The Book On Metadata

by Laura J. N. Dawson

This is just a little booklet based on my Metadata Boot Camp sessions, for those who can't attend them or who want reminders of what we discussed. Most of it is based on my own experience in my 30 years in the book industry. I've tried to link to sources for historical notes, but they're not always available and I'm relying on memory in some cases.

The point of this project is not necessary to offer a how-to guide on metadata best practices - you can get that from the Book Industry Study Group, or by reading Renée Register's "Metadata Handbook". Instead, I'm attempting to contextualize metadata use, looking at how it has evolved since computers began permeating the industry in the 1960s - and to extend the implications of that evolution to understand how we might use metadata in the future.

The book industry is its own universe. From monasteries to printing presses to Amazon, there are a lot of legacy reasons why we do things the way we do. Other industries look to us as an example of a business that makes good use of metadata, while we ourselves see so much room for improvement. That dichotomy is a rich source of inspiration to me, and I hope it is for others as well.

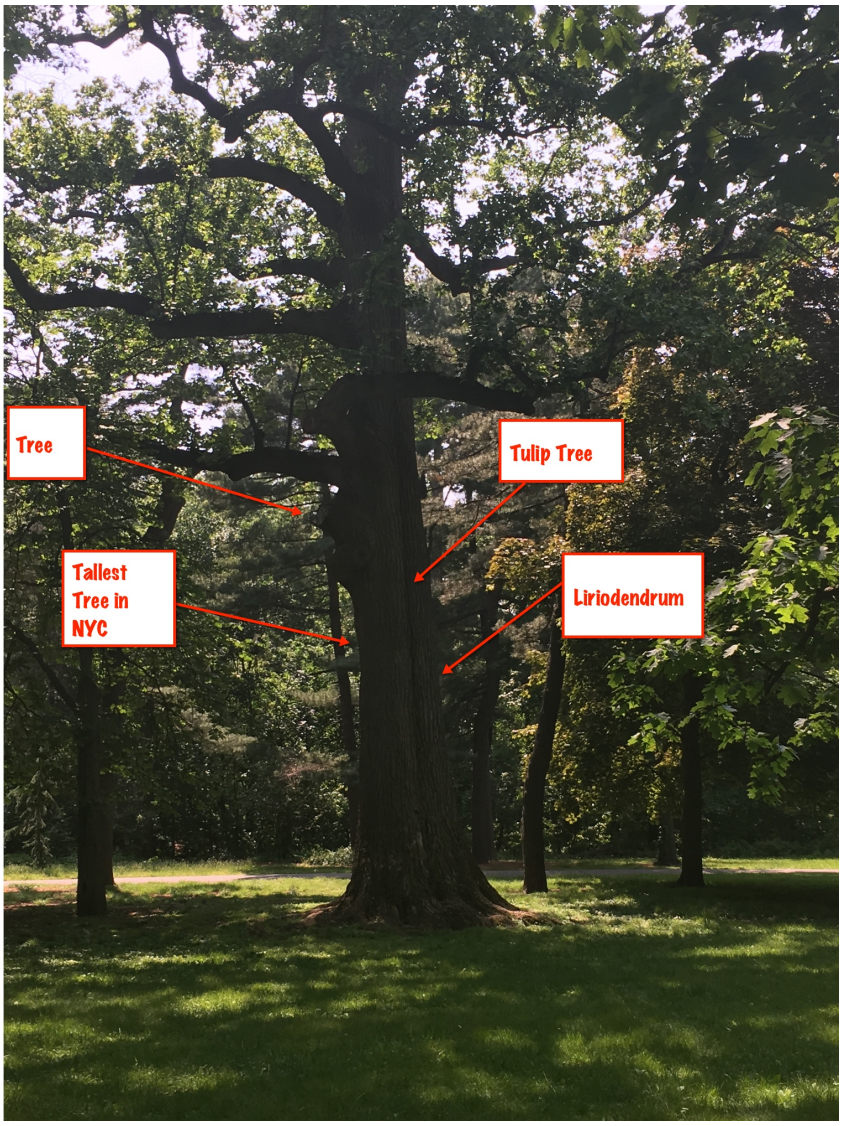
— Laura Dawson, 2016

CHAPTER ONE

Foundations

This first chapter is about bringing us all up to speed on the basics of book industry metadata - what metadata is, how complicated and ambiguous it can be, and how we arrived at this period in 2016 on the verge of ONIX 3.0.

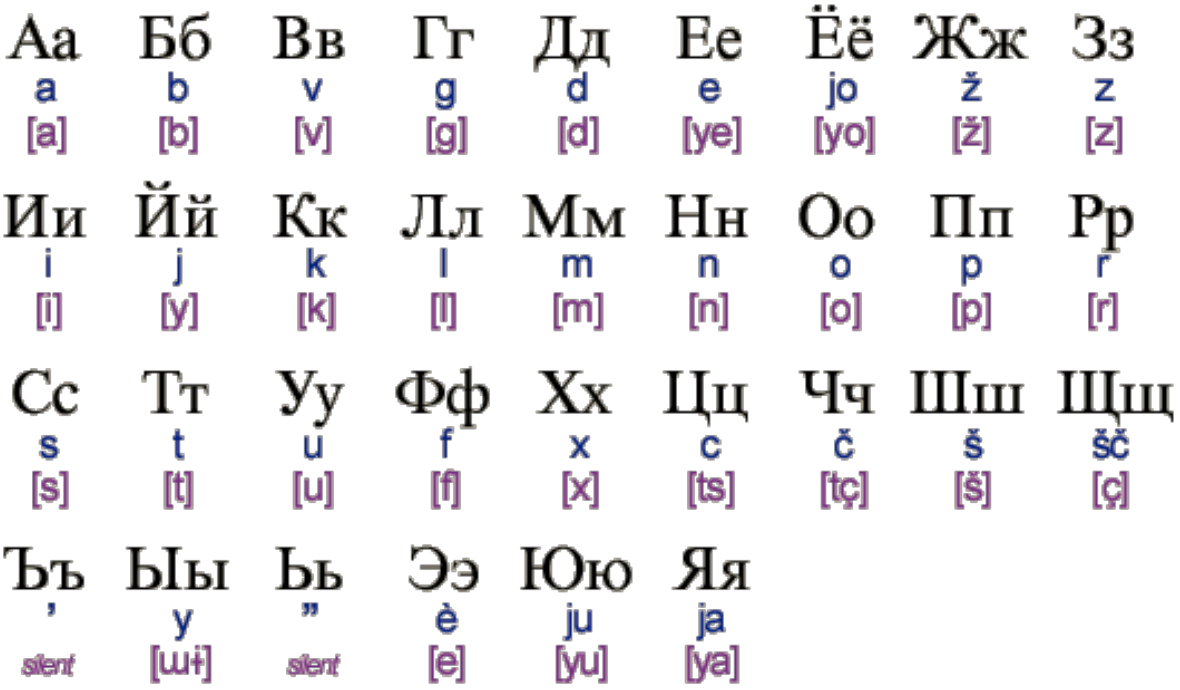
We create metadata by describing things. Behind every metadata term is a descriptor - a word that we use to describe something.



But humans have a lot of descriptions, and not all descriptors suit all circumstances. In this picture of the tree, there are scientific descriptors, botanical descriptors, and regional/local descriptors.

And of course, we speak many languages, so let's multiply those descriptors by the number of languages in the world (currently, depending on who you ask, somewhere between 6500-7100).[1]

There are also multiple character sets – different alphabets and scripts that also determine how things get described. (The example below is the Cyrillic alphabet, used in Slavic languages.)



Even within the same language, there are regional descriptors. In American English, we have many words to describe a large sandwich, for example - hoagie, grinder, sub, po’boy. We have a rich vocabulary to describe soft-soled athletic shoes - sneakers, tennis shoes, kicks, pumps, running shoes, etc.

Maps, knitting patterns, recipes, ledgers, inventory lists, genealogy - as long as there are things to describe, count, and measure, in any language, there’s metadata. Through all this morass of languages, alphabets, and types of descriptors, we’re looking at a real mess of words.

Standards are NOT perfect. But they do impose a solution. By definition, standards are impositional – they are created by representative consensus and are a top-down application. Standards don’t perfectly handle nuance. Or context. Standards always leave something out, have exceptions, edge cases…

But standards are the signposts we have to navigate the mess of meaning in our world.

A Detour Into Some History

Born in Romania, Emery Koltay grew up in Hungary and was active in the Hungarian resistance to Communism. He spent a lot of time in and out of prison camps before eventually fleeing to the US, where he began working…for R. R. Bowker.[2]

Bowker’s Books in Print database has long roots, going back to the 1800s.[3] It was (and still is) intended to be a catalog of all the books being published in the

United States. (Later versions include the UK and Australia as well.)

While he was there, the UK bookseller and wholesaler WH Smith was building a new warehouse that was going to be computerized. This system required every book in the warehouse to be numbered. WH Smith recruited a man named Gordon Foster, who had worked at Bletchley Park during World War II and later worked with Alan Turing at the University of Manchester. In 1965, Foster developed the SBN, a nine-digit number that WH Smith could use to identify editions of books in their new computer system.[4]

Koltay followed the SBN's progress from the US, and introduced the concept to US publishers as part of his work at Bowker. That set the stage for the number to be refined and ratified by ISO – the International Standards Organization. Within 4 years, ISO published the standard for global use – the fastest an ISO standard had ever been approved.

The ISBN remained at 10 digits until 2005, when it evolved into a 13-digit number to align with the EAN barcode, which was becoming the standard in retail.[5]

It's important to remember that the ISBN was initially created to solve the problems that digitization was bringing to the book world. Our so-called "digital disruption" is actually a culmination of events that began in the 1960s, when computers began to be more widely used outside of military and academic settings.

The Road to ONIX

BISAC originally was created in the 1970s. It stood for Book Industry Systems Advisory Committee.[6] As warehouses and retailers began using dial-up to communicate about inventory and orders, a standard format was needed to make communications easier. A fixed-format electronic data interchange was ideal – the 10-digit ISBN, a truncated title name (just for corroboration against possible keying errors in the ISBN), possibly the name of the author, the status, and the price. All of this was entered manually at the publisher's, the distributor's, and the retailer's. The term "fixed format" refers to the fact that the field lengths were limited – 10 digits for the ISBN, and limitations on the other fields as well.

As computing power grew, fixed fields became more cumbersome than they were worth. ANSI X12[7] allowed for variable field length. But it was (and remains) an inter-industry standard. In the 1980s and 1990s, we saw the rise of the book superstore, which carried items not strictly limited to books – gift items, beverages, toys. Bookstores needed to get information from non-book suppliers and needed a standard format that both publishers and manufacturers

would adhere to. This drove the adoption of the X12 standard for business-to-business transactions in the supply chain.

ANSI stands for American National Standards Institute. So ANSI X12 was a US-only standard. You may have heard of NISO – the National Information Standards Organization. NISO serves as ANSI's designated secretariat to the International Standards Organization's committee that oversees ISBN and other book-related standards.

BISAC itself evolved during this time to stand for Book Industry Standards And Communications[8] – which developed categories for books so that bookstores would know where to shelve them. A book could only have one BISAC subject heading at that time – stores wanted a definitive category for each title, so there was no confusion about where the book was located.

In the 1980s, BISAC merged with BISG, which had, until then, been simply a research group studying the trade book market.

In 1994, Tim Berners-Lee and the W3C openly released the technology that harnessed HTML to the Internet (each of which had been developed separately from one another), creating the World Wide Web. That same year, a company named Cadabra was founded by Jeff Bezos.

