XML Publication Workflows for Standards

by Bruce Rosenblum

Introduction

Publication of standards is a vital part of the work of any standard development organization. Never has this mission been more important than in the fast-moving Internet era, when new or revised published standards need to be available in weeks, not the many months of the 20th century.

In the 21st century, not only speed but also new electronic features are critical. Standards must work well on a wide variety of devices and require support for a myriad of new formats (e.g. ePub, Kindle). They must have reflowable, richly hyperlinked text that is accessible to the visually impaired. Users need faceted, full-text searching to rapidly find *exactly* the information they require. In short, you need to offer your customer flexible content retrieval and presentation.

How can you meet all these new demands? The key is eXtensible Markup Language (XML). With XML, you can re-engineer your publication workflow to address all of these requirements. And, when implemented correctly, a publication workflow that incorporates full-text XML can actually cut publication time *and* costs.

XML Primer

XML,[1] based on SGML (ISO 8879),[2] is a World Wide Web Consortium (W3C) recommendation that has been widely adopted by technical and scientific publishers. It incorporates the Unicode[3] standard for representing text characters in all languages and MathML[4] for markup of mathematics.

To anyone who has looked at HTML, XML will seem familiar because XHTML is actually one "flavor" of XML. Whereas HTML is designed to format information for on-screen presentation, XML is meant to represent the semantic structure of a document, which can then be presented with a style sheet in a formatted view. Figure 1 shows the relationship between XHTML, with tags that indicate structure, and the formatting that displays once a style sheet has been applied.

XML goes farther than HTML because you can define your own tag names (that's the eXtensible part) and also your own business rules, such as creating special tags to mark terms and definitions in standards.

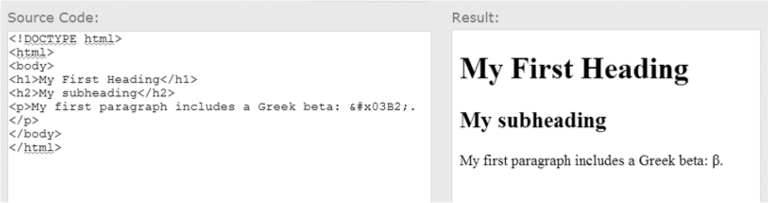
Introducing XML in a Publishing Workflow

Adding XML to a publication process introduces the question of how and when to introduce XML into the workflow. The author has worked with a wide range of publishers and has found that XML is generally created at one of these four points in the workflow. Each point has pros and cons:

- ***Have authors write publications in XML instead of Word***. Since SGML was created in the 1980s, it has been a dream of many to have large numbers of authors originate publications in SGML or XML. While this approach works well for small numbers of in-house authors, it presents much larger challenges when working with large numbers of outside authors. Equipping outside authors with XML-authoring tools is often impractical. More importantly most authors, while brilliant subject matter experts, are not XML experts, and they may not be able to reliably deliver high-quality XML documents.
- ***Convert Word documents to XML immediately upon receipt at the publisher***. This solution removes the XML authoring burden from authors, and it has the advantage that editors work in a continually validating XML environment. However, it requires retraining production staff to work in XML rather than in Word, and most editors prefer working in the comfortable Microsoft Word environment. An additional conversion from XML back to Word is also necessary at the end of the publication process if a Word document must be returned to authors for future editions.
- ***Convert Word documents to XML after editing, but before composition***. This solution requires in-house software to create XML from Word files and to create PDF files from XML, but it has many advantages. First, because the PDF is created from the XML, it is not necessary to re-proof XML files, as is required with post-publication XML conversion (see next bullet). Second, with automated composition from XML, editors can work in Word, regenerate XML at any time, and create a new PDF in minutes. With the tight Word → XML → PDF integration, the Word file contains the final edited content, and that file can be returned to the author(s) for future revisions.
- ***Convert final PDF files to XML after publication***. This solution is often seen as inexpensive and easy because the pre-existing workflow can continue undisturbed; just pay an outside (typically offshore) firm to convert PDF files to XML. However, this solution adds production time to the end of the workflow and can introduce quality problems because the resulting XML is rarely re-proofed with the same degree of care as the PDF from which it has been created.

