Using HTML for Book, Journal, and Magazine Repositories and Workflows

BILL KASDORF

Vice President and Principal Consultant, Apex Content Solutions
General Editor, *The Columbia Guide to Digital Publishing*
Chair, BISG Content Structure Committee
Metadata Subgroup Lead, IDPF EPUB 3 Working Group

Participant's Perspective

In my consulting practice, I work with a wide variety of publishers and aggregators in the areas of trade, scholarly, educational, STM, reference, and magazine publishing. This work is primarily focused on the areas of XML modeling, vocabulary development, content management, and the design and implementation of XML-based workflows for multichannel publishing (print, online, and ebook).

I also devote significant time to standards development, education, and the development of best practices. I am an active participant in the IDPF EPUB 3 Working Group (I serve as Metadata Subgroup Lead); I chair the Content Structure Committee for the Book Industry Study Group, and am on the BISG Coordinating Council; I am a leader in the development of the IDEAlliance PRISM Source Vocabulary for magazines (I chair the "Packaging PSV as EPUB" committee); and I am an active speaker and writer on these topics. My bio is provided as an appendix to this position paper, but I highlight these activities here because they have given me both a deep and broad insight into the needs and practices of a variety of publishers.

Participant's Viewpoint

Over the past two years I have particularly emphasized the use of the Open Web Platform and HTML as a foundational infrastructure for publishing. Most publishers see HTML as simply one output from a repository and would never consider basing their workflows and infrastructure on it. However, I have successfully guided several leading publishing organizations—e.g., the World Bank, Harvard Business Publishing, the University of Toronto Press, and SAGE Publications—to do exactly that. In addition, I have worked with major players in educational publishing, most notably Pearson and CourseSmart, to develop next-generation models and workflows based on Open Web standards.

While this work has demonstrated the significant positive benefits of this strategy, it has also given me insight into the issues these publishers encounter in moving from more traditional models like DocBook, NLM/JATS, TEI, or proprietary models to HTML. I have become firmly convinced that the Open Web Platform has already evolved to a point that makes its use throughout an editorial, production, and distribution workflow possible for many publishers, and its expansion to more aspects of their workflows imperative for all publishers.

My participation in this W3C Workshop on Publishing Using the Open Web Platform will enable me to contribute these insights and bring a practical, real-world perspective to its work.

Examples of Concrete Observations and Suggestions

In the context of my participation in this Workshop, I would be able to contribute insights based on the following real-world experiences:

- The development of a rich and flexible vocabulary for the wide variety of publications produced by European Union Office of Publications.
- The development of an XHTML-based model for the World Bank that accommodates richly designed print-centric publications like the World Development Report while enabling a majority of their publications to be published only in online, ebook and POD formats. This model is particularly rich in both structural and semantic metadata, which is crucial in the context of the World Bank's mission to make this content discoverable and useful.
- The development of an end-to-end infrastructure and workflow for the University of Toronto Press for both books and journals that accommodates editing in Microsoft Word, Digital Asset Management in North Plains Telescope, print production in Adobe InDesign, automated generation of EPUB content documents, all based on a XHTML Hub model designed for transformation to HTML, NLM/JATS/BITS, TEI, DocBook, or other models.
- The experience of designing an XHTML-based infrastructure for Harvard Business Publishing that is flexible enough to enable automated conversion to well-structured XHTML and EPUB from Word files authored by hundreds of faculty that exhibit the inconsistencies and errors typical of relatively undisciplined Word-based authoring.
- The development of XHTML/EPUB 3-based specifications for CourseSmart, one of the world's largest aggregators of textbooks, that guide their clients to proper and yet practical production of reflowable, browser- and device-agnostic files for complex textbooks.

Key insights generated from these and other such projects that could lead to concrete improvements in the Open Web Platform for its use in publishing include:

- ***The need to accommodate very rich semantic and structural vocabularies while aligning with the inherent semantics in HTML5.*** The vocabulary for the University of Toronto Press required nearly 200 terms; the World Bank required nearly 300; and the EU Office of Publications required nearly 400.
- ***The need for flexibility.*** For example, it is typical of publishers like these to need to decouple heading levels from section structure; creation of "fake sections" is common when attempting to make them align, which results in incorrect structural modeling.
- ***The need for certain elements (e.g., headings, labels) to be both phrase-level and block-level.*** For example level 3 and 4 headings can often be run-in; while this is a presentational issue, it causes difficulties in the course of editing and production if they can't be treated as phrase level elements when desired.
- ***The need to include very rich metadata.*** Most of the models mentioned involve extensive metadata headers about the publication and accommodate the inclusion of metadata at an extremely granular level within the content documents.
- ***The need to be practical in layout-driven environments.*** Particularly in complex textbooks, it is difficult (though not impossible) for publishers to focus on structure and semantics and defer presentation. In magazines, it is virtually impossible.

Appendix: Bill Kasdorf Bio

Bill Kasdorf, General Editor of *The Columbia Guide to Digital Publishing*, is Vice President and principal consultant of Apex Content Solutions, a leading supplier of data conversion, editorial, production, and content enhancement services to publishers and other organizations worldwide.

Active in many standards initiatives, Bill serves on the IDPF Working Group developing the EPUB 3 standard (he was coordinator of its Metadata Subgroup and is now active in the Indexing Working Group); the IDEAlliance working group developing the nextPub PSV source format for magazines and other design- and feature-rich publications (chairing its Packaging PSV as EPUB Committee); he is Chair of the BISG Content Structure Committee; and he is a member of the Publishing Business STM/Scholarly Advisory Board and the NISO eBook SIG.

Past President of the Society for Scholarly Publishing (SSP) and recipient of SSP's Distinguished Service Award and the IDEAlliance/DEER Luminaire Award, Bill has led seminars, written articles, and spoken widely for publishing industry organizations such as SSP, O'Reilly TOC, NISO, BISG, IDPF, DBW, AAP, AAUP, ALPSP, STM, Seybold Seminars, and the Library of Congress. Most recently, he is the author of the chapter on EPUB metadata and packaging for O'Reilly's *EPUB 3 Best Practices*.

In his consulting practice, Bill has served clients globally, including large international publishers such as Pearson, Cengage, Wolters Kluwer, and Sage; scholarly presses and societies such as Harvard, MIT, Toronto, ASME, and IEEE; aggregators such as CourseSmart and netLibrary; and global publishing organizations such as the World Bank, the British Library, and the European Union.