

The Supply and Demand of Metadata Competency in the Publishing Industry

A White Paper

January 14, 2013

Pace University M.S. in Publishing Program
Michael McGill
mike@intelpreneur.com

The Supply and Demand of Metadata Competency in the Publishing Industry

Table of Contents

Introduction	1
The Status of Metadata Competency & Its Impact by Industry	2
Book Publishing	2
Information Technology	3
Science, Technical & Medical	3
Online	5
The Ideal Metadata Candidate for a Publishing Company	9
Industry Recruitment Shortcomings	10
Conclusion and Recommendations	12
Appendix A	
Industry Workforce Comparisons	13
Relevant Course Descriptions from Publishing MS	14
Relevant Course Descriptions from Library and Information Science MS	15
Relevant Course Descriptions from Information Systems MS	16
Appendix B	
Trade Book Publishing Financial Analysis	17
Sources	17
Deducing the Values of the Online Sales Charts	17
Deducing Enhanced Sales Estimates	19
Projecting Trade Online Sales Growth - Metadata ROI	19
Trade Publishing: Total Market & Non-Fiction Market	21
Sales vs. Percent Backlist Matrix to Estimate Additional Operating Margin	22
Concluding Remarks on the Financial Analysis	24

The Supply & Demand of Metadata Competency in the Publishing Industry

Introduction

The objective of this study is to explore the publishing industry's metadata requirements from an employee proficiency standpoint and to determine whether the workforce has sufficient talent to meet those needs.

Unlike the stock market and baseball there is limited publicly available data on metadata (meta-metadata) with which to conduct a definitive analysis on the impact that complete and accurate metadata (or lack thereof) has on sales and operational effectiveness within the publishing industry. The scope of the publishing industry, which ranges from trade books to online websites to scientific, technical, and medical publications (STM), further complicates the issue because each market has its own unique relationship with metadata. I have pieced together some qualitative data from metadata professionals both within and outside the publishing industry to get a descriptive sense of their pain-points and recommendations for improvement. Insights from these interviews were then combined with data mined from LinkedIn profiles to get a broad sense of the educational backgrounds, skill sets, and work experience from metadata employees in the Publishing, Libraries, and Information Technology industries. In addition to this I've conducted an analysis of the trade book industry from a financial perspective to calculate a rough estimate of how much more revenue that particular industry stands to gain from a more proficient metadata workforce. The Trade Book Publishing Financial Analysis section will go into further detail on these drawbacks.

[Return to TOC](#)

The Status of Metadata Competency & Its Impact by Industry

I was able to conduct interviews or receive responses from five individuals who deal with metadata on a regular basis, both inside and outside of the publishing industry.

Thad McIlroy	Co-Author of <i>The Metadata Handbook: A Book Publisher's Guide to Creating and Distributing Metadata for Print and Ebooks</i>
Adrienne Tannenbaum	Author of <i>Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand and Implementing a Corporate Repository: The Models Meet Reality</i>
Melissa Fulkerson	Senior Channel Manager at Elsevier
Aaron Deutsch	Chief Technology Officer at Psychology Today
Evan Whalen	Tech Lead at chloe + isabel, inc.

Book Publishing

Thad McIlroy knows a great deal about metadata in relation to the book industry and his feedback suggests that while there is a need for a more highly trained workforce, the problem is much larger and compounded by additional issues. Metadata is a complex issue and simply raising the training standards of employees won't solve the problem. He agreed with my conclusion, which I develop in a later section, that the average publishing professional has not been adequately trained in data management and that those who are (Information Systems & Computer Science Majors) are not likely to enter the publishing industry. An additional problem is that the executives at the major publishing companies also fall under the category of not having a strong technology background and as a result these decision-makers haven't been emphasizing the technology skills required to adequately transform and adapt their companies. On the bright side, as Thad pointed out, the industry does not need the cream of the crop from Silicon Valley to achieve its objectives - all it needs is a reasonably trained staff with an interest in publishing. This group does exist and a segment of this talent supply can be found at the [CodeToPrint](#) meetup group. For

more information on the many issues and problems of metadata in the book publishing industry I recommend Thad's website [The Future of Publishing](#).

Information Technology

Adrienne Tannenbaum can also be considered a metadata guru, having authored two books on the subject for the Information Technology field. While Thad McIlroy's tone expressed frustration at the current state of metadata affairs, Adrienne's vexation over the issue reached a feverish pitch. She first broke metadata into two distinct categories: 1. Libraries, Periodicals and Publications; and 2. Information Technology. In Adrienne's opinion libraries have been very successful in their use of metadata specifically because of the content they use. On the other hand, IT is a mess because they don't have any standardized content. Because of this, the focus has recently been on data governance which aims to standardize data definitions, data names, data classifications, etc. This headache of affairs has caused Adrienne to do some 'soul searching' and her new aim is to move metadata out of the Information Technology industry and into industries, such as publishing, that have a business need for search and retrieve classifications. Her selling point as of late has been to encourage information-heavy organizations to think about how hard it is for them to actually find their data. In Adrienne's opinion, data governance has failed at this and continues to be on the losing end of the cost/benefit wars. Her advice: bring the technical skills normally found in IT and move them into publishing while at the same time transition metadata knowledge into taxonomy focused classification awareness.