ISO Case Study—XML Publication Workflow

Like many standards publishers, the International Organization for Standardization's (ISO's) workflow for the first decade of the 21st century focused on submission of Word files, ideally authored according to a standardized template.[5] Once received at ISO, the files were edited and often



```
<!DOCTYPE html>
<html>
<body>
<h1>My First Heading</h1>
<h2>My subheading</h2>
<p>My first paragraph includes a Greek beta: &#x03B2;.
</p>
</body>
</html>
```

My First Heading

My subheading

My first paragraph includes a Greek beta: β.

Figure 1: XHTML, a subset of XML on the left, specifies the structure of a document; the formatted view is shown on the right.

(Continued on page 16)

(Continued from page 15)

reformatted due to errors in the application of the template. When the content was finalized, a PDF was created from the Word file and published. Full-text XML was not part of the publication process.

By 2010, it was clear that this publication process was not sustainable. Increased security and compatibility issues in different regions of the world in the Windows and Word environment made it more difficult for committee members to install the ISO template on their PCs. Even when installed, template use varied widely by committee, often necessitating costly reformatting of documents by national standards bodies and the ISO production team. The production team found Word to be an excellent word-processing program but an often frustrating tool to use for "typesetting" standards in preparation for PDF creation. Publication times were very slow and measured in months. The final PDF publication was not an acceptable format for modern eReaders, and it had no hyperlinking.

In 2010, ISO began a new initiative to revise the publication process and incorporate full-text XML. ISO looked at possible workflow routes for XML creation and they settled on the third of the four workflows described above—to convert Word documents to XML after editing, but before composition.

This new workflow required licensing new in-house software to create XML from Word files and to create PDF files from XML. To implement the workflow, ISO licensed three customized off-the-shelf software (COTS) applications. ISO adopted eXtyles[6] for editorial cleanup and conversion of Word files to XML and Typefi[7] to automatically convert XML files to composed PDF pages in InDesign and to ePub. The resulting XML is stored in a MarkLogic[8] repository.

This workflow brought many advantages. First, typesetting was moved from Word to InDesign, eliminating much of the time and frustration in using Word for page layout, and effectively automating typesetting. Second, with automated composition from XML, editors could continue to work in Word, regenerate XML at any time, and have a new proof PDF in minutes. Finally, with tight Word → XML → PDF integration, the Word file actually contained the final edited content of the standard, and that file could easily be returned to the committee for the next round of revisions.

DTD Selection

As part of the incorporation of XML into their workflow, ISO needed a new Document Type Definition (DTD) to mark up full-text XML rather than just metadata. The ISO team decided to derive their DTD from an existing model rather than build a new DTD from scratch. The models they reviewed were the Text Encoding Initiative (TEI),[9] which ISO had used for some earlier projects; DocBook (an OASIS standard),[10] Darwin Information Typing Architecture (DITA, an OASIS standard),[11] and the Journal Article Tag Suite (JATS, a NISO standard).[12] All of these DTDs are freely available, have active community support, can be custom-modified to suit specific requirements, and have commercial support from tool vendors.

The team selected JATS as the foundation and then made some custom modifications to accommodate the requirements of standards publication, such as incorporating the TermBase eXchange (TBX) model for terms and definitions. ISO has made the DTD freely available, and it may be used by any organization.[13]

Back Catalog Conversion

The new workflow at ISO was designed for current content, but ISO also had a large library of previously published standards. ISO wanted to have these standards available as part of a complete repository of ISO content. It was decided to send previously published standards to an offshore vendor for conversion. An initial test conversion of 200 documents was carried out and a full-scale quality assurance (QA) process was set up for a successful conversion project. The production team at ISO built a rigorous QA process, supplemented with extensive written instructions for correct XML markup and software tools developed with Schematron (ISO/IEC 19757),[14] to automate a portion of the quality checking. The team found the same Schematron process could be used for both the back catalog conversion and the new production workflow as a way of checking that all XML markup requirements have been correctly applied.

