# Initial Report of the DPIG Metadata Task Force

Background

In order to assess the “pain points” with regard to metadata for publishers, the co-leaders of the task force, Madi Solomon of Pearson and Bill Kasdorf of Apex, conducted a number of interviews with publishers, service providers, and representatives from related organizations. The summaries of those interviews are available as an appendix of this document.

The interviewees were selected to provide insight from a variety of perspectives, and were individuals known to the interviewers to be knowledgeable and authoritative within their spheres. Ms. Solomon took a “vertical” approach, interviewing a broad range of individuals within Pearson, a large global educational publisher. Mr. Kasdorf took a “horizontal” approach, interviewing experts from diverse types of publishing (book, journal, magazine, and news) and representing diverse roles within the digital publishing ecosystem (publishers, metadata service providers, consultants, and representatives from other organizations that are addressing the issue of metadata in publishing).

The interview strategy was to conduct casual, open-ended interviews with a single individual[[1]](#footnote-1) without an agenda or a prepared set of questions. The reason for this strategy was to avoid steering the discussion in particular directions. Instead, in this initial phase, the goal was to elicit what each interviewer would perceive as the key issues and pain points with regard to metadata from their own point of view. Thus the interviews deliberately did not focus on the issue of what the W3C could do—and what changes could be made to the Open Web Platform to address them. Instead, the interviews stayed on the general level. Since many of the interviewees were not technical, framing the discussion in too technical a manner would have impeded the ability to obtain authentic responses. As expected, few of the interviewees felt able to identify specific “pain points” with regard to the OWP. They spoke instead of general issues of concern to them in their work. The hope was that with an understanding of these issues and pain points, the DPIG could then assess where the W3C and the OWP could potentially help address them—and could avoid addressing theoretical technical issues that might not in fact align with publishers’ priorities.

While the published interview reports cited above will provide the best understanding of both the common themes and diverse perspectives revealed by the interviews, this report attempts to summarize key observations and offer initial recommendations for subsequent strategies.

Primary Observations

If there is a single overarching lesson revealed by these interviews, it is that the issues with regard to metadata for publishers and their clients and partners differ significantly between publishing sectors.

* *Trade book publishers are primarily concerned with discovery:* providing metadata to the supply chain that will attract readers to their books and result in increased sales. It was observed that at the present time book content itself is not typically online; it is available digitally only as products (e.g., EPUB), and typically only at the title (book) level, and the metadata regarding those products (primarily in ONIX) is created, maintained, and disseminated independently of the products themselves.
* *Educational publishers see metadata primarily in the context of asset and content management:* the identification and characterization of granular components of content within a large repository of content in order to facilitate the creation of publications, the delivery of content in “chunks,” the repurposing of content to create new editions and new products, and the ability to guide students and teachers to highly targeted components of content. The ability to personalize content to individual students, the ability to associate learning objectives with arbitrarily granular components of content, and the ability to monitor and assess the use of content by students and the learning outcomes that result are all key metadata priorities for educational publishers.
* *Scholarly publishers see metadata mainly as a “solved problem”:* while none would assert that the current situation is perfect, the standards-based consensus in the scholarly publishing world—consisting of nearly universal participation in CrossRef and CCC (the Copyright Clearance Center), the ubiquitous use of the JATS/BITS XML model[[2]](#footnote-2) for markup and metadata, and the reliance upon the DOI as a persistent, actionable identifier (initially for journal articles but now increasingly for book chapters and components, reference content, conference proceedings, and other publications, as well as the data sets that support research)—has enabled the development a rich ecosystem of services and platforms that has made the Web the primary mode of publication, dissemination, and access for scholarly content. It has also led to the development of other standards—such as ORCID, the Open Researcher and Contributor ID, and FundRef, a system for making public the funders of research—that continually refine the sophistication and utility of metadata in the scholarly publishing world, solving what were previously significant pain points (e.g., disambiguating contributor identities, revealing potential conflicts of interest or reliably documenting the absence of such conflicts).
* *Magazine publishers and news organizations are currently focused on rights metadata:* the rapid shift to the Web as a source of news, entertainment, and information has created an urgent need to identify the owners of content and the rights associated with very granular components of content—both the rights that apply broadly to a particular unit of content (e.g., copyright) and the usage rights conveyed by the rightsholder to specific parties, in specific contexts, for specific purposes, and in specific modes. While there are well established metadata models in each of these areas—PRISM is a very rich metadata framework virtually universal in the magazine world, and standards such as IPTC photo and media metadata, rNews, newsML, and schema.org are widely used by news organizations—standards bodies in both of these areas (IDEAlliance for magazines and IPTC for news) are each actively working on developing and refining rights metadata models.

While all of these issues—discovery via subject metadata and other metadata characterizing content and products, management of content via metadata, development and participation of cross-publisher platforms and services via metadata, and the communication of rights via metadata—cross all sectors of publishing, it is clear from the interviews that the priorities in distinct sectors diverge significantly.

Another major theme heard in virtually all of the interviews was that metadata is “too complicated.” Book publishers, for example, recognize that ONIX is the standard way to communicate supply chain metadata; as such, it is an extremely rich, complex, and useful standard. Similarly, the BISAC standard is a rich vocabulary used in the US for subject classification; there are similar such standards in most other countries or regions, and also a new global standard, Thema. While publishers recognize the value of these standards, they often characterize them as “too hard”; yet when pressed for what an individual publisher needs to communicate (to the supply chain, or about the subjects of its books), they often wind up asking for more complexity. (E.g., a U.S. publisher may want to describe a book as being about “the Battle of the Bulge,” within the topic “World War I” which itself is in the category of military history; this can be done with BISAC but not with Thema.) The truth is that these systems are complex because what they are designed to do is complex. The desire for an “ONIX Lite” expressed by several interviewees may prove to be unrealistic, because a significantly simpler model would be significantly less expressive.