In 1995, Cadabra evolved into Amazon.com and proceeded to disrupt the book industry with remarkable speed. You might remember Amazon's early days (unfortunately the Wayback Machine only goes back to 1998) – misspelled, all-caps, truncated titles, the lack of descriptions about the books, no cover images.

This is because ONIX hadn't been invented yet.

ONline Information eXchange

By 1997, superstore behemoth Barnes & Noble had launched their website in competition with Amazon. (As a matter of disclosure: I began working there in 1998, directing the database that served both the web and the stores.) Borders, Hastings, Books-A-Million, and others soon followed. There were numerous start-ups dedicated to selling specific sorts of books – Varsity Books for textbooks, FatBrain for business books. Companies were acquired, rose up, shut down – the late 90s and early 2000s were absolutely chaotic in the web world, and the book industry certainly saw its fair share of disruption.

In the midst of all this, the problems presented by back-office, transactional metadata were abundantly clear to consumers – these websites were ugly, clunky, and not very enticing. Publishers noticed. Distributors noticed. Everyone saw an opportunity to increase sales.

ONIX was developed as a joint effort by the Association of American Publishers (AAP) and EDItEUR (which originally stood for EDI-to-Europe, but which has evolved more broadly into a London-based standards body for the book industry). It was created to solve two problems: (1) that consumers were now looking at this data so it had to be more robust, descriptive, accurate, and reflective of what they needed to see, and (2) that ANSI X12, as a US-based standard, was insufficient for international communications about books.

In the interest of context, it was hotly competitive between Amazon, Borders, and Barnes & Noble. There were lawsuits, front-page news articles, insults, and shade thrown. It was nasty. But everybody could agree that the metadata was causing us all the same headaches. So there was a parley.

In 1998, in the conference room at AAP on 5th Avenue, a group of publishers, Barnes & Noble, Amazon, Ingram, Baker & Taylor, a handful of startups, and a number of other interested companies sat down at a large conference table and laid out the problems. It was the first time B&N and Amazon had allowed representatives to sit in the same room together. It was clear that the competitors needed to present a unified front to persuade publishers to adopt this new standard that would benefit all of us. After two years of negotiations, ONIX 1.0 was published and its maintenance in the US was handed to the Book Industry Study Group, which created a metadata committee to handle changes and fixes and additions to the code lists.[9]

Discovery

So that's how we got to where we are. But where do we go from here?

As our worlds become increasingly virtual, effective identification and description is even more critical. A virtual world is an invisible one…until it's described to you. When you're on the subway, you're not going to know what someone's reading on their phone. But if someone tweets a link to an Amazon book page, a publisher will want to be sure that the metadata for that book is complete, clean, and robust.

You'll want to remember that there is no such thing as one monolithic "METADATA". The evolution of book metadata has taught us that there are different data sets for different audiences. The stuff we used to send around in BISAC Fixed Format was transactional – what is the thing, how much does it cost, is it available to buy – this information was sent around from the warehouses of publishers to the warehouses of distributors to the warehouses of retailers. Office workers – book buyers at retailers, publicity folks, acquisitions librarians – relied on much less efficient communication: the paper publisher catalog. The visit from the sales rep. As late as the early 1980s, some publishers

were still selling door-to-door!

So when you think about metadata as a concept, you need to ask, "Who's it for? What do they need to know? When do they need to know it?"

There is a set of metadata elements that are considered "must-haves" by the industry:

- Product identifier (ISBN)*
- Title*
- Contributor*
- Contributor identifier (ISNI)
- Format*
- Price*
- Status
- Company
- Dimensions/file size/weight/page count
- Release date
- Language

This list is the basic information that retailers, distributors, librarians, consumers, and publishers need to begin trading on a book. (And by trading I mean selling, buying, setting up records in systems, etc.) The identifier (99% of the time it will be an ISBN), the title (which can change over time), who wrote it/illustrated it/etc., the identifier of that contributor if it's available, what format it comes in, how much it costs (and in what currency), what the status is (available, not yet available, no longer available), who's publishing it, how big the thing is, when it's coming out, and what language it's written in. So EVERYBODY needs this stuff to even *begin* to make a decision about the title. Whatever the nature of that decision is going to be.

The elements that I've asterisked are those that are the BARE MINIMUM you can possibly get away with in communicating with vendors.

"Enhanced" metadata is what ONIX was invented for – it's the consumer-friendly marketing information that really sells a book online (and also to a book buyer at a retailer or an acquisitions librarian at a library). Most distributors and vendors ask for a short description (maybe 3 sentences), and then a longer one. They'll ask for review excerpts if the book's been released for review (or you can add them later). An image is essential. What follows is a list of metadata elements considered to be "enhanced" by the industry:

- Descriptions
- Reviews
- Image

- Categories (BISAC)
- Contributor bio
- Contributor image
- Endorsements/quotes

These days, we are permitted more than one BISAC! Most vendors require at least 3, some like 5. Many specify that one of those BISAC codes needs to be identified as the "primary" one – the main thing that the book is about. Try to stay away from "General" BISAC categories – these don't give you much of a competitive advantage at all. All of the specific categories roll up under the general ones, so by sticking with general categories, you're a tiny fish in a really huge ocean.

A contributor bio helps a great deal because, especially if it's a nonfiction title, it gives the reader some indication of the author's credentials and why his book should be given attention. An author photo is helpful as well – in the same way a cover image is. And, of course, endorsements and quotes from other authors or public figures are a major selling point.

In general, both "basic" and "enhanced" metadata is going to come from the publisher itself, because that's where the book is being created. However, downstream stakeholders (distributors and retailers, as well as libraries) tend to append even more data in their records. So we also have vendor-created metadata:

- Location (libraries, stores)
- "People who bought x also bought y"
- Personalized recommendations
- Sales ranking
- Editorial descriptions
- Categories (mapped from BISAC)

Libraries and stores need to ensure that staff knows where to find the book – in which section, on what shelf. Online stores rely heavily on the "people who bought x also bought y" algorithm to upsell titles. Stores will use user preferences to create specific recommendations for an individual consumer. Sales ranking or circulation statistics also get generated and are used as tools to promote titles. In many cases at an online retailer, a merchandising staffer will write their own description of the book in question (similar to the "shelf talkers" that are in some bookstores, where the bookseller writes a little summary stating why they think the book is worth buying).

And, of course, there are categories. Very few vendors (and even fewer, if any, libraries) use the BISAC codes as they are supplied. Each vendor is going to have their own taxonomy, one that they've curated specifically with their

customers in mind. Barnes & Noble's is different from Amazon's. Both are different from Powell's or other online retailers.

What happens is, these vendors receive ONIX feeds from publishers that include the BISAC codes. These codes are mapped to the vendor's proprietary codes. Mapping is a little complex – many categories can't be matched one-for-one, and sometimes the BISAC codes lag a little behind what consumers are actually looking for and the language they use to look for it. We'll delve more deeply into "why BISAC is the way it is" in a future chapter, but for now, just know that it's not always going to be a one-to-one match and you might have to tweak your categories or get your vendor contact on the phone to figure out what the best BISACs are for your books.

And then there's consumer-generated metadata:

- Customer reviews
- Ratings

This metadata is kind of a wild card for publishers, who can't control what happens here.

B&N introduced customer reviews in 2000, in response to Amazon's similar functionality. At the time, this was shocking to many – to authors especially, but also to professional reviewers and even editors and publicity departments. The idea that any random person could stick up a negative review and affect the sales of a book was just horrifying. Many authors insisted that their work had to be reviewed only professionally – and we couldn't, of course, toe that line. Again, there were lawsuits, and threats of lawsuits. And there was the endless bickering over what was acceptable and what was not.

The difficulty here is that before customers had the ability to write reviews of the books they read, publishers and professional reviewers really served as gatekeepers to Literature with a capital L. A more democratic landscape was difficult for them to navigate. Little by little, book review publications began to shrink and then go out of business. Goodreads arose (and was ultimately bought by Amazon). And now we're accustomed to seeing reviews, we actively WANT them, and we can do things like sort by ratings. Consumer-generated metadata has become an integral part of bookselling.

Libraries have their own metadata elements:

- Basic + enhanced metadata
- Copyright year
- Location (call number)
- Classifications (Library of Congress, Dewey, etc.)

- Access levels (public, restricted, reserve-only, etc.)
- Dates (check in, check out, due date, etc.)
- RFID (for physical objects)

The larger libraries make use of publisher metadata (what's found in both the basic and enhanced levels), but because they have other functions, they require a host of other data points that retailers and consumers might not. Because they are archivists, they want the copyright year of the book. Like physical bookstores, they need to note the location of the book (which they call the "call number"). Classification systems for libraries are more precise than retail classifications, so librarians have developed the Library of Congress subject headings as well as the Dewey Decimal System.

And libraries don't sell books. They provide access to them. In some cases, any member of the library public can check out a book; in other cases, access is restricted to within the library only. In universities, instructors might put books "on reserve" for students who don't wish to purchase them – then access is limited to students of that instructor and usually has to be within the confines of the library building.

For books that can be checked out of the library, there's additional metadata specifying when the book is checked in, what the due date is, when it was checked out. And, of course, for physical objects, most libraries now glue an RFID tag that identifies each object separately (so if they have 10 copies of a certain title with the same ISBN, each copy gets its own identifier).

In the next chapter, we'll do a deep dive into BISAC subject headings, other taxonomies, and category mapping.

CHAPTER TWO

Words Mean Things

What do I mean by "words mean things"?

The world of books represents the world of human thought. Concepts articulated, written down, codified, published. But of course, our understanding of these concepts can vary – by nationality, cultural background, experience, philosophy of life. The word "alienation," for example, can mean different things to different people. It can be expressed differently in different languages – by a single word, or by a phrase rather than a word. And, in fact, in cultures all over the world, many words can be used to describe phenomena like "snow", "walking" – think of how we describe colors in the Crayola box, for example, or the Pantone chart.

Words carry nuance that's not always immediately apparent, which is why non-native speakers of languages tend to struggle, and why translations nearly always lose meaning.

Taxonomies, in particular, are inherently political and authoritarian. They are hierarchical. Taxonomies are, essentially, what we call "controlled vocabularies". Which begs the question: Who controls them? Do we trust those people to express what we mean? What if we disagree?

Words carry more than just their semantic meaning. They carry allegiance, alliance, politics. Whether you call a country Myanmar or Burma is very dependent on your political outlook.

As in politics, taxonomies evolve as society evolves. What used to be "Occult" became "New Age", which became "Body/Mind/Spirit".

Taxonomies reflect our understanding of phenomena. And that understanding is deeply colored by our culture, our experience, our politics, and our vision of the world. It varies from person to person. Taxonomies are a compromise, a consensus.

They're the result of committee work. They are rarely finalized. They shift and change depending on cultural mood, society's evolution, and market trends. They are living things.

For those of us in the book trade, BISAC is the most familiar controlled vocabulary.

BISAC

BISAC began as a standalone initiative, designed to create standards for EDI transmissions in the book supply chain, largely between ordering departments and warehouses. These were early versions of what ONIX would eventually become – a way of communicating among trading partners.

With the rise of book superstores such as B&N, Borders, and Books a Million,

it became apparent that an additional standard was needed to determine where in these stores books should be shelved. BISAC took on the responsibility of creating standardized codes that publishers could use to suggest to bookstores which section of a store a book would be a good fit for.

Initially there were only a handful of codes. By 1995, there were around 50 general codes, with a few "sub-codes" under each – forming a 2-level hierarchy. The codes were rather cryptic – 3 letters followed by some numbers – because they were developed for machine-to-machine processing. The actual names of the codes were only used by those doing the assigning and those receiving the books and deciding where to put them. (There were those nerds who had the codes memorized, because there were so few of them.)

