The Metadata Survey

*a small controlled survey of Pearson’s thoughts on metadata*Madi Solomon, Director, Semantic Platforms & Metadata, Pearson plc  
May 28, 2014

**Executive Summary**

Metadata, data about data, has been a conversation piece in the publishing industry for years now but proving its usefulness to the businesses has remained elusive. While most people can speak of the importance of metadata and how it has been successfully monetized by the likes of Facebook, Amazon and Google, in local practices it is still perceived as a labour intensive manual effort with few redeemable benefits. Even with past efforts of applying minimal metadata, our assets remained un-discoverable, rights information are locked and difficult to obtain, and federated search provides scant returns.

This persistent reality has remained despite several enterprise deployments of different asset management systems over the past ten years. The systems were not at fault, but our process was. Many large publishing organisations have never had a global strategy. Investments in digital asset management and rights systems tended to concentrate in the US and UK leaving significant international regional offices such as Brazil, Australia, and Mexico, completely isolated, without tools or access.

This approach to asset and content management must change if we are to unite around a single suite of technologies that streamline global access, embed efficacy measures, and enable digital distributions across all devices and platforms.

In preparation for a Pearson Metadata & Taxonomy Roadmap, Madi Solomon conducted 12 anonymous interviews with education publishing representatives across many different Lines of Business (mostly from US, UK, and Canada). This report synthesizes the results of the survey.

**Sponsorship**

These interviews were commissioned by the Semantic Platforms & Metadata team of the Core Platforms and Enterprise Architecture (Pearson Technology) and the W3C Digital Publishing Interest Group, Metadata Taskforce, in the quest to answer the question “What are the metadata pain points for publishers as they evolve to digital distribution?”

**Goal**

The results of this survey inform two goals:

1. Results have been combined with another set of interviews that were conducted by the Co-Chairs (Madi Solomon – Pearson, Bill Kasdorff – Apex CoVantage) of the W3C Digital Publishing Interest Group, Metadata Taskforce, to provide a broad view of publishing challenges around metadata.

2. Inform the Metadata & Taxonomy Roadmap to be created by Ian Piper, Chief Enterprise Taxonomist and Madi Solomon, Semantic Platforms, to ensure business relevance across the enterprise.

**Methodology**

One-to-one half-hour interviews were conducted over a four week period in March-April 2014. The interviews proceeded as anonymous, casual and candid conversations on experiences or observations around metadata.

**Results**

When the results of the W3C Metadata Task Force were combined with those from this report, differences in metadata expectations between Trade and Education publishing surfaced. The majority of those interviewed by Bill Kasdorf were trade publishers while Pearson, still in its transformative state to digital, were more focused on modularising content. Some of these differences are exampled in the table below.

|  |  |
| --- | --- |
| Trade | Education Publishers |
| Trade publishers stated that metadata complexity, mostly with ONIX, was a challenge to their business | Not a single interviewee mentioned ONIX or any other industry standard. |
| Trade publishers lamented the many metadata vocabularies (ONIX, BISAC, PRISM, etc.) and the difficulty in keeping up-to-date on all of them. | One interviewee mentioned multiple metadata standards and vocabularies as an issue |
| “Few books are online anyway,” was a general response from Trade. Other than STEM journals and articles, traditional publishers considered books as whole products and rarely modularised or componentized content. Metadata was relegated to Title and Author and not much more. | This did not apply to any representative as modular education content, personalised content, learner outcomes and efficacy measures were all based on data. The ability for personalisation (by a student, teacher, or institution) was a top priority for education publishers. |
| ONIX vs. Subject Metadata was a common debate in Trade. ONIX was originally created for the supply chain (retailers, etc), primarily for physical books and has since been updated (ONIX 3,0) for eBooks. There was general resistance to ONIX 3.0 because publishers believed 2.1 was fine and what the supply chain demanded. Subject or descriptive metadata on the intellectual content of a book was not easily supported or embedded in ONIX. | Subject metadata was a key entry point for educational content and information such as Learning Outcomes/Objectives and Learner levels were considered essential. |
| Trade publishers recognized the need for Keywords for books, chapters and component discovery, but were not necessarily interested in a controlled vocabulary. Trade publishers were only just beginning to realise that search engines did not use Library Catalogues to find book titles. | Many interviewees stated that many controlled vocabularies were required to optimise discovery and to re-purpose existing content. |

**Education Publisher’s Results:** percentages refer to the number of respondents who mentioned these topics as top metadata priorities for their business.

**Governance - 100%**

Every respondent cited the lack of metadata governance and authority as a major issue in their daily interactions with metadata. There wasn’t an authority to help dictate metadata requirements or to help embed or impose standards across the workflow. Instead, the “right to refuse” remained steeped in the traditional business culture where editorial had the authority to reject anything. The right to reject was also scattered across the workflow.

QUOTES:

“We’re nowhere on this. Every publisher right now adds and amends metadata at will so there is no cohesive approach.”

“Governance is critical in ensuring that customers have a good experience with our platforms.”

“We need to standardise on format for dates, for example, and make everyone use the ISO standards. This would have a very positive impact overall.”

“Governance is a pain point. In the past, governance boards were formed by non-metadata experts. We should be adopting industry standards and *impact* should not be a decision choice.”

“There are no mechanisms, governance, or levers that apply metadata, so the businesses just remain frustrated.”

**Rights – 60%**

A majority of respondents stated that Rights information was one of the most important metadata issues facing publishers at large. Without trusting Rights data, the businesses would rather err on caution and re-create or re-commission content before re-using or re-purposing existing content. The lack of Rights information on content and assets was costing organisations a small fortune in duplication and litigation. A means for querying rights information from source content/titles to the many derivatives was a priority need. While rights information was available in internal rights systems, the difficulty and the long wait in getting requested information was a source of frustration to many.