Science, Technical & Medical

Melissa Fulkerson points out that the science, technical & medical (STM) publishing industry faces the same challenge of finding adequately skilled talent to meet their metadata needs.

More so than the rest of the publishing industry, the discoverability of STM publications is highly dependent upon the accuracy of its metadata. This is due to the highly specialized nature of the content (i.e., Global depletion of dopamine using intracerebroventricular 6-hydroxydopamine...) which in turn requires that the metadata be extremely articulate and granular. Melissa found that it's been extremely difficult to attract candidates who understand 1. What metadata actually is at its core, 2. How and why it's used and how integral it is to the electronic sales and marketing process, and 3. What the technical requirements are by all parties (publishers, authors, resellers, etc.). The company is currently in a learning/adaptation process where they are training their employees to better understand metadata and they've also recently created and filled a new position, "Technical Marketing Manager," which focuses almost exclusively on metadata and online content discoverability - she states that this has already proven valuable and expects it to continue to do so.

Going forward Melissa hopes to find more statistical information on the effectiveness of metadata (she referenced the Nielsen report as being a great resource, but as I point out later in this paper the Nielsen data is somewhat misleading). She also would like to get a better understanding of how resellers such as Amazon, Google and Barnes & Noble use the metadata she provides. Unfortunately, according to Thad McIlroy "Amazon, B&N, et al. pay lip service to metadata, but only support a small subset of the whole, and their support only works when there are no irregularities in the metadata instance." His blog entries and metadata book site cover this topic in detail and he also suggests reading the [Yahoo ONIX](#) group to "find the real heart of darkness documented."

Online

Aaron Deutsch, the Chief Technology Officer of Psychology Today magazine and the owner of Haus Interactive, was able to provide some insight into how metadata affects online discoverability and the difficulties faced when dealing with a consumer-base responsible for tagging their own profiles. According to Aaron, metadata is not as important as one would think in the online world. He suggests that the top 1,000 websites can gain significant traffic by investing time and effort into metadata, however small sites are likely to see gains so small that it's not worth the effort. Google, he says, does a very good job of automatically extracting content from a website and deciding where it should land on the search results. This was actually a disappointing lesson Aaron learned when he and his team invested a great deal of time into metadata. They eagerly awaited the results from Google only to find that it is predominantly the quantity of content that determines search rankings, not the quality of that content. Fortunately, the scales should be tipped in favor of quality content going forward as Google has recently revised its search engine algorithm in its latest Panda upgrade .

There are however, instances where it's worthwhile to invest the time and effort into metadata. He specifically points to [Psychology Today's online Therapist Directory](#) which offers searchable online profiles of roughly 50,000 therapists and health professionals. For Aaron and his team, this was an intensive, multi-year effort that has resulted in the Therapist Directory consistently taking the top 3 results in a Google search ([Click for Example](#)). Currently, they have a dedicated employee, Charles Frank, who has so far done an excellent job directing the online directory's SEO and SEM. Charles has a finance background which is worth mentioning because the general consensus has been that anyone with a data background can easily transition into a metadata job.

[Return to TOC](#)

Like all the other people I interviewed, Aaron also finds aspects of the metadata process frustrating, such as relying on his customer base to tag their own profiles. Each therapist in the online directory is responsible for populating their profile. Aaron has done everything he can to make sure they get it right the first time, including making the system as rigid as possible to avoid mistakes and actively encouraging each therapist to dedicate more time to their profile. Despite this, he and his staff spend countless hours editing their profiles to fix THEIR NAME'S WRITTEN IN CAPS or to fix obvious mistakes such as when a therapist lists their phone number in the address box. Aaron believes there are two reasons for this problem and that they will never go away. First, he estimates that only 0.2% of the population is detail oriented*. While PhD's in Psychotherapy may not fall under that category, they do fall under the second category, which is that the vast majority of people, regardless of their educational background, have no business skills. So, to draw a parallel to the book publishing industry, a brilliant best-selling author will have a difficult time filling out a billing statement or documenting appointments. This same skill set carries over into metadata competency. Unless at some point in their learning path they received a data-related education, they otherwise don't have the skills to accurately tag and classify their publications. Aaron's conclusion on the topic is that problems such as this will only be fixed by improvements in technology.

I spoke with Evan Whalen, the tech lead at chloe + isabel, inc. who deals heavily with metadata, about his recommendations for a 'bare bones' education in metadata. He found this to be an interesting question because he's never heard anyone describe themselves as a 'metadata guy' - it's always been something a person does in unison with other tasks. Unlike SEO, which

* The percentage of the population that is detail-oriented may be larger than Aaron's estimate. According to the [Myers-Briggs personality test](#) the Introversion-Sensing-Thinking-Judging (ISTJ) type is detail-oriented and makes up about 10-14% of the population. Defining characteristics are: organized, orderly, tidy, strong sense of duty, hardworking, detail-oriented, dependable and has a preference for security and tradition over the unknown and theoretical.

he suggests you can literally just read a book about to understand, metadata requires a certain data-oriented background and a way of looking at things. In Evan's opinion it doesn't take much to reach an adequate level of proficiency, though he admits some bias as he's dealt with metadata and SEO for much of his life.