But with the emergence of online retailers, BISAC experienced a period of rapid change. BISAC codes were developed with an eye towards discovery on the web as well as in-store placement of books. Whereas bookstores required a single BISAC code, web stores could "shelve" a single book in multiple categories. Most guidelines now recommend 3-5 codes per title.

You might notice, for example, that the "Body, Mind and Spirit" BISAC categories begin with the characters "OCC".

This is because that category used to be called "Occult" back in the 80s and 90s. It was where books about UFOs, spiritual healing, crystals, Wiccans, and other titles were shelved. The OCC prefix evolved into the "New Age" category. As "Body, Mind and Spirit", it has been expanded to includes books about mindfulness, meditation, reiki, "inspiration and personal growth", and feng shui, all of which are fairly mainstream, in addition to continuing with more obscure topics such as astrology and numerology.

So if the BISAC prefix doesn't match up to the name of the category itself, it probably had a previous life as a category more appropriate for the 80s or 90s cultural landscape. Books reflect our landscape, and their subjects evolve over time.

Library Classifications

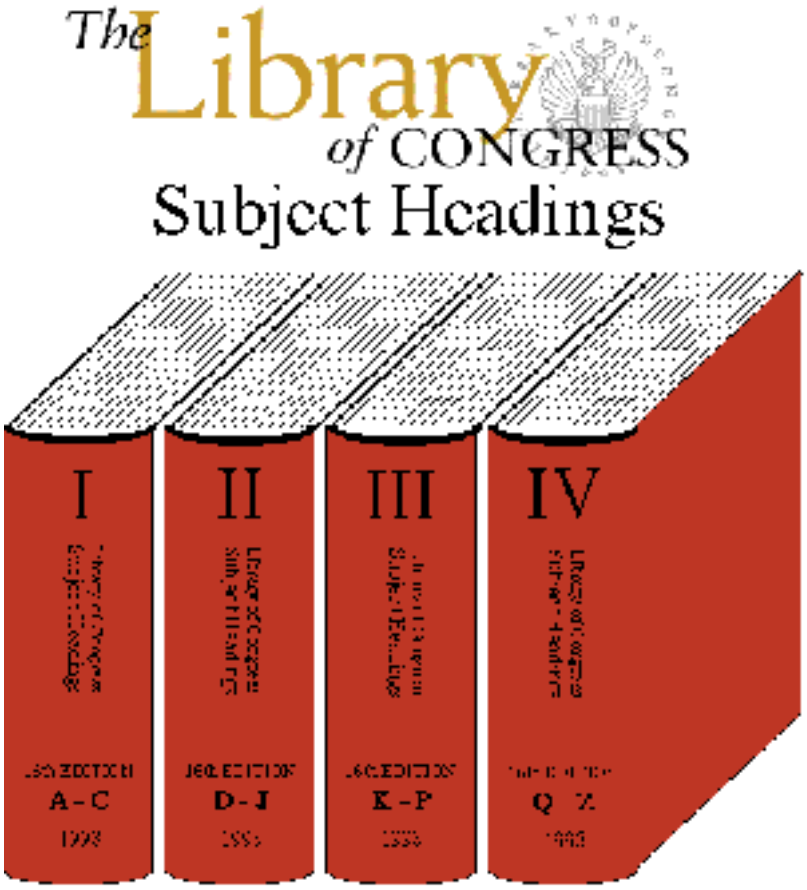
It's helpful to understand the way libraries categorize their books (and other materials) as well.

001-099	Generalities	PREQUEL TO THE DEWEY DECIMAL SYSTEM	Encyclopedias, curiosities and wonders, unexplained mysteries
100-199	Philosophy	WHO AM I?	Books about the self, feelings, dreams, witchcraft
200-299	Religion	WHO MADE ME?	Christianity, Judaism, Buddhism, Hinduism, etc., and Mythology
300-399	Social Science	WHO'S THE GUY IN THE NEXT CAVE?	Customs, cultures, laws, manners, costumes, fairy tales
400-499	Languages	HOW DO I TALK TO THAT GUY?	Dictionaries, parts of speech, sign language, foreign language aids
500-599	Natural Science	LET'S TALK ABOUT THE WORLD WE SEE	Mathematics, earth, astronomy, chemistry, plants, wild animals
600-699	Applied Science	NOW LET'S MAKE STUFF OUT OF WHAT WE SEE	Inventions, robots, transportation, pets, recipe books
700-799	Arts and Recreation	NOW LET'S HAVE SOME FUN	Art, drawing, comics, handicrafts, music, games, sports
800-899	Literature	**LET'S TELL OUR CHILDREN HOW WONDERFUL WE ARE**	Poetry, plays, classic literature, jokes, riddles
900-999	Geography and History	**LET'S TELL OUR FUTURE CHILDREN HOW WONDERFUL WE WERE**	Landforms, travel, atlases, exploration, countries
92 and 920	Biography and Collective Biography	FIND OUT ABOUT FAMOUS PEOPLE	Single person: filed by last name of subject **Multiple people: by author**

Grand View Elementary School Library – MBUSD – October 2009

Most of us of a certain age had to learn about the ***Dewey Decimal System*** in school. The Dewey system is a type of "enumerative" classification system. Numbers indicate what a given book is about. If the string of numbers after the decimal is particularly long, then it's a book about an extremely specific thing. It's fairly simplistic, and tends not to be so flexible for libraries cataloguing detailed or nuanced material.

But there's more to library categories than Dewey!



The ***Library of Congress subject heading***s were formalized in 1898 and have been actively maintained ever since. There are hundreds of thousands of LC subject headings, and there is usually some debate about their precision, reflection of cultural reality, and political intent. LC subject headings are displayed as strings, but there is an implied hierarchy to the elements in the strings.

LC subject headings not only take into account what material is about. They also take into account chronology (e.g. "this book was written about the 1920s") and place ("this book was written about life in Paris"). LC categories are complex strings, refined by dashes. The order of words is important and standardized.

Examples:

- Great Britain—History—Anglo-Saxon period, 449–1066
- Civil rights—Religious aspects—Buddhism

MeSH are specialized subject headings for the medical community. Library of Congress headings were not precise enough for this constituency, so the National Library of Medicine created these and maintains them in a committee-like environment similar to our BISAC committee.

Facets

Just as online bookstores changed the way we search for books to buy, online library catalogs changed the way we search for information to access. The long strings of LC subject headings proved to be cumbersome and unwieldy in a user interface. So in 1998, OCLC developed the FAST system – Faceted Application of Subject Terminology. This allowed Library of Congress codes to be displayed in a hierarchy for easy browsing.[10]

OCLC, for those who aren't familiar with it, used to stand for Online Computer Library Center (it no longer includes its full name and just goes by OCLC as a brand, somewhat like Xerox). It is a membership organization where librarians create records for books they've acquired, and contribute those records to a large catalog called WorldCat. OCLC itself does quite a bit of data processing from a variety of sources, in addition to its own manual cataloguing. And its research division does a lot of interesting projects with book metadata – including FAST, and laying the groundwork for the ISNI initiative.

Mapping in Commerce

When a publisher communicates information about a book to a distributor or retailer, that publisher will assign a series of BISAC codes. Online retailers, as we

know, have their own proprietary codes, based on how their users search and browse for books. Retailers such as Amazon and Barnes & Noble.com tend to look at BISACs as useful suggestions. They map BISAC codes to their own codes.

Just as we lose meaning in translation, mapping data points always results in the loss of some meaning or context. Some categories don't cleanly line up to others. So mapping one taxonomy to another is yet another compromise – one we have to live with in a taxonomic, hierarchical world.

Categories on the Web

Outside of traditional "Amazon-like" commerce, the web upends this hierarchical, one-to-one model. It allows for complexity, multiple simultaneous states, and nuance. And we're beginning to see this in a variety of initiatives, so it's a good idea to look at categorization on the web and how that has evolved.

Most of us are familiar with meta tags. These were the first bits of web metadata, describing a web page so that search engines would pick up on it. But web developers started gaming the system, inserting irrelevant keywords to drive traffic, and search engines no longer look at meta tags with the same degree of importance.

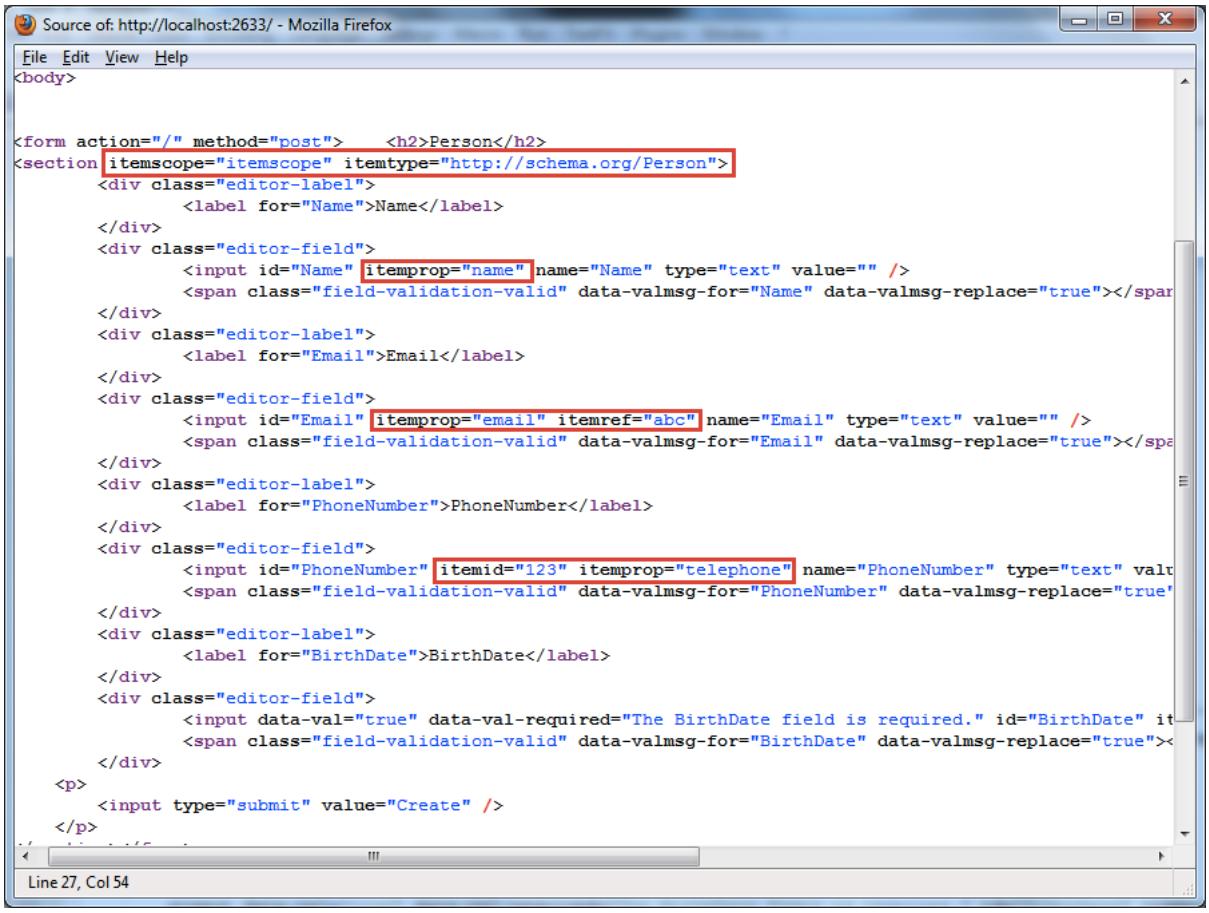
Insofar as there was a standardized vocabulary for the meta tags, it was Dublin Core.[11] Named NOT after the Irish city but after a Columbus, OH suburb where OCLC is located, Dublin Core was developed to describe resources you might find on the web – video, audio, a web page itself. Initially there were 15 key elements, but as the web grew in complexity, so did the size of the elements list.