QUOTES:

“Our legacy systems are intractable and need to be abandoned.”

“We are six months behind in rights clearances for digital delivery and this is a major bottleneck.”

“There really shouldn’t be a separation anymore between the technologies that handles rights information. We need a new global rights strategy.”

“U.S. Rights can be cleared in our rights system … but there is no consideration for data governance.”

“We’re getting better at this by getting getter rights interactions between people. Now we need to get our systems to interact.”

**Flow – 60%**

60% of the respondents think that the flow of metadata was seriously compromised. There were many opportunities for metadata to be inherited, but no measures or mechanisms were created to capture them in the content lifecycle. Metadata was an afterthought in most workflows. Automation existed only in scraping exercises long after the content or assets were created, leaving some guesswork to vital information such as rights or identifying source originals to the distributors. The onus of applying metadata, data clean-up, and format conversions were left to the receiving platforms. This could be ameliorated with a more holistic coordinated view.

QUOTES:

“Metadata doesn’t flow upstream! We should fix this whatever we do.”

“A hybrid approach would be ideal, where humans populate some fields with controlled vocabularies and the rest auto-populated or scraped.”

“There is a general confusion between metadata, the process and flows of metadata, and what and when it’s captured.”

“The current workflow is more bothered than helped by metadata.”

“We end up being in the conversion business rather than the content delivery business. There are glaring inefficiencies, if not outright broken components, in our workflow and our ways of working.”

“Metadata is currently a myth, it simply does not exist. So what is its value? Metadata’s value can only be measured by its application so we’re stuck in a Catch 22: there is no metadata so there is no value so there is no use case.”

**Lack of Skills – 50%**

In our current changing culture, modular content demands more disciplined data care, but the businesses have not caught up to this. In general, the businesses don’t have the knowledge or skills to design a metadata growth model and by default, expect it to be done by someone else. This option, while valid, has not been formalised in any way and consequently, no such entity exists (other than vendors). There were many recommendations for a centralised service to help with this (See Centralised Service - 40%). Overall, the businesses requested more help in defining the new rules of engagement.

QUOTES:

“We are not data specialists, and this is all about data.”

“There are missing links between the workflows and information.”

“We need more education in order to unlock the potential of metadata.”

“It’s a chicken and egg thing: how to innovate and still support key business functionality?”

**Lack of Authority – 50%**

Half of the respondents collectively wanted more authority around metadata. The businesses wanted to know who “owned” metadata. Rather than leaving it up to the businesses, or the editorial process, they requested a stronger authority that could better support and enforce the required standards and could extend this authority to influence technology measures that ensured compliance. This related directly to the Governance issues around metadata, which further substantiated the need for a centralised entity to fully manage, monitor, and govern metadata standards.

QUOTES:

“Consensus takes too much time.”

“There are no governing principles so we are capturing metadata but without a good story. By story, I mean we don’t know the worth of the effort.”

“There would be more acceptance if we had a stronger top-down mandate.”

“We need a central function of metadata, taxonomy, and vocabularies with an authority that manages it.”

“We need a balance between enforced standards through tools and extending standards for specific business needs.”

“Metadata was the responsibility of the engineers and developers of the system. The role of metadata should be shifted to a specialist.”

**Standards – 50%**

Metadata standards were cited as something that should be identified, imposed, and managed. This included all forms of metadata from technical, structural, descriptive, search optimisation, educational standards and curriculum, all the way to online delivery standards. The use of standards was the solution for keeping content fluid enough to be shared and distributed across multiple platforms and devices. These were not solely for content, however, but also applied to consumer and learner data. These standards were recognised as essential in realising goals for personalisation and recommendations.

QUOTES:

“We need to implement minimum metadata standards and vocabularies.”

“The problem is that terms are dictated by owners so there are conflicts between systems because they are not standardised.”

“Rather than working groups coming up with standards, we should use educational standards already in existence.”

“If I had a magic wand, I’d build collaborations between product and services. Product has been isolated and really, 75% of metadata should be relevant across all platforms. I’d normalize metadata.”

“No one is aware of what standards we’re supposed to use.”

“We need to make standards adoptable. Right now it’s too difficult to get people to change.”

“Fundamental change in our process is necessary.”

**The following topics were referenced by 40% of the respondents:**

* **Inconsistency**

Multiple interpretations of standards, multiple definitions of terms and regionalisms have complicated the field.

* **Lack of Incentives**

Organisations often lack incentives for the businesses to work towards a “greater good.” Metadata was consistently sacrificed to aggressive delivery dates when the effort of applying it offered little or no benefits outside of the program.

* **Need for Learning Objectives**

Learning Objectives/Outcomes were referenced as key metadata for ensuring compliance with specific curricula. This was considered as important as descriptive metadata.

* **Need for a centralised authority or entity**

Closely related to Governance and Authority, a little less than half the respondents wanted a centralised group to solve metadata problems and offer services to accelerate access to and ingestion of metadata, vocabularies, learning objectives, and format conversions.

* **More education and guidance**

Many of the interviewees wanted more educational resources around metadata. Recognising that publishing was fairly new to content/asset/product management based on metadata, they requested more guidelines on the subject.

**Conclusion**

Metadata touches many parts of the digital supply chain, yet a comprehensive approach to its application and its value has been poorly executed. Reasons for this is includes a long list of exhausted efforts with good intentions. The businesses, however, may now be ready to embrace changes to their traditional approaches to content creation and are particularly open to the prospect of data-driven workflows that extend to efficacy metrics and personalisation of learning objects