Technology Standards Update

A Summary and Links to Resources STM E-PRODUCTION | LONDON, 4 DECEMBER 2014

Bill Kasdorf

Vice President, Apex Content Solutions

This document provides a brief overview and summary of the key technologies and standards discussed in the presentation "Collaboration & Convergence: How Today's Technology Standards are Working Together to Make Things Work" given at the STM eProduction Seminar in London on December 4, 2014.

The purpose is to help make STM publishers aware of important recent technology standards developments. Some of these are core STM standards and technologies; others are not considered core to STM but are directly or indirectly quite relevant to STM publishers. Also included are unreleased standards or technologies that are in development and deserve a degree of attention by STM publishers even though their adoption and implementation would be premature at this time.

An important theme is the extent to which standards are increasingly built on top of other standards, beginning to break down the siloed standards landscape of the past. Collaboration between standards organizations results in a convergence on fundamental technologies that makes the publishing ecosystem increasingly interoperable, reliable, and adaptable over time.

Web Standards

Today's digital publishing ecosystem is largely based on, or depends on, the group of standards developed and maintained by the W3C, the Worldwide Web Consortium (http://www.w3.org/) and referred to collectively as the Open Web Platform (OWP). Comprising over 100 standards, the OWP is the basis for markup, managing, transforming, rendering, and disseminating publications via the Web. Those most commonly used by STM publishers are the following.

XML

XML, the Extensible Markup Language, has become so entrenched in STM publishing that people may forget that this is indeed a W3C standard. It is the basis of much we do, from production to hosting to citation resolution and much more.

It is a good standard to begin with in the context of a standards update because of all the standards discussed here it is by far the most stable. Virtually all XML implementations are still based on XML 1.0, originally published as an official Recommendation in 1998—over 15 years ago. The current version is XML 1.0 (Fifth Edition), available at http://www.w3.org/TR/REC-xml/. This essential standard has remained so stable that the earliest implementations of it still work.[1]

[1] Full disclosure: there is in fact an XML Version 1.1; but it was soon realized that this was a misguided development, and except in very specialized cases, XML 1.1 is virtually never used. Stick with XML 1.0 (Fifth Edition), which was actually developed after XML 1.1 to address, in a bit of a clean-up (not a revision warranting an incremented version number), some of what was intended to be accomplished by XML 1.1.

HTML

HTML, the markup language of the Web, is just the opposite: it has evolved significantly over time. Originally very simplistic and presentation-oriented, it has evolved into a much more sophisticated vocabulary that stresses structure and semantics and relies on related technologies—primarily CSS, Cascading Style Sheets—to handle presentation.

The current version is HTML5 (http://www.w3.org/TR/html5/), which, after literally years of development and debate, was finalized as an official W3C Recommendation on October 28, 2014. But don't be daunted by how recent that date is. Because of the arduous W3C approval process, which requires implementations to be done in order for a specification to become final, HTML5 has become the fundamental markup scheme of the Open Web Platform. All modern browsers now handle HTML5, and a great many technologies (e.g., EPUB 3) are based on it.

Why is this so important? Because browsers and other web-based technologies based on HTML5 are required to *understand HTML5 semantics natively* and handle things properly. And although HTML5 does not require adherence to the strict rules of XML, it can be expressed as XML and thus benefit from XML's rigor. That's what's meant by XHTML, which in tech-speak is known as the XML serialization of HTML. That's what EPUB 3 requires.

MathML

Another W3C standard important to STM is MathML, the standard for markup of mathematics in XML. Its current version is MathML 3.0 2nd Edition (http://www.w3.org/TR/MathML3/), a recent (April 2014) update of MathML 3.0, which was finalized in 2010. Of particular note in this context is that MathML is a "built-in" complement to—and in a sense a part of—HTML5.

W3C Digital Publishing Interest Group

Although not a standard, the current work of the W3C's Digital Publishing Interest Group (DPIG) was discussed. Its mission is to identify aspects of the Open Web Platform that need to be addressed or improved for use in publishing. Like all W3C work, its work is public; the DPIG wiki is available at https://www.w3.org/dpub/IG/wiki/Main_Page. Of particular interest to STM may be the work of its Metadata, Annotations, and STEM Task Forces.

Standards People Think Are Web Standards, but Aren't

Some standards are so firmly embedded in Web technologies that they appear to be official W3C standards but are in fact governed by separate standards bodies.