In 2008, Dublin Core spun off from OCLC and became its own organization, called DCMI (Dublin Core Metadata Initiative). Its controlled vocabulary is still in use, and forms the basis for ontologies being developed now.

Given the unreliability of meta tags, something had to help search engines get beyond the alternative - analyzing enormous blobs of text, which are notoriously unreliable. Metadata helps to fine-tune these blobs, confirming their subject matter. It helps search engines distinguish between Mercury the planet and mercury the element, for example.

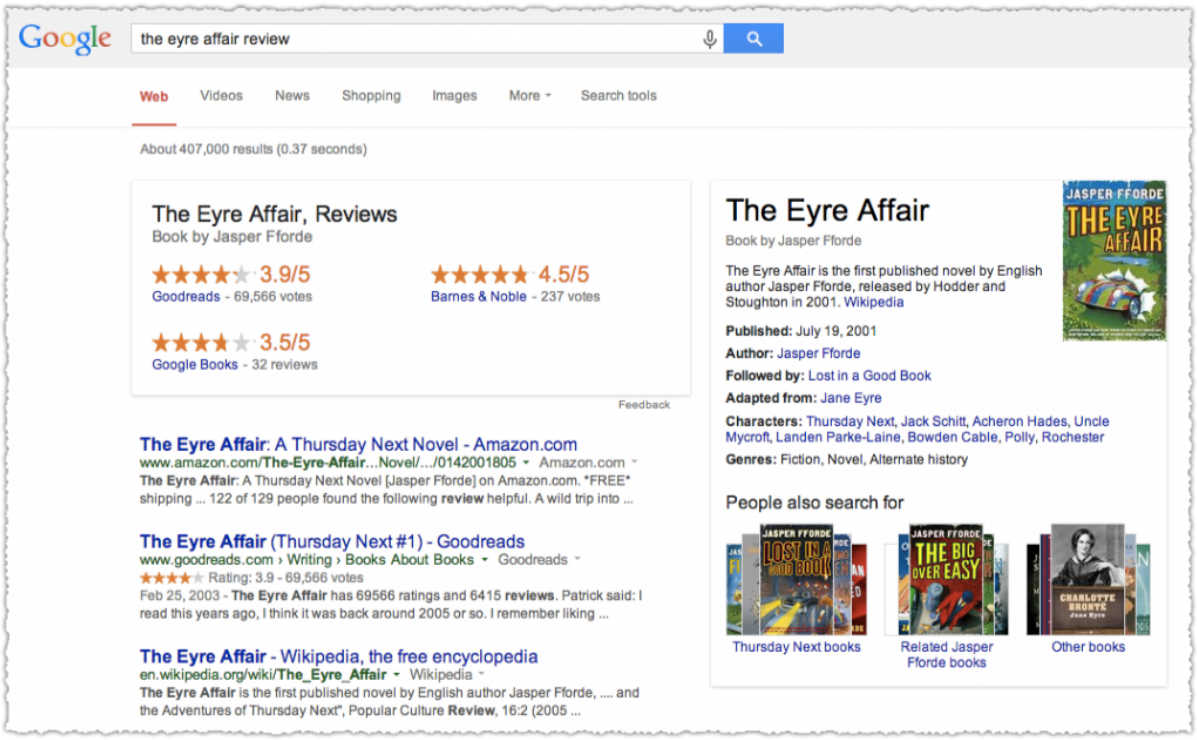
In the late 1990s, there were initiatives to provide what was called "structured markup" for web pages. Ultimately, the standard that won out was RDF. RDF is a W3C HTML standard for tagging websites so that search engines can display their contents in a structured, organized way.[12]

But what goes IN the RDF tags?



Schema.org is a relatively new initiative that launched in 2011 as a combined effort by Yahoo, Bing, and Google. Yandex, the Russian search engine, joined a few months later.[13]

Schema.org consists of mutually-agreed-upon (by the search engines as well as by volunteer contributors) vocabularies to describe persons, things, places, and events. These categories are given preference by search engines and are used to create…



…displays like this one.

You'll note that Wikipedia is nearly always cited in the Knowledge Panel, to the right. This is because Wikipedia content is marked up with Schema.org RDF elements for optimal mining by Google's spiders.

Amazon and B&N are already using a few (in my opinion not enough) Schema.org elements in their markup. But publishers are slower to implement. Penguin Random House does not appear to be using it at all, and Hachette only seems to be using it to define that "the thing we're talking about here is a book".

We continue to continue. BISAC categories aren't going away – they'll continue to evolve as we require them. Retailers will map BISAC categories to their own proprietary codes. But there's a lot of room for improvement when it comes to describing books on the open web – on publisher websites, on author websites, at reader destinations and online book groups. Here, we don't see as much attention to the potential that vocabularies like Schema.org provide.

And then there's the book itself. As books are published digitally, there is no reason why they can't be marked up with Schema.org tags as well. The EPUB specification allows for this, and the merging of IDPF (who maintains the EPUB standard) with W3C is a glimmer into the potential that web developers see in ebooks. That metadata can be harvested and used in discovery efforts – not just from book to book, but WITHIN books.

We're already seeing some of that with efforts like Amazon's X-Ray, or with

Booklamp, which was acquired by Apple a few years ago. These are services that data-mine books, essentially. For that to be effective, there need to be standards, controlled vocabularies. It's happening. The question is whether or not we, as "book people" want to be at the table developing those standards or not.

If you want to have a say in how these developments are going to happen, you have to show up at the meetings. At the rate at which technology is evolving, we can't afford stasis. There are mobile technologies to consider. There are AI and VR potentials to think about.

Words mean things. And we see every day how words can be distorted, their meaning depleted or turned around completely. As book people, we need to exercise our bookishness and think seriously about the words we use to describe our books.

CHAPTER THREE

True Names Are Numbers

We think of identifiers in the book supply chain as being limited to the ISBN. But, in fact, publishers, distributors, retailers and libraries encounter all kinds of products – not just books – and have to grapple with those identifiers as well. In this chapter, we'll have a look at these numbers, and talk about how they're used, what they're used for, and I'll give as much history as I'm able. These identifiers are all in the same family – the International Standard Organization's TC 46/SC 9 subcommittee of informational identifiers. So they're similar, but not identical.

ISBN

It all begins with the ISBN. We covered its history in the first chapter, but there's more to say about it!

The ISBN is not just a "dumb number" – the digits actually mean something. The first three – always 978 or 979 – designate the product as a book. Most EAN bar codes begin with a prefix that designates the geographical location in which they are sold – in the case of books, "Bookland" is a fictitious country created so that the global book industry could continue to use the ISBN as the basis for its EAN bar codes.

The second group of digits indicates the country/language in which the book is published. US and UK titles generally have a group of 0 or 1, which indicates that they are published in English.

The third set of digits is the publisher prefix. Larger publishers get smaller prefixes, and vice versa. However, the publisher prefix begins to lose its meaning when publishers merge, split, or acquire one another – the ISBNs for their products remain the same, even if the publisher's name changes.

The next group is the actual number assigned to the title itself. So in the example above, you can see that this is the 766th book that the publisher has published.

The final number is a checksum that is based on a formula applied to the previous digits. This ensures that the ISBN is a valid number.

There is no such thing as an "e-ISBN". An ISBN is an ISBN is an ISBN – and it identifies both physical and digital books. Any organization using this term is incorrect. Years ago, a start-up website tried to fool naive publishers into purchasing "e-ISBNs" that were in no way related to the ISO standard. This posed serious data integrity issues at online booksellers. That start-up no longer exists, but unfortunately the term has stuck around, causing confusion.

ISSN

The ISSN identifies "serial publications" – journals, magazines, annual conference proceedings. Developed on the heels of the ISBN, the ISSN was drafted in 1971 and published by ISO in 1975. It was clear from the ISBN's success that numbering publications was a good idea.

The ISSN, unlike the ISBN, is a dumb number. The digits carry no intrinsic meaning; there are no prefixes or groups. The ISSN is an eight-digit number – seven digits and a checksum.

Now, to sow confusion: There *is* an e-ISSN.

As you might expect, it identifies the digital version of the journal. There is,

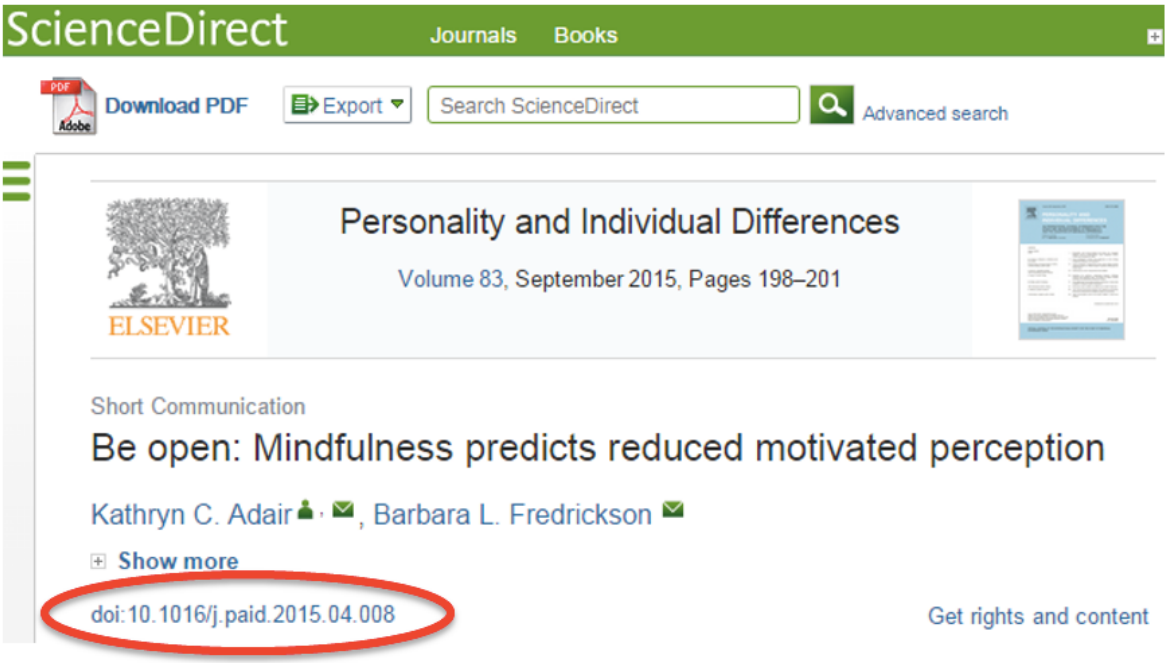
therefore, a p-ISSN, identifying the print version. Linking the two is the ISSN-L, which is assigned to whichever version is published first.

DOI

The DOI is not in fact an identifier of digital objects. It's a digital identifier of objects – DOIs can be assigned to physical items. But they are most frequently assigned to digitally-published journal articles.

The DOI was created in the 1990s, as the Web began growing in scale. Before the DOI, journal articles were identified by their URLs. But URLs change, or go defunct. Articles needed identifiers that would withstand redirection.

The DOI was developed using existing technology – the Handle system. This system was created by Bob Kahn, one of the inventors of the internet. Essentially, the Handle system stores a persistent identifier, as well as the location of where that thing can be found on the web. By decoupling the URL from the identifier, Kahn was able to provide a solution for when files get moved around on the web.