Opposed to Evan's opinion that "once you know it, you know it" and that a mere XML class will suffice, Thad has developed a more thorough scale of proficiency. To meet minimum requirements Thad proposes a two-day training course, with Day 1 focusing on an introduction to metadata principles, followed by an introduction to ONIX on Day 2. Beyond this basic introduction Thad defines metadata as "any information that can be encoded in alphanumeric text that makes a work discoverable" and therefore sees the metadata scale of proficiency as follows:

1. Basic metadata expressed via Excel or a Title Management System
2. Basic metadata expressed as ONIX (the 2.1 spec)
3. Enhanced metadata expressed as ONIX (the 2.1 spec)
4. Enhanced metadata expressed as ONIX (the 3.01 spec)
5. Full metadata expressed as ONIX (the complete 3.01 spec)
6. Basic SEO
7. Intermediate SEO
8. Advanced SEO
9. Introduction to web analytics
10. Intermediate web analytics
11. Advanced web analytics

One question I posed to Evan, who has computer programming skills was "If you, for whatever reason, decided to work for a publishing company to handle their metadata and you discovered that they relied entirely on Microsoft Excel to document their data, would you eventually transform their processes and create your own system?" I was surprised to hear that his answer was 'no.' Essentially, if he were to create his own program he ultimately would be recreating Excel. Excel by itself is an excellent program and should suffice for most data needs.

In addition to this, he mentioned that his company uses a dedicated database for business analytics, however they produce many of their reports in Excel mainly because almost everyone is familiar with and knows how to use Excel. User experience is key.

Relating this back to publishing industry That suggests that anyone serious about meta-data should be using a Title Management System (TMS), the best of which (and most expensive) are offered by companies such as [Stison](#) and [Klopotek](#). These systems typically allow for meta-data creation, management functionality, and the ability to export and distribute data feeds to multiple partners.

The Ideal Metadata Candidate for a Publishing Company

“My personal experience is those with the most natural (for lack of a better word) skillset in meta-data are information professionals if, for no other reason, our training in controlled vocabularies. I was hired by an eHealth publisher as a medical librarian and taxonomy specialist.”

-Mary Capps, MSIS Information Studies (Library Science)

In interviewing metadata professionals I have discovered that there is a definite need for a better trained workforce. I will now turn to exactly what the ideal candidate would be to fulfill the role of “Metadata Manager” at a publishing company.

I compared three industries (Publishing, Libraries, and Information Technology & Services) and tried to develop a profile of the typical staff member in each industry. To develop this profile I searched LinkedIn to find the typical educational background, skill set and work experience from each profession. I also explored the course descriptions for these educational backgrounds to see how relevant they were to metadata. The following page provides a quick visual of the results of this search. My findings suggest that most people who follow the Publishing track have an insufficient exposure to both the theory and hands-on use of metadata. On the opposite extreme, those who major in Information Technology or Computer Science have a strong background in data, however their additional skill sets such as the ability to write programs means that they’re unlikely to enter the publishing industry. The sweet spot, which came up multiple times in my interviews, was those who pursue a Masters of Science in Library and Information Science. Unfortunately, targeting this talent base leads to an additional set of problems. First, while library science teaches metadata, the curriculum is tailored towards metadata for retrieval - not metadata for selling. Second, libraries currently use the MARC standard and personnel are largely unfamiliar with ONIX. With all of this in mind it appears that the education and training requirements that would produce a suitable candidate for the publishing industry is

dispersed amongst multiple professions and it falls to educators and employers to wrestle these scattered components into a cohesive program.

Industry Recruitment Shortcomings

It should be noted that while the current publishing workforce is largely lacking in metadata proficiency, the publishing companies themselves have not done an adequate job of articulating the skill sets they need. Thad McIlroy and I have each done separate online searches for job openings that deal specifically with metadata in the publishing industry.

Thad's approach was to find relevant job titles that would require metadata proficiency, such as Senior Marketing Analytics Manager or Digital Distribution Coordinator. Of eight relevant job openings the word metadata appeared only twice, ONIX did not appear at all and SEO was mentioned in one job description. Figure 1 below gives a breakdown of each job title and the occurrence of relevant skills.

Figure 1

Title	Metadata Mentioned	ONIX Mentioned	SEO Mentioned
Digital Content Specialist	No	No	Yes
Digital Distribution Coordinator	Yes	No	No
Marketing Coordinator	No	No	No
Digital Publicist	No	No	No
Production Supervisor	No	No	No
Senior Marketing Analytics Manager	No	No	No
Senior Production Associate	No	No	No
Website & Marketing Asset Associate Manager	Yes	No	No

My approach was to search LinkedIn Jobs using the key terms 'metadata' and 'XML'.

Figures 2 & 3 show available jobs per industry when searching for the key terms 'metadata' and

‘XML.’ As you can see, publishing and libraries return a small number of results. The XML search returns more results, however these are typically for jobs beyond the skill set of your typical publishing employee. Job titles include: Salesforce Application Developer, Android Engineer and Associate User Interface Designer. The job descriptions for these titles typically required additional skills such as Java, C/C++, SQL, etc. It’s interesting to note that in each search result the publishing industry accounts for less than 1% of job openings. One would expect more from an industry that is expected to attach metadata to virtually every product it sells.