Schema.org

Schema.org (http://schema.org/) is a collection of vocabularies and properties that enable semantic enrichment of content in a manner that is natively recognized by Web browsers and search engines (and can also be used by other non-Web technologies like JSON). It was developed by, and is governed by, the major search engines—a collaboration of Bing, Google, Yahoo! and Yandex. It's an excellent, and quite surprising, example of such fierce competitors working together for the common good. Although it is not a formal standard, it is so useful and becoming so ubiquitous (in that it "just works" in all the major search engines) that it has also been incorporated by reference in HTML5.

Schematron

A technology that has become increasingly important in the context of XML is Schematron (http://www.schematron.com/). It's invaluable for testing XML files against business rules and context-based requirements that parsing against a DTD or schema can't do. It is thus becoming an essential complement to XML parsers in quality control of XML. It is closely related to XSLT and based on XPath (both among the W3C Open Web Platform family of standards). While it is not actually a W3C standard, it is a very formal standard: ISO/IEC 19757-3:2006.

ODRL

Another standard that appears to be a W3C standard but isn't is ODRL, the Open Digital Rights Language. It's the creation not of an official W3C Working Group but a W3C Community Group. Whereas a Working Group is composed people from W3C member organizations and develops official W3C Recommendations, Community Groups are less formally structured, more open, and can produce standards that are useful but not "official" in the W3C sense.

ODRL is an example of one of these. It is very useful but not yet widely used except in certain sectors (for example, news; and it is being rapidly adopted for magazines). Given the suddenly compelling need for interoperable, machine-readable rights metadata—important to STM publishers as they make available, or use, content in a much more granular form over a much broader array of platforms and systems—ODRL may soon become much more widely used. Its latest version, Version 2.1, published as a Final Draft on November 10, 2014, is available at http://www.w3.org/ns/odrl/2/.

JSON

JSON, JavaScript Object Notation (http://www.json.org/) is an increasingly widely used standard that appears to be part of the Open Web Platform. A subset of JavaScript, the primary scripting language of the Web, it's a lightweight data interchange format, described in its documentation (http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf) as "a text format that facilitates structured data interchange between all programming languages." Governed by an international industry association known as Ecma, it was initially developed in 2001 but was formally adopted as ECMA 404 in October, 2013.

JSON is widely embraced among users of Web technologies because, in addition to its close relation to JavaScript, it is very simple to understand and use, both by humans and by machines. Another quote from that documentation is a clear indication of its appeal: "Because it is so simple, it is not expected that the JSON grammar will ever change."

One important potential use of it for STM publishers is for Linked Data as part of the evolving Semantic Web: "JSON-LD" stands for "JSON-Linked Data." It's one way, for example, in which RDF can be expressed on the Web (another is RDFa, RDF in Attributes, which is an extension of HTML5).

Although RDF (the Resource Description Framework, a fundamental component of the Semantic Web) was considered beyond the scope of this presentation, a very useful primer on RDF 1.1 can be found at http://www.w3.org/TR/rdf11-primer/ (updated June 2014) and one on RDFa 1.1 can be found at http://www.w3.org/TR/rdfa-primer/ (updated August 2013). Both are W3C standards and are considered part of the Open Web Platform.

Standards that are Built on Web Standards

Many of the standards considered essential to STM publishing are not themselves Web standards, but many of them are built on Web standards.

JATS/BITS

The XML model most fundamental to the STM digital ecosystem is JATS, the Journal Article Tag Set that is the successor to the ubiquitous "NLM XML" family of markup standards. Previously managed by the NLM, it is now an official NISO standard, ANSI/NISO Z39.96-2012. The previous "NLM Book DTD" has also been succeeded by a new model for books, BITS, the Book Interchange Tag Suite (not yet a NISO standard, currently maintained by NCBI).

BITS 1.0 was formally adopted in December 2013, at which time JATS was updated to version 1.1d1 to align with BITS 1.0. The important thing about this for STM publishers is that below the chapter level, BITS XML markup is virtually identical to JATS markup for journal articles. JATS version 1.1d1 is available at http://jats.nlm.nih.gov/1.1d1/ and BITS version 1.0 is available at http://jats.nlm.nih.gov/extensions/bits/.

While JATS and BITS may not appear to be "Web standards" (and they aren't, strictly speaking), what is important in this context is that they are both based on XML, the most fundamental of Web standards, and they incorporate other Web standards like MathML and the HTML table model.