The DOI began its development in 1998, and was published by ISO in 2012. The DOI is, like the ISBN, not a dumb number. There is a prefix and a suffix, separated by a slash. Most of the time, the prefix begins with the number "10", followed by a period – this indicates that the identifier, while part of the Handle system, is specifically a DOI. After the period, there's a number indicating who registered the DOI (similar to a publisher prefix). Following the slash, the actual identifier of the article can be alphanumeric.

The main DOI registration agency is CrossRef[14], but there is also one for the entertainment industry called EIDR[15]. Most book publishers who are using DOIs register them with CrossRef, however - these are largely academic and scientific publishers who also publish journals.

ISRC/ISWC

The ISRC identifies a specific musical recording, regardless of the format in which it is issued. Thus the studio recording of Adele's "Hello" has the same ISRC regardless of whether it's issued on vinyl, MP3 or CD. The ISWC identifies the composition and lyrics.

ISRC is in use at some of the biggest distributors of music worldwide – Apple, Spotify, and Pandora all use ISRCs to identify music. While it was developed in the 1980s originally, it has proven to be quite useful over time.

The ISRC is a smart identifier. It's alphanumeric, with prefixes and suffixes – the country code, the code indicating who's registering the recording (usually the artist or label), the year it's being registered, and then the identification number itself.

COUNTRY CODE	REGISTRANT CODE	YEAR OF REFERENCE	DESIGNATION CODE
US	XXX	11	12345

ISAN

The International Standard Audiovisual Number was developed in the year 2000 and published by ISO in 2002. It applies to films, TV shows, video games, etc. ISANs are used by film and television studios, services such as iTunes and HBO, and technology companies like Microsoft.

The first twelve numbers of the ISAN form the "root" of the identifier. The root is assigned to the core work. The next set of numbers applies to the episode or part (if there is one – if not, the next four numbers are zeroes). The next character is a check character. The next four numbers identify the version of the work. The last number is also a checksum.

ISTC

The ISTC began development in the early 2000s as a way of collocating editions of textual works. It's not technically a "Work ID" for books, but was misperceived that way – in actuality, it's an identifier of text strings. So (broadly speaking) the hardcover, paperback and ebook edition of a book would all receive the same ISTC, but the French translation would not, because the text strings are different.

The ISTC is not necessarily assigned by a publisher, or a library, or a bookseller. It can be, but it doesn't have to be. There is no "ownership" of the ISTC like there is with other identifiers. Anybody who wants to register a textual work – whether it's an author, an agent, a publisher, or whoever - must submit a request to an ISTC registration agency with the necessary metadata needed to distinguish that work from others. The registration agency determines whether or not that request qualifies for a new or an existing ISTC.

This is one of several issues that prevented widespread adoption of the ISTC. Members of the book supply chain were uncomfortable with a registration agency determining the relationship of one book to another, for a variety of reasons. Publishers didn't want consumers being directed to competing editions of books (especially public domain titles, which are published by more than one publisher). Retailers wanted to be able to group books together in their own way. Agents were not willing to change their workflows to accommodate an additional step of registering for ISTCs.

There are currently 7 ISTC registration agencies in the world. Their mandates are not territorial – applicants can apply at any of them. But ISTC is not widely

in use in the book supply chain, and is an example of a standard that didn't get sufficient support from its intended users.

ISNI

The ISNI is the latest in our little family of identifiers. It was developed in 2010 and published by ISO in 2012. While we have identifiers for just about every form of intellectual property, the ISNI identifies the people and organizations that create that intellectual property. Authors, composers, musicians, actors, directors, producers, public figures, music labels, publishing companies – anybody who has created or contributed to the creation of intellectual property is eligible for an ISNI.

ISNI is particularly helpful in two cases – differentiating people with the same or similar names, and collocating people who have made multiple contributions across a variety of media. Large publishing houses, for example, frequently have similarly-named authors in their stables and royalty tracking can be made easier with an identifier like ISNI.

In addition to helping with rights tracking, the ISNI plays a central role in connecting various different sorts of systems together. In use by Wikipedia, Musicbrainz, the British Library, Harvard University, and a number of other organizations, ISNI provides a bridge between all these data sets, a way of linking all these different collections together with a common identifier. This enhances discoverability across the web.

A given registration agency doesn't do the actual assigning of numbers to names – that's done by the Assignment Agency, which reports to the ISNI-IA Board. The current Assignment Agency is OCLC. So they do the work of assigning the numbers to the names.

The registration agency works with clients to package files for bulk assignment. Registration agencies can also process individual applications.

Registration agencies form a liaison between clients and the Assignment Agency, so that the AA can focus on the the assignments, the algorithm that governs automated assignments, and other specialized issues.

Unlike ISBN, an ISNI registration agency is not bound by territory. There can be more than one per country. ISNI registration agencies tend to form around specific interests.

The Bibliothèque nationale de France, for example, is focused on French-related interests. Ringgold registers only organization names. The British Library is focused on UK-related contributors. Iconoclaste specializes in French-

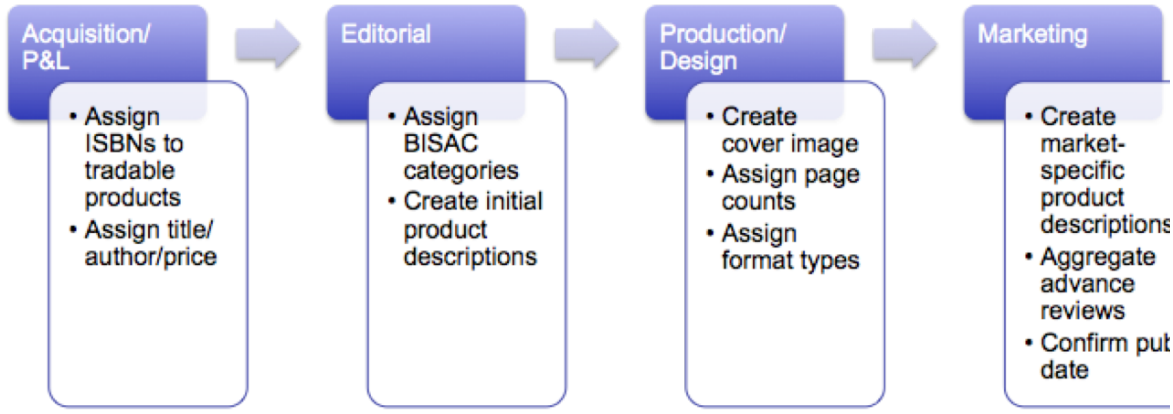
Canadian musicians. Numerical Gurus[16], the only US-based registration agency, is focused on creators of books, music, video and film, and other media and entertainment content. It is also the only registration agency to process individual ISNI applications.

CHAPTER FOUR

Into the Pipeline

E-commerce

In this chapter, we're going to take a look at how the book industry creates and distributes metadata.



Above is a functional map of the various divisions within a publishing house, and the metadata they create. ONIX, as we covered in the first chapter, stands for ONline Information eXchange. It's an XML format, a series of tags with specialized codes. It's a communication tool – you don't store data in ONIX. You send and receive data in ONIX.

In the US, we're still using version 2.1 – version 3 is ready to implement, and many organizations are ready to receive it or send it, but no one has made the first move to actually do so yet. As with so much in book publishing, ONIX 2.1 is "good enough", and the pain of moving to a new version seems to be greater than the pain of staying put.

ONIX tags are human-readable in long form. There are, however, abbreviations to the tags, which are easier for computers to process. ONIX consists of "composites", which are groupings of metadata – so in this example, the "Supply Detail" composite includes things like who's supplying the book, how to contact them, what rights they have to sell the book, the availability status, the pricing information, discounting information.

```
<SupplyDetail>
  <SupplierName>Gardners</SupplierName>
  <TelephoneNumber>+44 (0)1323 521777</TelephoneNumber>
  <FaxNumber>+44 (0)1323 521666</FaxNumber>
  <EmailAddress>sales@gardners.com</EmailAddress>
  <SupplierRole>03</SupplierRole>
  <SupplyToTerritory>WORLD</SupplyToTerritory>
  <AvailabilityCode>NP</AvailabilityCode>
  <ProductAvailability>10</ProductAvailability>
  <Price>
    <PriceTypeCode>01</PriceTypeCode>
    <PriceQualifier>05</PriceQualifier>
    <DiscountCoded>
      <DiscountCodeType>01</DiscountCodeType>
      <DiscountCode>ALBSB001</DiscountCode>
    </DiscountCoded>
    <PriceStatus>00</PriceStatus>
    <PriceAmount>7.99</PriceAmount>
    <CurrencyCode>GBP</CurrencyCode>
    <TaxRateCode1>S</TaxRateCode1>
    <TaxableAmount1>7.99</TaxableAmount1>
  </Price>
</SupplyDetail>
<SupplyDetail>
  <SupplierName>Amazon</SupplierName>
  <SupplierRole>08</SupplierRole>
  <SupplyToTerritory>WORLD</SupplyToTerritory>
  <AvailabilityCode>NP</AvailabilityCode>
  <ProductAvailability>10</ProductAvailability>
  <Price>
    <PriceTypeCode>01</PriceTypeCode>
    <PriceQualifier>05</PriceQualifier>
    <DiscountCoded>
      <DiscountCodeType>01</DiscountCodeType>
      <DiscountCode>ALBSB001</DiscountCode>
    </DiscountCoded>
    <PriceStatus>00</PriceStatus>
    <PriceAmount>7.99</PriceAmount>
```

Within the composites are frequently "codes" – so you see the "Product Availability" code in the example above is a number. That number corresponds to a particular status, and these codes are outlined in code lists that get added to (or deprecated from) by the various ONIX maintenance committees around the world. In the US, the code lists are maintained by the BISG Metadata committee.

Sound complicated? It can be. From the first day we began creating ONIX code lists, in 1999, there were terms no one could agree on. "Page count" – the physical count of the number of pages in a book? The NUMBERED pages in a book? What about forewords? What about indices? Do those count too? "Pub date" – is that the date the book hits the stores? Is it the date the book can be made available if a store chooses to? Is it the date the book is scheduled to be manufactured? What if it's actually a date RANGE?

Discussions like these are what it's like to sit on an ONIX metadata committee.

You don't have to use ONIX per se. Smaller publishers tend to use Excel templates, supplied by their trading partners, to communicate the same information (the information in the ONIX code lists) that larger publishers use the XML format to communicate with. Formats are important, but they are just the envelope. The codes are the information inside the envelope – your actual message. So there are really two factors you have to pay attention to – the *format* (if you are using XML, the file has to validate the way any XML file needs to, or your trading partner can't process it) and the *contents*.

At the Vendor's

What happens when it leaves the house?

The data gets ingested by various trading partners, on whatever schedule they have decided to adhere to. Proprietary data gets added. Not all the data you send gets used. Data points get mapped – as we discussed in Chapter Two. So what appears on any given screen will differ somewhat from what you've sent out.

There are so many different players in the metadata arena that can affect what a book record looks like. When you send your information to Bowker, they add proprietary categories, massage author and series names, add their own descriptions, append reviews from sources they license – and send out THAT information to retailers and libraries. The same thing happens at Ingram, at Baker & Taylor – so what appears on a book product page is a mishmash of data from a wide variety of sources, not just you.

At an online retailer, different data sources get ranked differently. This happens over time, as a result of relationships and familiarity with data quality, and these rankings can change. The data can also get ranked on a field-by-field basis. So a publisher might be the best source of data for title, author, categories, and cover image. But the distributor might be ranked higher for price and availability. And an aggregator might be ranked higher for things like series name – especially if they specify to the retailer that it's something they're focusing on standardizing and cleaning up. It's important to remember that in the eyes of the retailer, not all data feeds are equal. You'd think the publisher would be the best source of data about its own books but I can assure you, having worked with publisher data my entire 30-year career, that isn't always the case.