Figure 2

Industry - Metadata	Jobs
Information Technology & Services	172
Computer Software	106
Financial Services	65
Accounting	54
Internet	40
Management Consulting	37
Computer Hardware	35
Hospital & Health Care	17
Insurance	17
Entertainment	15
Publishing	3
Libraries	2

Figure 3

Industry - XML	Jobs
Computer Software	1,328
Information Technology & Services	1,298
Internet	422
Computer Hardware	331
Financial Services	270
Publishing	31

Conclusion and Recommendations

The results of this investigation show a clear need for the publishing industry to proactively address, and improve upon, the current low standards of metadata competency that have thusfar been neglected. This will require a concerted effort on the part of decision-makers and educators to stress the importance of metadata proficiency and make available the resources necessary to train and educate personnel. The changing conditions of the publishing industry demand an equally transformative organizational shift on the part of publishers in order to maintain and increase revenue and competitiveness. Unlike today's publishers who are still bogged down by updating their backlist, the publishers of the future will consist of a technically agile workforce that feeds on innovation. Listed below are some recommendations for publishers to reverse course and ease their way out of the metadata predicament.

1. Make an in-house assessment of your organization's supply and demand of metadata competency. Try to determine how many personnel, and at what level of proficiency, would be required to climb out of the metadata deficiency and remain out of it.
2. Be articulate when posting job descriptions. This first requires that you can articulate to yourself the problem and solution.
3. Establish and seek out employee training programs to raise metadata proficiency standards in your organization.
4. Communicate with publishing industry educators on the need to introduce metadata into the curriculum.

Industry Workforce Comparisons

The following information was obtained by data mining [LinkedIn](#)

	Publishing Industry	Libraries	Information Technology & Services
Gender			
	~55%/45% Female/Male	~62%/38% Female/Male	~10%/90% Female/Male
Education	Typical Major -English -Journalism -Publishing -Fine Arts	-Library & Information Science -Law -History	-Information Technology -Computer Science -Management Information Systems
	Metadata Training Theory: Low Hands-On: Low	Theory: High Hands-On: Medium	Theory: High Hands-On: High
Course Descriptions	Click to View Sample Publishing MS	Click to View Sample Library & Information Science MS	Click to View Sample Information Technology
Sample Skill Sets	Editorial Copywriting Proofreading	Metadata Classification Information Retrieval	Information Retrieval Program Languages Data Mining
Conclusion	Underqualified	Sweet Spot	Overqualified

Sample of Relevant Courses from Pace University's Publishing MS

Cost Per Class: \$2,961

http://appsrv.pace.edu/academics/crsdesc_all.cfm?subj=PUBIntroduction to Supply Chain Management

This course examines the strategic methodology of supply chain management; primarily in the book publishing industry. Supply chain models of other print as well as electronic publishing will be discussed. Supply Chain Management is an interdisciplinary subject and students will be exposed to many aspects of publishing – after the original work is completed and ready for publication. The topics this course will cover include: basic economic principles; supply chain models; forecasting and analyzing consumer demand; procurement and global sourcing; inventory planning; ordering and fulfillment; logistics.

Marketing Principles and Practices in Publishing

This course provides an in-depth analysis of how both fiction and nonfiction books are marketed, including hardcover trade books, children's books, and mass market books. Students learn general marketing principles and implement them by designing marketing plans using both traditional methods and new digital methods (social networking, blogs, twitter, e-mail, downloads, podcasts, and other marketing strategies using new technology). Students practice professional communication skills by writing marketing and publicity copy. The role of the marketer in the publishing process and the interaction between the marketer and other publishing professionals is examined in detail.

Practical applications of Product Management in Digital Media

The book and magazine publishing industry has undergone tremendous changes in only a few short years. With the explosion of pure online content sites, interactive tools and ebooks, and media-centric mobile applications on the market, roles that were once more common in technology fields, are now becoming standard in publishing houses as well. In the last few years we've seen a new role in particular emerge in publishing. Digital product management is no longer just for computer science or engineering majors working in software companies. With user experience, return visits, and content quality becoming the predominate drivers of successful digital media sites and apps, **publishing companies are now turning to professionals with traditional liberal arts, and publishing skills to help develop engaging media products.** Examples include everything from Travel and Leisure's destination guide, to Bride's magazine's wedding planning tools, to Richard Branson's newly announced iPad app-only magazine, called Project. Now the digital project manager-the person who oversees the creation of all of these content-driven sites, tools or mobile applications-often plays a key role in developing all of the kinds of features for publishing companies. The downside is that non-technical print publishing who transition to digital product management roles, or are in some way involved in the creation of online or digital content, **often find themselves playing catch up when it comes to understanding the requirements of digital product management as a discipline.** This course will help **take the mystery out of technical product development and methodologies,** give students **hands-on, highly sought after skills,** and bridge the gap in ways that publishing professionals can immediately put into practice. This course addresses, in both books and magazines , interactive media content. It will teach students **how to read and write requirements documentation, speak with software developers, and understand analytics, user experience** and more. Most importantly, it will give publishing professionals the working knowledge to understand this emerging field in publishing, and an edge up in this fast-changing industry.

