantly, such variations of the same information undermine the metadata's machine readability, which requires precision and consistency in its creation.

The Ins and Outs of Metadata

The God Janus Is Pictured Facing in Opposite Directions

The role of metadata has expanded in ways not previously imagined. As noted above, a publisher's most critical quality of discoverability is dependent on how that publisher's key word metadata is captured and prioritized by various search engines. The library market is increasingly demanding more about how content is cataloged and how that information is made machine readable (MARC records). Pricing and transactional metadata is variable across multiple channels and territories, and it needs to be kept up to date in as close to real time as possible. Metadata defines content at the chapter and article level, making it possible to sell more granular information with more specificity.

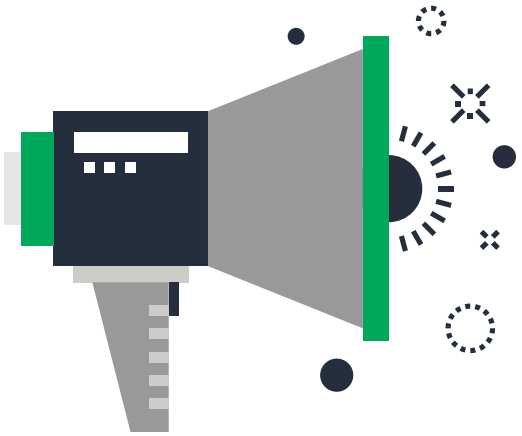
Access to data has never been greater. Some would characterize it as overwhelming. We are creating and adding to online inventories of content at a rate of 5 exabytes a day. How does a book or journal—much less a chapter or article—get discovered with so much diversion and direct competition? It is metadata that provides the best shot—metadata in many forms and aligned to a broad universe of interpretive engines. Broadly speaking, metadata comes in two categories: metadata external to the content object and metadata internal to the content object.

METADATA EXTERNAL TO THE CONTENT OBJECT WOULD INCLUDE:

- All descriptive properties, such as the product registry ID number (ISBN, ISSN, EAN), the author, editor, contributors, publisher, extent, trim size, format, abstracts. This also includes overall product descriptions and product assembly instructions for the warehouse.
- Associated properties such as author biographies and product reviews.
- Discoverability elements such as key words and BISAC subject codes.
- Commercial/transactional metadata, such as price, currency, territory rights, digital rights.
- Cataloging metadata, including MARC records and KBart.

METADATA INTERNAL TO THE CONTENT OBJECT WOULD INCLUDE,

- Digital object identifiers (DOI's) and DOI registries, such as CrossRef and DataCite. [It is the use of DOI's and DOI registries that make it possible for the publisher's online system to capture individual book chapters for sale.]
- Industry ontologies with embedded definitions and references. [Wherever there are well-defined communities--societies, trade associations, government regulatory entities—there has been the drive to devise an ontology of terms to describe metadata. This has resulted in a science/discipline of semantic language searches and content retrieval that takes metadata to a whole new level.]