Updates are Disruptive

For a publishing house, updating old metadata records is a break from normal workflow, so it doesn't happen as often as it should for optimal

marketing purposes. It's important to remember, though, that the job doesn't stop once the book leaves the house – there are reviews, awards, and other events that are worth making stores and readers aware of through your metadata feed.

Just another quick word on terminology when it comes to updates – a "delta file" is what we call these updates – additions, changes, and deletes only, rather than a full file. Most publishers will send an initial full file, and then supplement with delta files for a time, and begin the cycle again just to make sure that their trading partners are in sync.

But on the retailer/aggregator end, there's no guarantee that your updates will get processed in a timely way (without a phone call). Companies ingest on their own schedule, and if they have a very heavy processing week, they might skip your delta file and wait for the next one, which means there might be gaps in data updates. This is why publishers find themselves occasionally sending a full file – just to be sure all their records are brought up to date.

Library Processes

I want to get into library data as well, because while it's not as sexy as retail, libraries do form a critical market that we often don't pay so much attention to.

There is no functional map.

Most publishers, unless libraries make up a significant portion of their customers, don't create their own MARC records. They leave that to intermediaries – library distributors (Baker & Taylor does a huge business creating MARC records), OCLC (whose member librarians catalog millions of items for WorldCat's database – in addition to the literal warehouse of cataloguers they have on staff), and the library customers themselves.

MARC stands for Machine Readable Catalog Record. It was developed in the 1960s by Harriet Avram at the Library of Congress, as they began using computers (with punch cards!) to store and transmit information about their holdings.

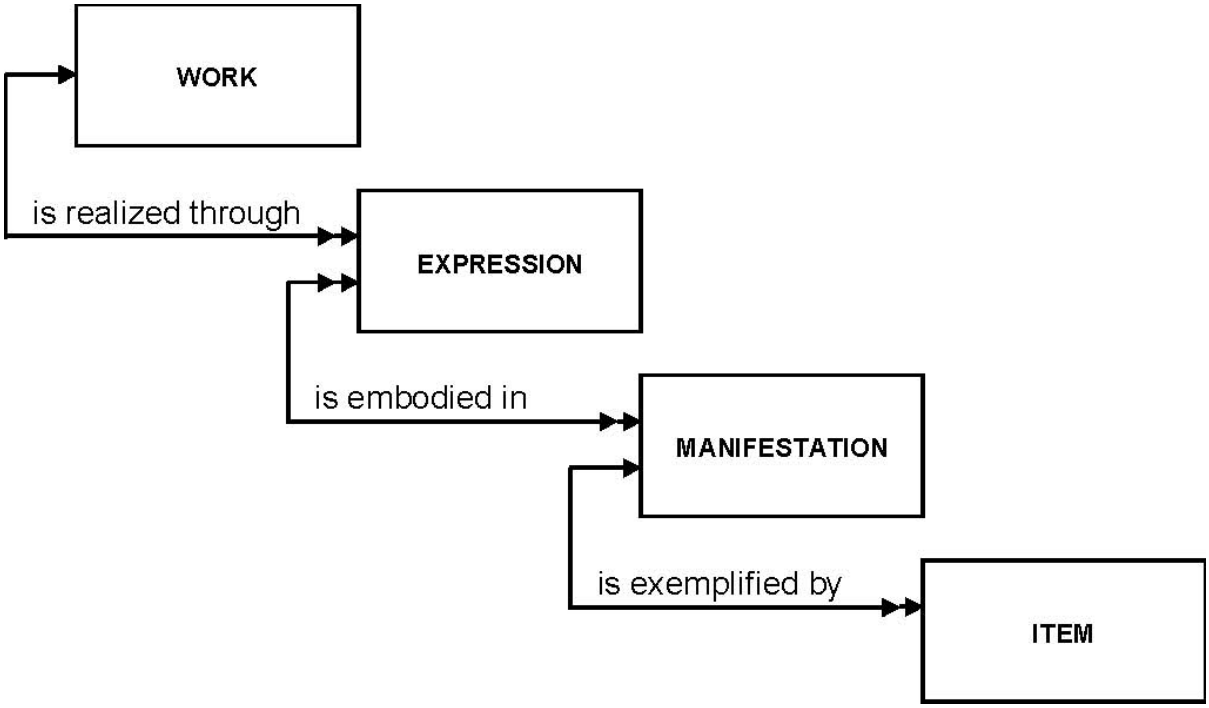
```
03465njm 2200589Ia 4500
001 ocm65198904
003 OCoLC
005 20060324022622.0
007 sd bsmenn-----
008 060324s197u ru opn rus d
028 12 $a C10-06739/3-1--C10-06740/3-1 $b Melodii˜a˜
035    $a (Sirsi) o65198904
040    $a NGU $c NGU
041 0  $d rus $d ita
049    $a NGUU
245 00 $a Poi˜u˜t solisty Bol'shogo teatra Soi˜u˜za SSR $h [sound recording] = $b Soloists of the Bolshoi Theatre.
246 31 $a Soloists of the Bolshoi Theatre
260    $a [Moscow] : $b Melodii˜a˜, $c [197u?]
300    $a 1 sound disc : $b analog, 33 1/3 rpm, stereo. ; $c 12 in.
500    $a Opera excerpts; sung in the original languages.
505 0  $a Side one: Eugene Onegin. Onegin's area / P. Tchaikovsky (Yuri Masurok, baritone) -- Boris Godunov. Marina's
       Aria / M. Mussorgsky (Yelena Obraztsova, mezzo-soprano) -- Queen of Spades. Hermann's Arioso (Vladimir
       Atlantov, tenor) -- The mermaid. The miller's aria / A. Dargomyzhsky (Alexander Vedernikov, bass) --The
       enchantress. Kuma's aria / P. Tchaikovsky (Tamara Milashkina, soprano) -- Ruslan and Lyudmila. Farlaff's Rondo /
       M. Glinka (Yevgeni Nesterenko, bass) --
505 8  $a Side two: The maid of Orleans. Jeanne's aria / P. Tchaikovsky (Irina Arkhipova, mezzo-soprano) -- Tosca. Aria
       of Cavaradosi / G. Puccini (Zurab Sotkilava, tenor) -- Sadko. Volkhova's lullaby / N. Rimski-Korsakov (Bela
       Rudenko, soprano) -- Demon. Demon's romance / A. Rubinstein (Alexander Ognivtsev, bass) --The Marriage of
       Figaro. Cherubino's aria / W. Mozart (Galina Borisova, mezzo-soprano) -- Iolantha. Roberto's arioso / P.
       Tchaikovsky (Yuri Gulyaev, baritone)
650  0 $a Operas $v Excerpts.
700 1  $a Mazurok, I˜U˜rïï. $4 prf
700 1  $a Obrazt˜s˜ova, Elena. $4 prf
700 1  $a Atlantov, Vladimir. $4 prf
700 1  $a Vedernikov, A. $4 prf
700 1  $a Milashkina, Tamara. $4 prf
700 1  $a Nesterenko, Evgeniï. $4 prf
700 1  $a Arkhipova, Irina, $d 1925-2010 $4 prf
700 1  $a P'i˜a˜vko, V. $4 prf
700 1  $a Sotkilava, Zurab. $4 prf
700 1  $a Rudenko, Bėla, $d 1933- $4 prf
700 1  $a Ognivt˜s˜ev, Aleksandr $4 prf
700 1  $a Borisova, Galina, $d 1941- $4 prf
```

As you can see from this sample, MARC is not exactly intuitive or human-readable. The fields are numbered, and cataloguers eventually get to memorize which numbers correspond to what should go in them. MARC is also not simply descriptive metadata for the purpose of "marketing" a book to a patron; it is an inventory system. The intent behind MARC is different from that behind ONIX.

Also unlike ONIX, MARC is not book-specific. Libraries catalog a wide array of materials in addition to books – films, software, music, websites, streaming content, physical objects like ebook readers and laptops, maps, musical scores, periodicals and individual articles. Some libraries use their cataloguing system for booking conference rooms and creating events.

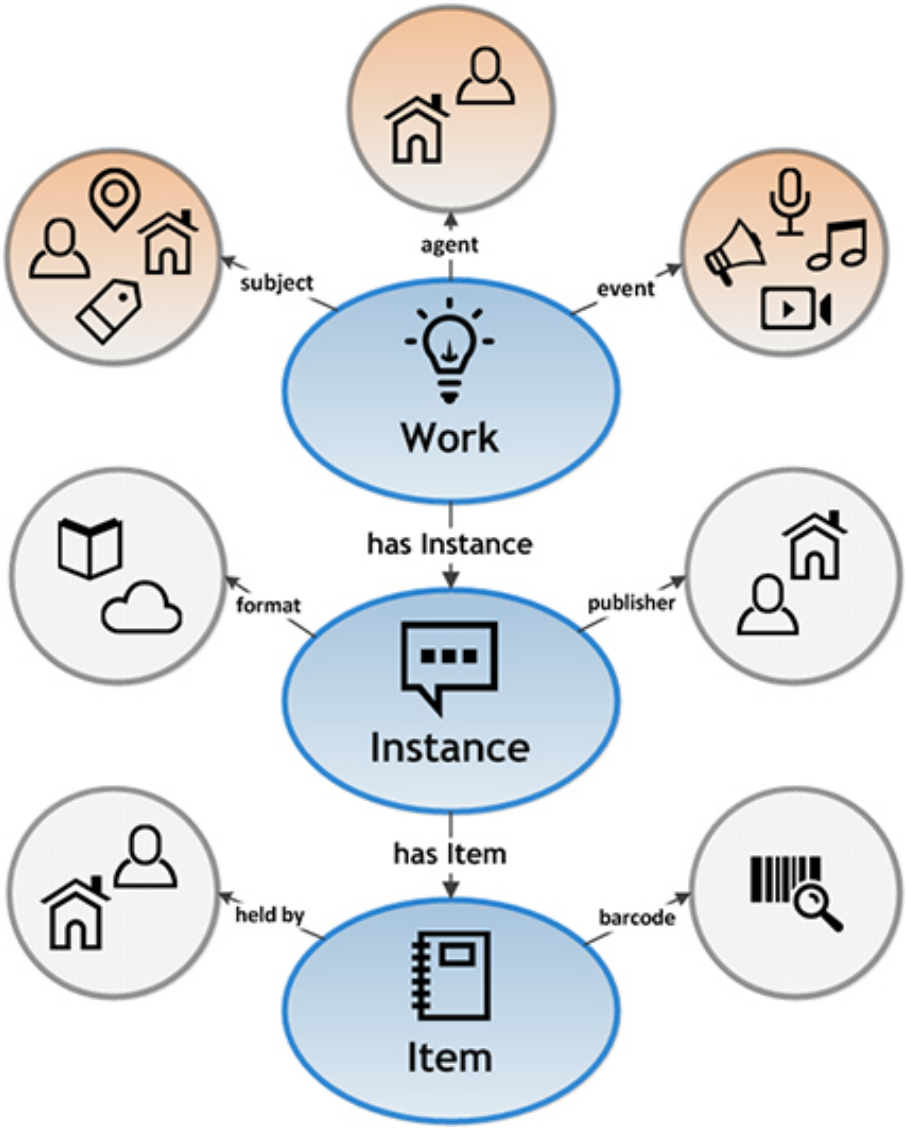
It has been revised and expanded (MARC21, MARCXML) but never satisfactorily replaced – global library infrastructure depends on it and, as with ONIX 3, the pain of transitioning to something new outweighs any constraints caused by a format based on 1960s computing standards.