[Return to Industry Comparisons](#)[Return to TOC](#)

Sample of Relevant Courses from Pratt Institute's Library and Information Science MS

Cost Per Class: \$4,206

http://www.pratt.edu/academics/degrees/course_catalog/METADATA: DESCRIPTION AND ACCESS

Students will explore the principles of archival description with an overview of **metadata applications** such as **Dublin Core**, **Text Encoding Initiative (TEI)**, **Metadata Object Description Schema (MODS)**, **Encoded Archival Description (EAD)**, **VRA Core**, **Categories for the Description of Works of Art (CDWA)** and **Cataloging Cultural Objects (CCO)**. The overview includes **metadata formats**, **descriptive detail for different forms of material**, choice and form of entry for creators, provision of authority control for creators, subject analysis and thesauri.

PROJECTS IN DIGITAL ARCHIVES

This class is a combination of **theoretical, practical and hands-on approaches to digital library creation**. Topics will include **metadata creation**, image capture, archival storage and Web presentation. Students will learn about the theories behind the practices that they will implement, and will gain an understanding about **the administrative issues associated with the successful implementation of a digitization project**.

HUMAN-INFORMATION BEHAVIOR

This course examines the concepts of information, information needs, and **the process of seeking information**. Models of information behavior and major theories, paradigms, and perspectives related to information seeking are addressed. Characteristics of information seeking behavior are explored as they relate to individuals and groups in various social roles, demographic areas and occupations as well as issues related to user-centered service and system design.

DIGITAL PRESERVATION & DIGITIZATION

The course introduces students to the concept of digital preservation. Issues and trends in **the creation, maintenance, and management of digital assets** that are accessible for the long term will be explored. Particular topics to be explored in-depth are **digital preservation standards, strategies, policies and systems**. As well as the selection of materials, **metadata**, copyright issues and **digital asset management**. This course aims to cover digital preservation and digitization from a variety of angles in order to expose students to the full breadth of issues raised by these new trends.

COLLECTION DEVELOPMENT

A study of methods and techniques for the development of a collection policy and practice to meet the needs of diverse user groups. The course includes **collection policy development**, censorship, methodology of user group identification, **standards, budget allocations**, selection, media and weeding.

[Return to Industry Comparisons](#)[Return to TOC](#)

Sample of Relevant Courses from Pace University's Information Systems MS

Cost Per Class: \$3,105

<http://appsrv.pace.edu/academics/view-programs/?School=GCS&Cred=MS&Maj=IS&Location=nyc&details>Database Management Systems

This course focuses on the theoretical and practical aspects of **file and database management systems**. Topics include **data models hierarchical, network, relational; data structures, storage structures, storage devices and their relation to data access; importance of data as an organizational resource; data management, sharing availability, security, integrity and consistency; data independence and conceptual data models**. Examples of database applications and software packages are selected.

Database Programming

This course is an introduction to **database programming**. Concepts and techniques of **data definition and data manipulation** using SQL will be stressed. Students will **design and implement a database in a relational database environment**. Topics covered include **creating database structure, populating the database, maintaining data, retrieving data, administering the database and optimizing queries**

Data Warehousing, Data Mining and Visualization

This course provides a foundation for teaching the basic concepts of **data mining, data warehousing, Online Analytical Processing and Visualization**. This course will focus on distinctly “real world” orientation that **emphasizes application and implementation** over design and development in most topic areas. The primary course targets are graduate information systems (IS) or operations and decision technologies electives.

Database Design and Development of Web Applications

This course applies **theoretical and applied** aspects of database design to **web-based applications**. This course will review the basics of **database technology, cover different development platforms, and develop projects that connect client-side interfaces to server-side databases**.

[Return to Industry Comparisons](#)[Return to TOC](#)

Trade Book Publishing Financial Analysis

One important measure by which to determine whether workforce proficiency is keeping up with industry demand is from a sales standpoint. To achieve this measure I've attempted to compare current sales to what sales would be if metadata competency was ideal. While attempting such a calculation can lead to arguable assumptions based off of a mish-mash of data, I believe that the analysis presented below is sound in its reasoning and based upon credible data. Considering that I haven't seen this done previously I felt it would be worthwhile to make an attempt - at the very least it might cause others to critique and improve upon my methodology.

Sources

The foundational source of information for this analysis is Nielsen's white paper on [The Link Between Metadata and Sales](#). Total revenue generated for trade books excluding religious titles for 2011 is sourced from [PublishersLunch](#)'s reference to a figure provided by Nielsen BookStats. The percentage of total sales derived from e-commerce (30%) was sourced from [Bowker's 2010-2011 U.S. Book Consumer Demographic and Buying Behaviors Annual Report](#). My estimate that on average 50% of a publisher's revenue is from its backlist is sourced from the article [What Happened to the Long Tail?](#) (*Publishers Weekly*). The estimate that one third of a publisher's revenue is from non-fiction titles is sourced from the article [A Changing Retail Marketplace](#) (*Publishers Weekly*). Finally, the estimate that 23% of an organization's backlist has incomplete BIC metadata was provided by Nielsen in an email correspondence.