The Emergence of ONIX

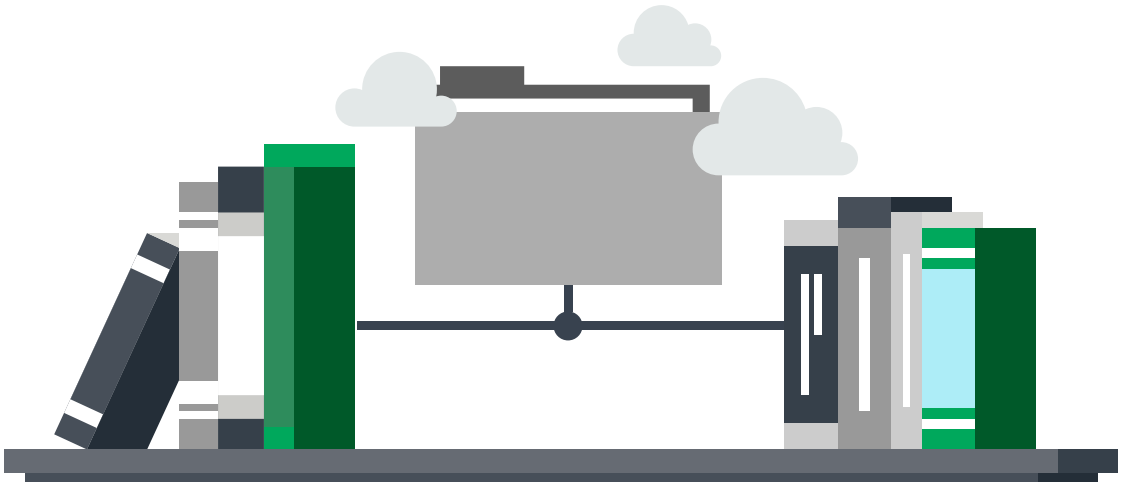
Communication Is Key

The exponential growth of metadata, as well as the increasing importance to keep it accurate and up to date, made obvious the need for a standard of communication. Such a standard was created in 2000 by industry groups EDitEUR and the Association of American Publishers (AAP), aided by the Book Industry Study Group (BISG) in the US and Book Industry Communication (BIC) in the UK.

ONIX, standing for online information exchange, is metadata as an XML instance conforming to a discrete set of international standards. ONIX computer-to-computer communication has become the most practical means of dealing with a plethora of metadata that now subsumes the market for content.

ONIX for Books made possible a consistent means for publishers to provide rich metadata to online retailers, supply chain partners, data aggregators, and other interested parties in the publishing business. With ONIX, publishers were finally able to provide metadata revisions without the cumbersome practice of developing EXCEL spreadsheets and waiting for each partner to get around to uploading the data. Updating the original ONIX versions required wholesale replacement of the entire record. 3.0 (the latest) version allows for more granular updates.

ONIX is an important tool for the publisher's ability to respond quickly, consistently, and with flexibility to market demands. Price changes and new product offerings can be made closer to real time than ever before. With ONIX comes the ever-pressing demand for a dedicated repository and managed platform for metadata. It is no longer practical to manage metadata in separate silos under the various business units and departments that create it. (See discussion below.)



Effective Capture and Management

Where's Waldo?

Whatever system a publisher adopts, the automatic processing of metadata requires having a consistently structured format for each element; in other words, a standard. There are a variety of standards in use, each having relevance for the task being performed (e.g., ONIX as discussed) or a particular market being served (e.g., MARC records for librarians).

With all the standards and wide array of uses, publishers can no longer afford to have metadata records in various "pockets," silos separated by different departments or different facets of the publishing business. It is easy to understand how publishers get into this situation. Aggregating and maintaining consistent metadata companywide is difficult when employees managing metadata are operating in different systems with different business rules. It is time (past time!) to get beyond disparate systems that don't talk to each other or (heaven forfend!) paper records. The digital age has brought metadata into the foreground and has demanded a better means of managing and leveraging it.

There is a growing need to provide new types of referencing, new types of cataloging. The Internet has brought about the emergence of outside information directories, indices, and exchanges that are driving the demand for more metadata and more refined metadata.

In the ideal situation, the publisher will have one system that houses all metadata under a single roof. Publishers still cobbling together metadata from multiple sources are already handicapped by a lack of flexibility and timely response to market opportunities. This situation will become even more dire as time goes by.

Within the ideal system, under that single roof, there still needs to be a systematic way to configure metadata to meet the exact demands or particular business needs of the publisher and business partners. For example, there is good reason to separate commercial metadata, where it can be kept current and routinely distributed to aggregators. Access to such metadata—and certainly the ability to alter it—should be strictly administered.

Information Exchange

Using the Correct Currency

Metadata is the currency of information exchange. In order to have information discovered and presented to the relevant audience, we need the right metadata in the right configuration. It is critical that publishers be able to quickly and accurately distribute and exchange metadata with a number of outside interests. As a result, they need a system that is able to retrieve metadata and structure it to the particular standard demanded.