You may also, in your travels, have heard of FRBR. It stands for Functional Requirements for Bibliographic Records, and was developed by a global library federation in the early 2000s. FRBR is a way of thinking about resources in a library, in a very hierarchical way. In the book world, a "work" would be something like the idea for "Alice in Wonderland", realized through the expression of that in the manuscript/original file as well as the Disney movie as well as the Tim Burton movie, which would then be embodied in various manifestations or editions (annotated, board book, streaming, DVD, etc.) The item is the actual thing a user accesses - the item you hold in your hand, or the file you're streaming. So to think about it in book terms, the ISBN gets assigned to the manifestation, and the item level is all the copies of that ISBN.



It's not perfect, and of course the jury is always out on what a "work" actually is, which is why it hasn't expanded beyond the library environment.

Given MARC's age, it's no surprise that the library community has been fishing around for more updated alternatives. BibFrame, or Bibliographic Framework, is a new-ish effort to come up with a way to describe resources using Linked Data principles. (It's helpful to note that in library-land, the term "bibliographic" refers not just to books, but all resources that might reside in or pass through a library and need to be described.) Library of Congress has just released version 2.0, but it has yet to be fully implemented – currently 13 library systems, globally, are experimenting with it. You'll notice some similarities to FRBR, and again, there are arguments as to how to define "work", but at least it is implement-able.



The book industry is not monolithic. It has pockets of quirk, changing rules, and polar opposite expectations from competing vendors.

Which means…you have to pick up the phone. Or email. Or whatever. Find out what your trading partners want from you. Do your best to provide that to them. There is no one-size-fits-all in this business, but it's a lot easier to manage that with digital communications than it used to be with paper catalogues and typewritten letters. ONIX was built to enhance communication. But it is not a perfect tool, nor are any of the others we've covered here. There's a lot of interstitial juggling and patchwork that we have to do in addition to standardized communications.

CHAPTER FIVE

Linking the World

In this chapter, we're going to take a look at the concept of Linked Data, and at the opportunity that we have, as an industry, to use existing metadata and identifiers to increase discovery for our books on the open web.

Search engines, as we mentioned previously, are heavily dependent on standardized metadata and identifiers - to the extend that Google, Yahoo, Yandex and Bing participated in creating the Schema.org standard. Search engines index and prioritize data that they find valuable - which drives page rank. Valuable data is both unique and authoritative - and authoritative identifiers like the ISBN and UPC are indicators of the uniqueness of a product.

When the web first became widely used, the <meta> tag helped search engines determine what a given website was about. But that tag got abused - spammers would embed keywords in the tag that had nothing to do with the content of the site. So search engines ignore that tag and focus instead on the RDF tags that Schema.org provides. These tags describe the elements on a website - reviews, people, products, businesses, recipes, events, music, etc.

Schema.org tags contribute to the formation of "snippets", which describe a web page in search results. They also contribute to the formation of the Google Knowledge Panel (or Graph, or Card), which calls attention to the product or person being searched for. Over time, the more attention a page gets, the more hits and links it gets, at which point the page ranking algorithm is affected.

From The Web To Books

In 1998, when I began working there, BN.com had 900,000 titles in its database for sale - representing the entire availability of books at that time.

Bowker reports there are over 28 *million* ISBNs in its database now. There are some caveats to this: some of these ISBNs don't represent viable products, some are assigned to chapters rather than whole books…however, there are also a sizeable number of books (via Smashwords, Kindle, and other platforms) that

never make it into Books in Print. We don't know if this evens things out, or if 28 million is the minimum number of books available in the US market today.

That's a 3000% increase.

Further complicating this scenario, we are living in a world where the content is born digitally. It can be produced and consumed rapidly, which is why there is so much of it, and why there is only going to be more of it. Lots and lots of information and entertainment. Lots and lots of, essentially, *data*.

Nothing ever goes away anymore.

Another factor is that the internet provides a persistence *even to physical objects*. With the web, nothing goes away, even physical objects – they only accumulate (on eBay, in vintage shops, and in libraries). They accumulate and accumulate. And books are very much a part of this accumulation. We don't order books out of paper catalogs anymore. We order books off the web.

And, in many cases, we order books that ARE websites – packaged into… ePub files.

We now have 28 million books to choose from. We also order music and movies over the web – and we do frequently don't see a physical medium for most of these things. Physical media get scratched, damaged, lost, borrowed and never returned. But digital is forever, and there's a freaking lot of it.

At some point (and remember, we're in a world of rapid development, explosion of content, and ever-more-sophisticated ways of consuming it – so "at some point" could actually be sooner than we think it ought to be), search engines and online catalogs go one step further than asking publishers (and other manufacturers) for product metadata in a separate (e.g. ONIX) feed. They are increasingly going to want to derive that metadata (and more detailed metadata) directly from the file representing that product itself. In our case, the EPUB file - a "website in a box".

This means that publishers are not only going to have to get good at creating and maintaining metadata at a pace that can sustain a 3000% increase over 18 years (so vastly more products to keep track of), they are going to have to get good at doing this *inside the book file itself*, which means not only grappling with markup language, but treating the EPUB file as a (really long) web page.

And at volumes that are unprecedented – because (a) publishing is easier than it has ever been before and (b) no book published now ever goes away.

This kind of rapid development doesn't just change your workflow – it

changes what and how you publish. And the more publishers understand about the web, the more likely they are to survive.

This is a different kind of survival than just holding on through a bad time, waiting out an economic downturn. This is survival that depends on evolution. On change. On new skills and abilities and ways of looking at things – while keeping in mind where we have come from and how we got here.

Let's go back to the problem of content proliferation. How are we going to manage it, organize it, feed the search engines in ways that they understand so that normal people who think Google is magic can actually find it, discern it, and read it?

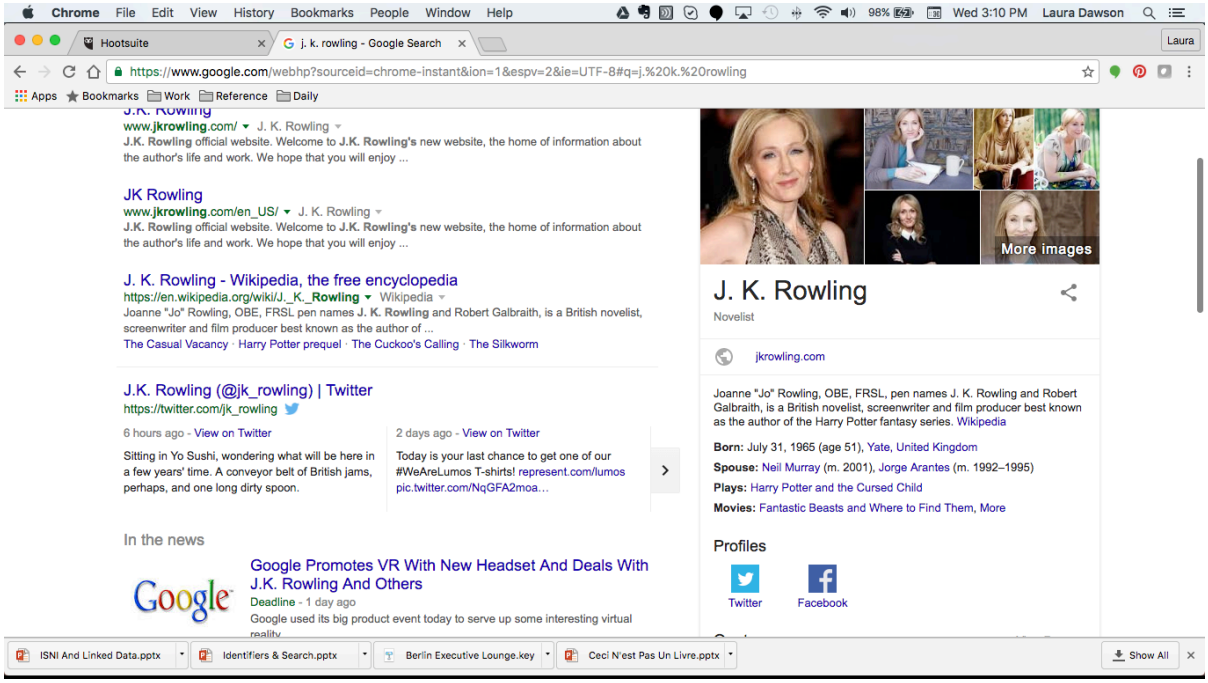
We impose a structure on it. We take that mess and organize the hell out of it. And yes, it has to be us - the book industry.

The search engine industry doesn't really care what results they display. Books are no more important to a search engine than anything else – it's *all* data. If we want to make the search engine work for us, we have to engage it. We have to understand how it searches, what's most effective on it. Just as the industry worked very hard in the 1990s to understand superstores and how they displayed books and what co-op could get us, so must we understand storefront of search.

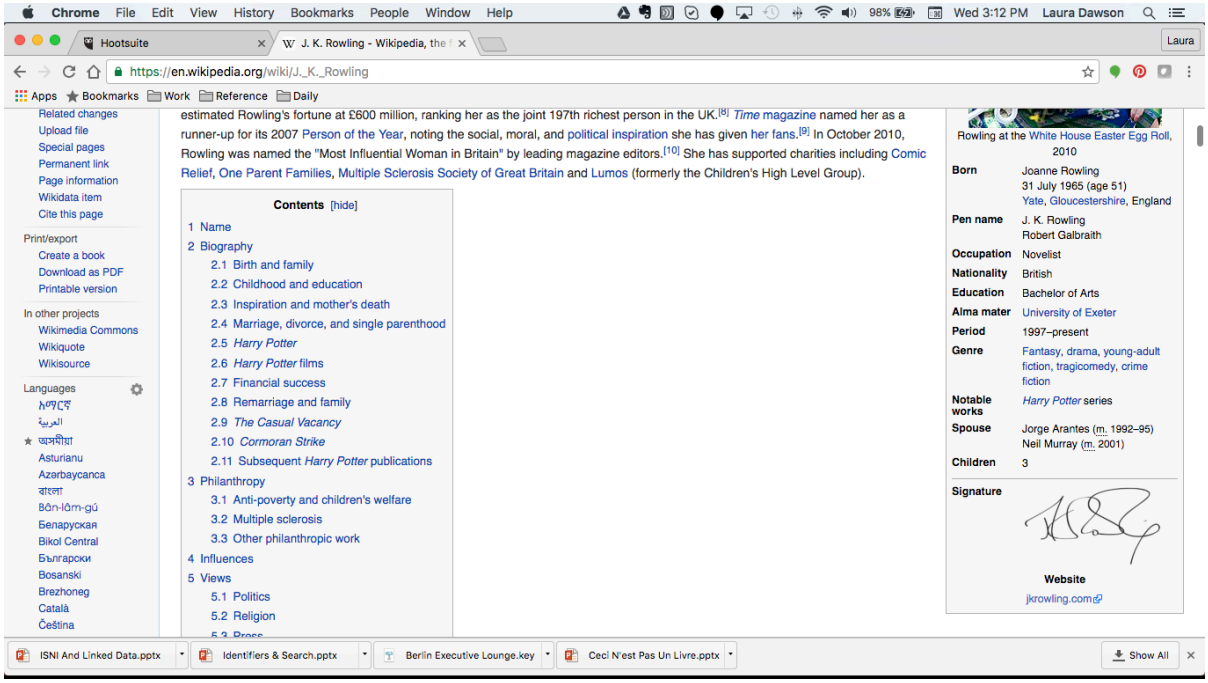
In an age of this much abundance, it's not enough to simply create a thing and then offer it for sale on the web. We have to understand how the market works. And the market revolves around search.