Deducing the Values of the Online Sales Charts

As a starting-off point, Nielsen provides average sales data for varying levels of BIC completeness which I've recreated in a chart below (Figure 4). In Figure 1.3 of Nielsen's report they then offer a rough estimate sales distribution across varying levels of BIC completeness,

[Return to TOC](#)

Figure 4



split into two categories: Online and Offline sales. Nielsen does not provide numerical values for these sales figures nor does their Figure 1.3 chart provide gradient values. Nielsen does not attach actual sales values to Figure 1.3 in their report, but assuming that the sum of the Online sales and Offline sales charts is equal to Total Trade sales, I was able to extrapolate those values. After precisely measuring bar length and grid line space I was able to plug in a gradient value of 200 for the Offline Sales chart and a grid value of 100 for the Online Sales chart with the aggregate of these values perfectly matching Total Trade sales. Figures 5 and 6 below list my calculated Online and Offline sales.

Figure 5

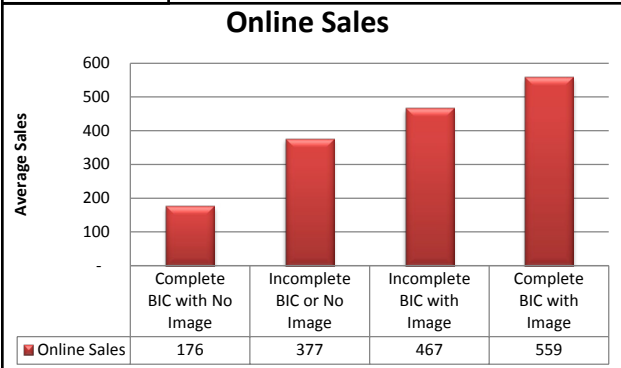
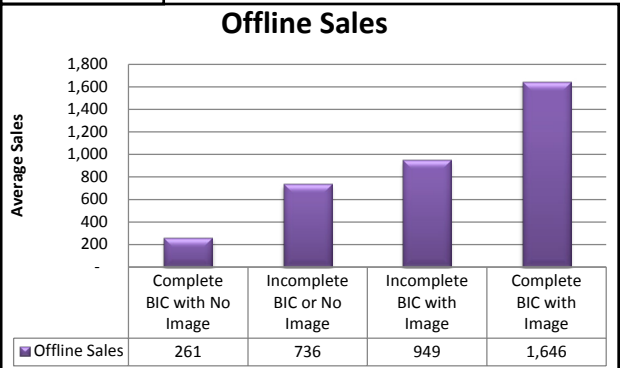
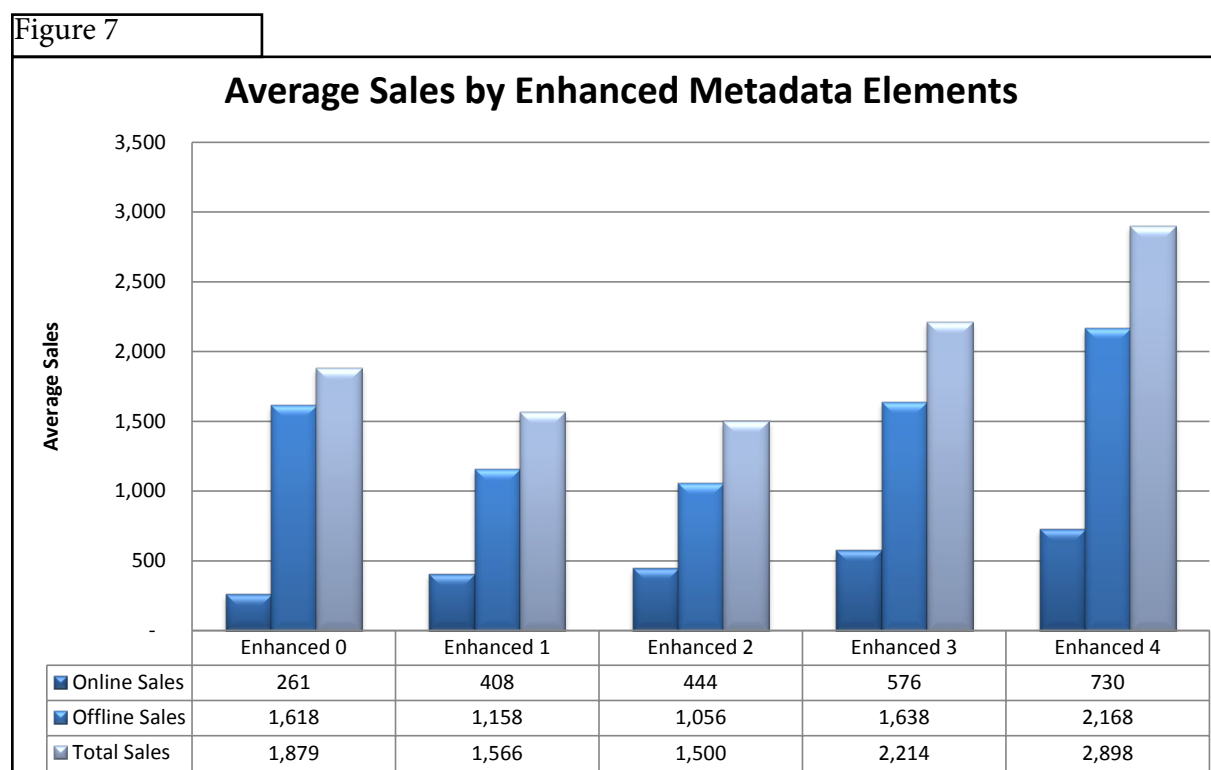