A responsive system is called for, one that includes master templates for each market segment or interest group. All metadata keyed in or uploaded will need to be validated against appropriate templates. If it is metadata associated with a title to be sold on Amazon, it should be validated against a template configured with the latest Amazon requirements for markets and market influencers. These requirements can be frustratingly precise and eclectic. More than one publisher has complained about Apple's pesky pricing grid for iBooks.

Consider the financial implications of a publisher hobbled by an inefficient means of distributing price changes. The publisher determines to make a strategic increase (or decrease) in prices across an entire suite of products but can only send out a series of spreadsheets to aggregators. The time delay in creating those spreadsheets is one thing, but then how long will it take each aggregator to actually ingest those changes? That situation is hardly ideal for having a nimble and flexible response to a rapidly changing market.

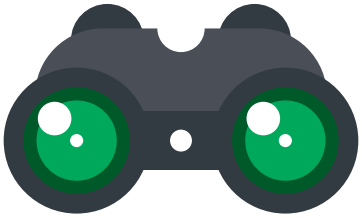
When Metadata Goes Bad

The High Price for the Imprecise

The price change scenario noted above is just one example of the numerous ways the mismanagement of metadata can lead to some very concrete painful results. It might help to take a brief look at some real life issues that occur in day-to-day business. Consider the following:

- A coffee table book extolling the many delectable variations of sipping bourbon is not selling. Turns out the subject code designating the title as having to do with alcoholic beverages had an unforeseen qualifier: it placed the title in a category of books dedicated to the prevention of alcoholism.
- A percentage of a publisher's ISBN's are completely missing product type and ultimately do not show up on pre-defined searches
- A major institution or channel partner has changed spec requirements. The publisher is not prepared to respond rapidly. The slow adjustment translates into lost sales.
- A percentage of customer orders have first pass failure at time of order entry, due to incomplete metadata. Delays result in lost revenue and sometimes lost customers.
- Multiple component orders are assembled incorrectly due to incorrect metadata, resulting in returns and unhappy customers.
- There is a mix-up in publication dates due to US style of show month/day versus UK style of showing day/month. A publication date listed as 12/1 is actually 1/12.
- Incorrect distribution metadata results in restricted books going to the wrong aggregators or unrestricted books being missed.

The above is just a short list to provide a sampling. The point is clear: when metadata is not properly managed, the publisher is at risk of losing time, money, and customers.



Seeking Outside Assistance

It Is Over-Whelming. Our Minds Are Boggled

The notion of consolidated metadata storage and efficient exchange can become overwhelming. Looking back to the initial discussion of the exponential growth of metadata, one can better appreciate the issue. Consider a simple math exercise. Imagine being steward to just 1,000 titles, each with 140 metadata fields. That alone means 140,000 fields to track. Now if one simply needs to manage 10% of that metadata universe in a single year, it would mean dealing with 14,000 data points. That works out to handling on average 54 metadata adjustments every working day, with no holidays.

There is no harm in a publisher seeking outside help. To quote Thad McIlroy from his article "Everything You Thought You Knew About Metadata…But Were Afraid to Ask,"

"If you feel that your inner pinball machine has just tilted and froze, it's time to look for a metadata partner. Potential partners offer services from basic to holding your hand, and their prices are thoroughly reasonable. You just need to find the right one for your scale and mission."

Often, a publisher is too close to its own issues or too removed from an understanding of the best technologies available. An outside perspective can help. In the case of metadata and metadata/content systems, there are a number of good potential partners.

Summary and Conclusions

As technology, the market, and the publishing business have evolved, it has become imperative for the publishing executive to recognize and respond to the new roles of content and metadata. Content has transformed from "archival" to "most valued asset." The associated metadata has not only grown exponentially and become increasingly complex, it is now an essential dynamic tool enabling what, where, when, and how content is used.

Metadata pervades all aspects of the publishing business. The challenge is not simply a matter of getting metadata right – it must actually be managed.

The first step for the publishing executive is to recognize this fact: metadata is far too important to slip under the radar. How content and metadata are managed will depend on individual publisher needs and budget.

If uncertain, an executive should get professional advice. It is always better to anticipate the pitfalls before stepping into one.

Brought to you by codeMantra