And Back to the Web

As we said above, Schema.org metadata is used by search engines to help prioritize and direct search. The tags get embedded in websites' HTML, the search engine crawlers look for it, and the search engine knows what to do with the content – how it's supposed to be organized, what it's about, what type of thing it is. And obviously those tags give the websites an added authoritativeness and uniqueness – attributes that all search engines prioritize extremely highly.



Let's look a little more closely at the knowledge panel for an author. There's a lot of information here – most of it derived from Wikipedia, which has incredibly structured data, and then turned into links for Google search results. Let's compare the knowledge graph here…



…with her knowledge graph in Wikipedia. All that information resides in a database (Wikidata) that Google crawls (with Wikipedia's permission) and harvests for use in their knowledge graph.

If your content is structured - tagged, identified - in standardized fashion, you

can support linked data environments such as the one between Google and Wikipedia. Graham Bell of EDItEUR and the BISG Identification Committee are working on creating an extension to Schema.org tags for book information that make use of elements from ONIX codelists. This will allow publishers to mark up their websites in a structured, standardized way so that Google can crawl them and use that information for the knowledge graph. Which is...like free advertising.

Thought Experiment

Let's think about the EPUB file as HTML "in a box". It is a website in a container.

From here on out you'll have to imagine what this would look like because… it doesn't exist yet. Although we are steadily working towards it!

Schema.org is currently working on representing books on the web, for discoverability. But the next natural step is to represent the *concepts* the books are talking about, and to link those to other concepts.

Basically, by opening the book – by shucking the container and making the content actionable on the web – books become APIs to other books. You can link to other books within that book. You can categorize ideas and characters, you can tag dates and locations – and if you are using an agreed-upon ontology to do this, you can build bridges to other books about the same things.

CHAPTER SIX

Putting Metadata To Work

From Systems To Metadata

There are many services out there that handle workflows. Some are comprehensive, like IngentaConnect and Klopotek. These cover every aspect of the publishing process, from title management to warehousing to metadata distribution. Some focus on specific parts of the publishing process – Firebrand focuses on title management and metadata extracts; Iptor includes modules for paper/print/binding and warehouse functionality; MetaComet focuses exclusively on royalty tracking. You may find yourself having to use portions of some put together – for example, Firebrand and MetaComet are able to integrate. Or you may find yourself using non-publishing-specific tools like SAP, which feed into systems like Firebrand or Klopotek.

Or, you may have your own in-house tools to manage workflows. Smaller publishers have made their businesses work on a series of spreadsheets stored in a central location, to which only a few people have access. Or a SQL database to handle title management with third-party tools for other functionalities.

All of which is to say that, in my experience, workflow management systems are somewhat Rube-Goldberg in nature, and there are usually systems talking to other systems. There's double-keying – entering the same information in multiple systems. And there's also a lot of sneaker-net.

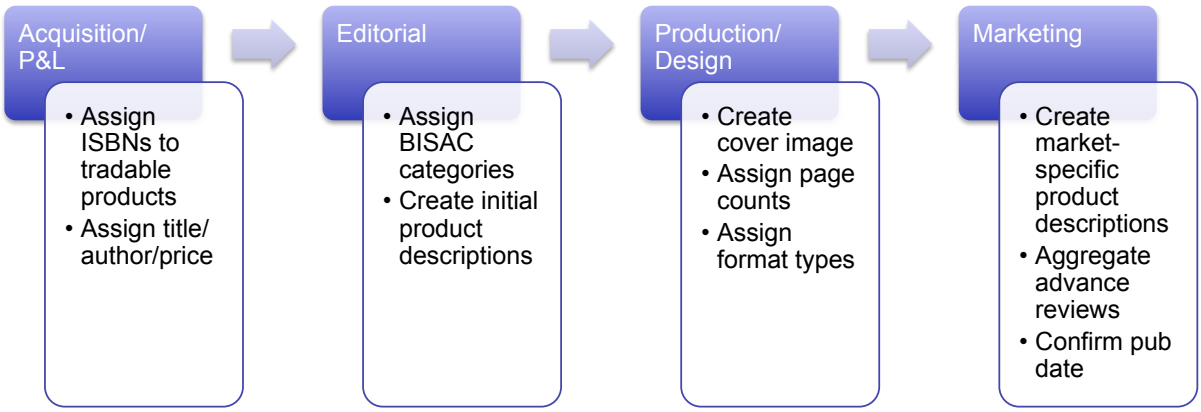
These systems' relationships to one another are complex. I've worked at McGraw-Hill, Bowker, Barnes & Noble, and consulted to many, many publishers and aggregators – and I have NEVER seen a smoothly running set of interoperating systems where everything worked elegantly and produced perfect and timely metadata with a minimum of effort. One or two components may be problem-free, but as a whole, there are ghosts in our machines.

And that affects workflow, of course. Certain jobs can only run at night, which means real-time data isn't available. Your warehouse data runs on an open source platform which isn't sufficiently supported. Your sales staff keeps

entering endorsements in the reviews field because their system doesn't have an endorsements field. Your company has acquired another company and the systems need to merge. Your digital asset management system is literally a box of CDs.

So there are lots of points where these systems don't align perfectly, and that is going to affect the quality of output.

Let's go back to our functional map for a moment.



It looks very neat. And, roughly speaking, it reflects most metadata workflow at most publishing houses. But it's not ironclad. And there's not necessarily a definitive end to the process – the metadata record might not leave the house after marketing gets through with it, because someone's still waiting for a cover image revision or a page count decision. Publishing processes can vary from imprint to imprint, from series to series, even from book to book. When you think about it, our job as publishers is to distinguish each book from all the other ones out there. That means innovations in marketing, in production, in subject matter – all of which affect metadata workflow.

Let's think about that a little bit. How is your workflow compromised by or helped by your systems?

One way to begin to tackle this is to structure your system so that there's a single repository everyone can tap into. Fran Toolan at Firebrand calls it the "Single Source of Truth" – having a central repository means you don't have competing spreadsheets on people's hard drives, or questions about whether you've got the latest version of information about a book.

It looks very neat, just like our functional map does. And it takes us further down the road than any other approach. But of course that's on the input side – on the OUTPUT side, you still have to deal with bespoke ONIX, competing requirements from vendors, and all the other frustrations that we've already discussed.

So let's think about that a bit – what are some of the issues around sending OUT metadata?

Relationships

We have systems, which can be outdated, chaotic, and incompatible. We have a functional process – who does what when – which is riddled with exceptions. And we have a single source of truth that holds all the most updated information, but has to export that information in a lot of different ways.

Relationships are what get us through the gaps in tech, staffing, and standards.

This industry was built on relationships: Agents knowing the tastes of editors. Marketers knowing the tastes of bookstore buyers. Editors knowing what works for library acquisitions. Human relationships – our work friends – keep this industry in business in the face of imperfect systems. And that's not appealing to some folks – which is also why we have a thriving indie scene with relationships of its own – some see these relationships as forming a gate-keeping system, a sense of exclusivity.

But for the most part, the relationships are there to smooth the bumps in the workflow. Sudden price change? Call up your vendor partner at B&N. Want to test out an idea for a project? Reach out to the collections development librarian to see if she thinks it's viable. Need to recall a book or cancel a publication, and the ONIX feed may or may not get ingested? Call up your trading partners and they'll do a manual update to their records.

Our relationships are the hack that makes it all work. The relationship between publisher and bookseller existed before there were systems to facilitate information exchange.

So, as with any relationship, communication is the highest priority. Much of that communication is automated with metadata transmissions – but as with anything automated, there are going to be exceptions that have to be dealt with manually. Which means an email or a phone call, or an in-person chat at a conference to gain or give clarification about what's expected. Maybe the answers aren't what we want to hear, but at least we know how to solve the problem.

I'd like to say that the one thing we can all count on is that everybody in the business wants it to work. And in most cases that's true.

You do have the "digital disruptors" – Apple, selling devices; Google, selling ads; Amazon, selling everything – who don't necessarily have the book industry's best interests at heart. And that's where we're experiencing the greatest friction. It's not necessarily with computer systems. It's with market systems – when our world is disrupted not by ebooks, but by (for example) a vendor using our product as a loss leader; another vendor who is uncommunicative and not at any conferences/standards meetings; another vendor whose founder didn't even want to sell books in the first place because "nobody reads anymore".

We can't NOT do business with these folks. But we are the little guys here, and that's a reality. This is WHY we have to deal with competing vendor requirements – books might be a small fraction of Amazon's business, but Amazon is a LARGE fraction of OUR business.

So how can we improve our relationships with these partners?

Well, on some level they do want things to work as well, or they wouldn't be selling our books. And maybe, in addition to trying to school Google in the wacky ways of book publishing, we need to learn from Google about the also wacky ways of tech.

The other thing to note is that these companies DO hire book people. The folks at Google Books are longtime publishing operatives – from HarperCollins, B&N, Scholastic and Harvard University Press. Amazon regularly hires out of the NYC book publishing pool. And the iBookstore is staffed by people from Simon & Schuster, B&N, Oxford University Press, and other traditional publishing companies. So the encouraging thing is…we are infiltrating. There really ARE like-minded people at these companies. They seem (and feel) like faceless organizations, but they are not. That's a myth that we've been telling ourselves for over 10 years.

Best Practices

What we'd consider best practices can vary depending on who's consuming the metadata.

BISG has published a document on best practices for product metadata in the book supply chain. That's a good foundation for bookselling – obviously the vendor requirements will differ as they compete with one another. Of course, Amazon doesn't participate in hammering out these best practices, so they're not perfect. But it's a good guideline to begin with.

For non-bookselling purposes – libraries and other institutions – the idea of "best practice" gets a little murky. Resource Description and Access (RDA) was created by an international association of library organizations as a set of guidelines for cataloguing resources (books and other things). RDA can serve as a good foundation upon which to build library metadata.

Some systems vendors also do educational sessions where best practices are reviewed: Firebrand, Klopotek, Iptor and Ingenta all have user conferences or webinars for their clients that go over market needs for metadata and standards. And of course there are trade shows like ALA and BEA which also have conference tracks.

Basically, if you get out there and start talking – and you might not have to go places physically, if budget is an issue – you can get a conversation going. Maybe others are having similar problems to you. Maybe there's a solution someone knows about. To some degree, we compete, but as an industry we're also really good at helping each other.

CHAPTER SEVEN

Bibliography/Resources

Book: A Futurist's Manifesto - Hugh McGuire and Brian O'Leary - book.pressbooks.com

The Book Industry Study Group - www.bisg.org

The Invisible Web : Uncovering Information Sources Search Engines Can't See - https://www.amazon.com/Invisible-Web-Uncovering-Information-Sources/dp/091096551X

---

[1] https://en.wikipedia.org/wiki/Lists_of_languages

[2] http://www.legacy.com/obituaries/lohud/obituary.aspx?pid=159414836

[3] https://en.wikipedia.org/wiki/R.R._Bowker

[4] https://en.wikipedia.org/wiki/Gordon_Foster

[5] http://www.bookweb.org/news/13-digit-isbn-and-2005-sunrise-and-gtin-initiatives

[6] http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1804&context=atg

[7] http://www.edibasics.com/edi-resources/document-standards/

[8] http://bisg.org/page/TutorialFAQ

[9] http://www.publishersweekly.com/pw/print/20000124/37602-pw-aap-unveils-e-retail-guidelines.html

[10] http://fast.oclc.org/searchfast/

[11] http://dublincore.org/

[12] https://www.w3.org/RDF/

[13] https://en.wikipedia.org/wiki/Schema.org

[14] http://www.crossref.com

[15] http://www.eidr.org

[16] http://numericalgurus.com