Figure 6



Deducing Enhanced Sales Estimates

Nielsen does not explicitly state the average enhanced sales values for their Figure 2.1 in their report, but using the same method described above and implying from Nielsen’s chart that Enhanced 3 is equal to 1,500, I was able to estimate the average sales values for Total Sales, then again provide a breakdown for both Online and Offline sales. Figure 7 below lists those estimates.

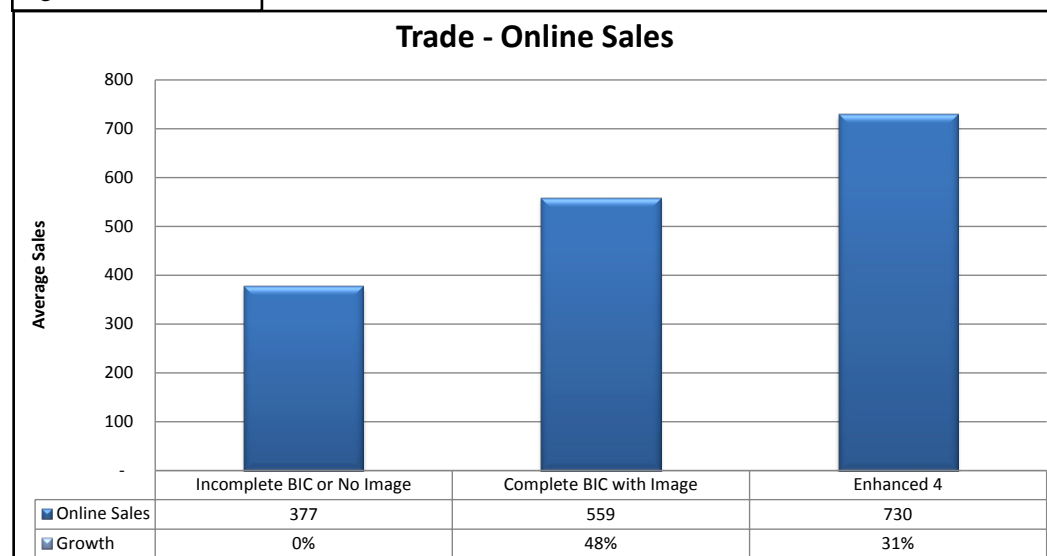


Projecting Trade Online Sales Growth - Metadata ROI

To consolidate relevant sales data Figure 8 below details average online sales for trade books with Incomplete BIC or No Image, Complete BIC with Image, and trade books tagged with the four enhanced elements stated in the Nielsen report (Short Description, Long Description, Review, and Author Biography). A 48% growth in sales is expected by upgrading a book’s metadata to have a complete BIC and an image. An additional 31% growth is expected when the

four enhanced elements are added (equivalent to a 94% total increase from Incomplete BIC or No Image).

Figure 8



To apply these values to a real world scenario I calculated the expected sales increase and resulting additional operating margin if these values were applied to 1. the trade publishing industry as a whole, and 2. a matrix listing the expected operating margin across varying base-line revenue figures (e.g., \$1 Million, \$500 Million, \$1 Billion, etc.) cross-referenced against the percent of backlist lacking adequate metadata (e.g., 10%, 20%, and so on).

This analysis was done twice, with the first analysis applied to the entire trade online backlist with inadequately tagged metadata and the second analysis applied to non-fiction online backlist with inadequately tagged metadata. Additional assumptions applied to the following analysis is that 1. 50% of a publisher’s revenue comes from its backlist, 2. 30% of a publisher’s revenue comes from e-commerce (both print and ebooks), and 3. the publisher has a 10% operating margin. Figures 9 and 10 assume that the typical publisher has 23% of its backlist inadequately tagged while Figures 11 and 12 provide a range from 10-40%.

Trade Publishing: Total Market & Non-Fiction Market

Figure 9 lists the expected industry-wide sales growth for all online backlist trade books with inadequate metadata while Figure 10 applies specifically to non-fiction online backlist trade books with inadequate metadata. To extract the total revenue that would benefit from improving metadata for backlist titles sold online I multiplied total revenue (\$12,517) by 3.5% for the total trade analysis and by 1.1% for the non-fiction analysis. To clarify I've listed the formulas below:

$23\% \text{ Inadequate Metadata} * 50\% \text{ Revenue from Backlist} * 30\% \text{ Revenue from Online Sales} = \underline{3.5\%}$ $23\% \text{ Inadequate Metadata} * 50\% \text{ Revenue from Backlist} * 30\% \text{ Revenue from Online Sales} * 33\% \text{ Non-Fiction} = \underline{1.1\%}$
--

Figure 9 shows that trade publishing had \$432 million in sales from books that were inadequately tagged. If these titles were provided with a complete BIC and an image then those sales would grow to an additional \$207 million. Adding the four enhanced elements would further grow sales an additional \$198 million, bringing total growth to \$405 million. Under this analysis the trade publishing industry stands to add \$41 million to its operating margin by properly tagging its entire backlist.