Another common theme was that in too many cases metadata may exist—or may potentially exist, if applied to a given publication—but it often “doesn’t do anything.” It is very frustrating to users if it is the case—or even if it is their perception—that going the work to adding metadata is futile because systems are not seen as using it. (This is of course true of some types of metadata but not others: clearly trade publishers know how their ONIX metadata is used by the supply chain, and scholarly publishers know how their CrossRef metadata is used for citation linking.) This particularly surfaced in the context of the Pearson interviews because complex educational content is created by a vast team of participants, each of whom may have the ability to provide some aspect of metadata but most of whom have no clear understanding of how to do so, no systems to enable to do that consistently, and no faith that if they “go to all that work,” it will actually be used for any purpose downstream.

In thinking about metadata, it is important to distinguish between metadata that is incorporated within a publication (an EPUB, a website); metadata that is separate from the publication or publications it describes (e.g., ONIX, which can continually change over time without requiring the publications it communicates metadata about to be altered); and metadata that is incorporated in systems designed to provide information about publications (e.g., a publisher’s, retailer’s, or aggregator’s website).

And finally, it should be noted that an important theme that *did not* emerge from the interviews was the importance of accessibility. Revealing this was one of the benefits of the interview strategy of not asking leading questions: when anybody is asked if accessibility is an important issue, they will almost always say it is. So it is particularly—and lamentably—of note that none of the interviewees mentioned accessibility as a priority issue with regard to metadata.

Important Themes

The key themes of the interviews conducted by Mr. Kasdorf are summarized in the following appendix. They are the following:

* Complexity
* Inconsistency
* Sacrificing Richness for Simplicity
* ONIX vs. Subject Metadata
* Few Books are Online Anyway
* Discovery (AKA Marketing) is the Priority
* Identifiers, Identifiers, Identifiers
* And Now for Something Completely Different: News

Please see the [summary in the appendix](https://www.w3.org/dpub/IG/wiki/Task_Forces/Metadata#Phase_1_Strategy) for a discussion of these themes, including important comments by members of the DPIG.

The key themes of the interviews conducted by Ms. Solomon are summarized in the report provided at [TK]. They are the following:

* Governance
* Rights
* Flow
* Lack of Skills
* Lack of Authority
* Standards
* Inconsistency
* Lack of Incentives
* Need for Learning Objectives
* Need for a Centralised Authority or Entity
* More Education and Guidance

Please see Ms. Solomon’s report below [link to come] for a more detailed discussion of these themes.

Initial Recommendations

On the basis of these interviews and the themes and priorities that they revealed, the DPIG suggests the following with regard to how the W3C could/should or could not/ should not attempt to address these issues in the context of the Open Web Platform. These should be considered in the context of the distinction between metadata embedded in content, metadata embedded in web pages about the content, and metadata that is entirely separated from the content. [To be augmented and refined by the DPIG]

* *Make no attempt to develop standard vocabularies.* There are a great many vocabularies (controlled vocabularies, taxonomies, and ontologies) that exist through the publishing realm and the constituencies (readers, scholars, booksellers, libraries, teachers, students, scientists, etc.) that publishers serve. Useful vocabularies almost always need to be discipline-specific and context-dependent. While schema.org (not a W3C standard, but one closely related to the Open Web Platform) provides a valuable high-level mechanism for standardizing certain properties and in some cases controlled vocabularies associated with those properties, the W3C should not attempt to go further.
* *The focus of the W3C should be to enable publishers to incorporate and use the metadata that is specific to their publications and use cases.* While the markup and metadata models that are widely used in certain domains—e.g., JATS/BITS for scholarly and STM publishing, DocBook for technical publishing, and TEI for scholarship and humanities research—all provide extensive “metadata headers” as fundamental components of their schemas, no such metadata component explicitly exists within the OWP (although the OWP—through HTML5, microdata, and RDFa—does provide generic mechanisms for managing and conveying such metadata). The result of such an activity might result in recommendations and best practices by which publishers can use the existing components of the OWP to manage and convey metadata rather than the creation of new features of the OWP.
* *The need to associate subject metadata and other metadata characterizing content with arbitrarily granular units of content is a priority.* While, as mentioned above, the ability to do this may already be sufficiently provided within the OWP, the lack of implementations retards its use in this way by publishers. It is ironic that something that is seen as a quite urgent need by publishers does not seem to be seen as an urgent priority by the developers of reading systems and tools.
* *The ability to associate rights metadata with arbitrarily granular units of content is a priority.* This goes far beyond the typical binary view of rights. It involves the ability to provide interoperable information about permissions and prohibitions associated with a unit of content by certain parties in a given context; the ability to convey or provide access to contractual information associated with the use of content components between the rightsholder and the end user; the ability to dynamically modify the rights information over time and in a context-dependent manner; and ideally the ability to monitor compliance with or violations of the rights associated with content.

1. The one exception was Mr. Kasdorf’s interview with Len Vlahos, Executive Director, and Julie Morris, Director of Standards and Best Practices, both with BISG, the Book Industry Study Group. [↑](#footnote-ref-1)
2. JATS, the Journal Article Tag Suite, and BITS, the Book Interchange Tag Suite—which share a common markup model below the article and chapter level and which have very rich metadata models and mechanisms—are the current versions of what were previously known as the “NLM DTDs,” the markup and metadata model on which virtually all publications, platforms, and services in the area of scholarly publishing are based. This is unique to scholarly publishing: in no other sector is there such universal consensus on a single markup and metadata model. [↑](#footnote-ref-2)