EPUB

The most obvious example of an important digital publishing standard built on Web standards is EPUB. Its governing organization, the International Digital Publishing Forum (IDPF) states it best: "EPUB is the distribution and interchange format standard for digital publications and documents based on Web Standards." Although it is most commonly thought of as a standard "e-book" format, it is important to realize that EPUB is not just for books; it is for packaging and distributing any and all types of publications.

At its core, EPUB 3 (the current version, EPUB 3.0.1, finalized on June 26, 2014, is available at http://idpf.org/epub/301) is "packaged web content." Its content documents are purely HTML5, expressed as XML (XHTML5). It is firmly committed to maintaining alignment with HTML and Web standards as they evolve. As an example, one of the recent updates in EPUB 3.0.1 is to accommodate schema.org (expressed as microdata or RDFa), because that became an official part of HTML.

A stable but dynamic standard, EPUB continues to evolve. While the underlying specification is unlikely to make any current EPUB 3s obsolete, new features continue to be added. Upcoming additions of particular interest to STM publishers are EPUB Indexes 1.0 (its all-but-finalized specification is available at http://www.idpf.org/epub/idx/); EPUB Dictionaries and Glossaries (http://www.idpf.org/epub/dict/); and Open Annotations in EPUB (about to be put out for final balloting and available at http://www.idpf.org/epub/oa/).

The current status of these and other various EPUB working groups and standards is available at http://idpf.org/ongoing.

EDUPUB

An excellent example of cross-organizational collaboration is the development of EDUPUB, the profile of EPUB 3 for educational content. While the formal development of the EDUPUB specification is being done as an activity of the IDPF EPUB 3 Working Group, this activity was prompted by and is integrated with the work of a loose collaboration of organizations known as the EDUPUB Alliance.

Initially launched a year ago by IDPF, IMS Global (an organization governing many key educational standards), and the W3C, the EDUPUB Alliance set out *not to create a new standard*. Instead, the concept was to enable existing and widely used standards to become interoperable in a way that enhances all of them.

For example, IMS Global governs three important standards used in the interchange of educational content and data: QTI (Question and Test Interoperability—available at http://www.imsglobal.org/question/), LTI (Learning Tools Interoperability—available at http://www.imsglobal.org/toolsinteroperability2.cfm), and Caliper Analytics, available at http://www.imsglobal.org/caliper/index.html. As the EDUPUB profile of EPUB is being developed, it is being engineered to accommodate these standards, but it does not change those standards. IMS's "Using IMS Caliper Analytics™, Question and Test Interoperability™ and Learning Tools Interoperability™ with EPUB3™: EDUPUB Best Practices" is available at http://www.imsglobal.org/edupub/EPUB3QTILTICaliper_BestPracticesvd8.pdf.

The EDUPUB profile of EPUB is to be published in a complete implementable draft by the end of 2014. The latest draft, published by the IDPF on November 27, 2014, is available at http://www.idpf.org/epub/profiles/edu/spec/edupub-20141127.html.

EPUB-WEB

Perhaps the most exciting and visionary development in the context of the collaboration and convergence of standards is the recently announced concept of "EPUB-WEB." First broached as a presentation at the October 2014 Books in Browsers conference, it was published as a White Paper, jointly authored by Markus Gylling (CTO of the IDPF and the leader of all EPUB development) and Ivan Herman and Ralph Swick of the W3C, entitled "Advancing Portable Documents for the Open Web Platform: EPUB-WEB" in an explicitly "*Unofficial Draft"* on November 17, 2014 (https://dl.dropboxusercontent.com/u/1007541/epubweb-snapshot.html).

This White Paper sets out a long-range vision—expected to take years to accomplish—that will ultimately result in an EPUB and a website as being two distinct "states" of the same thing. That is, instead of there needing to be a given set of content and resources for a website, and an almost identical set of content and resources packaged as an EPUB for offline access and distribution, there will ideally be no difference between them. Opening up such a document on a browser over the Web would deliver a virtually identical experience to opening it up in a phone, tablet, or e-reader. Or on future platforms or devices not yet even dreamed of.

To those of us in the publishing technology space, this is Nirvana. In no way should you expect this vision to be realized anytime soon; in fact there is a chance it will never be realized at all. But it is realistic and concrete. The vision is solid and well articulated. There is no better example of the benefits from the trends of collaboration and convergence in the publishing technology standards landscape.