Figure 9

Trade Publishing (2011) - Total Online Backlist	Value (000's)
Revenue	\$12,517
Percent of Revenue from Online Backlist with Poorly Tagged Metadata	3.5%
Sales of Poorly Tagged Online Backlist	\$432
Estimated Growth if BIC & Image were Complete (48%)	\$207
Additional Estimated Growth if All Four Enhanced Elements Included (31%)	\$198
Total Growth (94%)	\$405
Gain on a 10% Operating Margin	\$41

To provide a more conservative figure I conducted the same analysis for non-fiction titles (Figure 10). The rationale for this is that fiction books are more easily substituted - one can easily be replaced by another. On the other hand, if a consumer cannot find a specific non-fiction book

they are less likely to make an alternative purchase - in many cases only one book will “fit the bill.” This analysis reveals that from the \$144 million in sales from titles with inadequate meta-data, the industry can gain a total of \$135 million in revenue (\$69 from including a complete BIC and image; \$66 from further adding the four enhanced elements). Under this assumption the trade publishing industry stands to add \$14 million to its operating margin by properly tagging its entire backlist.

Figure 10

Trade Publishing (2011) - Non-Fiction Online Backlist	Value (000's)
Revenue	\$12,517
Percent of Revenue from Online Backlist with Poorly Tagged Metadata	1.1%
Sales of Poorly Tagged Online Backlist	\$144
Estimated Growth if BIC & Image were Complete (48%)	\$69
Additional Estimated Growth if All Four Enhanced Elements Included (31%)	\$66
Total Growth (94%)	\$135
Gain on a 10% Operating Margin	\$14

Sales vs. Percent Backlist Matrix to Estimate Additional Operating Margin

As publishers come in varying sizes and have differing degrees to which their backlist metadata needs correcting, I created a simple matrix that will allow a publishing professional to easily estimate how much they could add to their operating margin by removing metadata deficiencies from their online backlist. Figure 11 considers the total trade market while Figure 12 offers data for just the non-fiction market. The far-left vertical row indicates a publisher's total revenue. The top horizontal header indicates the percentage of backlist that is inadequately tagged. The body of the chart lists the corresponding anticipated increase in operating margin. As an example, if a publisher generates \$100 million in revenue and has 30% of its backlist inadequately tagged, then its operating margin can be improved by \$422,460 if it were to fix its

backlist. Focusing solely on non-fiction backlist would increase operating margins by \$140,819.

Figure 11 - Potential Operating Margin Gain by Updating Total Online Backlist

Sales vs % Bad Metadata	10%	20%	25%	30%	40%
\$100,000	\$141	\$282	\$352	\$422	\$563
\$500,000	\$704	\$1,408	\$1,760	\$2,112	\$2,816
\$1 Million	\$1,408	\$2,816	\$3,521	\$4,225	\$5,633
\$10 Million	\$14,082	\$28,164	\$35,205	\$42,246	\$56,328
\$100 Million	\$140,820	\$281,640	\$352,050	\$422,460	\$563,280
\$500 Million	\$704,100	\$1.4 Million	\$1.7 Million	\$2.1 Million	\$2.8 Million
\$1 Billion	\$1.4 Million	\$2.8 Million	\$3.5 Million	\$4.2 Million	\$5.6 Million
\$2 Billion	\$2.8 Million	\$5.6 Million	\$7.0 Million	\$8.4 Million	\$11.3 Million

Figure 12 - Potential Operating Margin Gain by Updating Non-Fiction Online Backlist

Sales vs % Bad Metadata	10%	20%	25%	30%	40%
\$100,000	\$47	\$94	\$117	\$141	\$188
\$500,000	\$235	\$469	\$587	\$704	\$939
\$1 Million	\$469	\$939	\$1,173	\$1,408	\$1,878
\$10 Million	\$4,694	\$9,388	\$11,735	\$14,082	\$18,776
\$100 Million	\$46,940	\$93,879	\$117,349	\$140,819	\$187,758
\$500 Million	\$234,698	\$469,395	\$586,744	\$704,093	\$938,791
\$1 Billion	\$469,395	\$938,791	\$1.2 Million	\$1.4 Million	\$1.9 Million
\$2 Billion	\$938,791	\$1.9 Million	\$2.3 Million	\$2.8 Million	\$3.8 Million

Concluding Remarks on the Financial Analysis

The reader should keep in mind that the metadata will not be fixed overnight nor within one year. Accomplishing this feat will take a concerted, multi-year effort. With that in mind it is not inconceivable for the industry to grow at the above rates over the course of multiple years. One must also consider the cost reduction benefits associated with the increased efficiencies made possible by e-commerce. An improvement in metadata will lead to reduced friction throughout the supply chain and ease consumer purchases.

As a final note, some readers may be skeptical of the degree to which improved metadata can positively impact a publisher's sales. Perhaps a more convincing exercise would be to try and consider the degree to which a *degradation* of an organization's metadata by 23% would *negatively* impact sales.