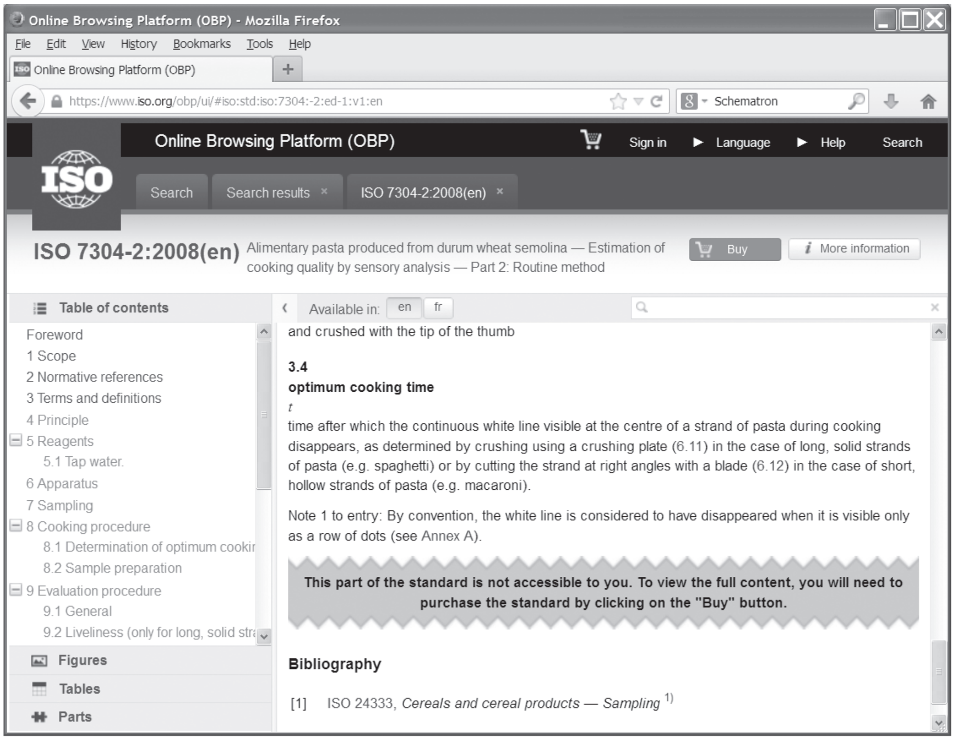
The Result

The new ISO production workflow went live in June 2012, and the back catalog conversion is scheduled as a two-year project in 2012 and 2013. The resulting full-text XML is loaded into a MarkLogic database and made accessible through the new ISO Online Browsing Platform.[15] The resulting online presentation of standards gives users access to the ISO standards library; this will then be integrated into ISO's online store. Figure 2 is a view into ISO 7304, found by searching for "pasta" in the ISO Online Browsing Platform. Users can freely read the scope, normative references, terms and definitions, and bibliography. The heart of the standard is available only upon purchase. This approach, implemented with full-text XML, allows purchasers to make a more informed decision. Standards in this view are richly hyperlinked both internally and externally; "6.11," "6.12," "Annex A," "ISO 24333" in the bibliography, footnote 1, and the table of contents are all active hyperlinks.

The new workflow at ISO built around full-text XML has been a successful catalyst for positive change in ISO's publishing operations. The codification of business rules as part of the XML production has allowed staff to begin to refocus on high-value content editing rather than formatting Word files for PDF production. Publication times are being reduced, with the ultimate goal being that no ISO standard will need to wait in a queue before being processed, and ISO will have achieved significant simplifications and savings in its publication operations.

Next Steps

The decision to go ahead with an XML workflow is fundamentally a business decision, not a technical decision. Before proceeding with an XML workflow, decide on the business goals you want to accomplish. If a high-quality ePub or a dynamic online presentations is among your objectives, then these goals can be best achieved with XML. Next, educate yourself and your organization about XML. Talk to other standards publishers who already work with XML. Learn about the challenges they faced in setting up their process and their best implementation practices. When you are ready to move ahead, consider hiring an in-house XML specialist who understands both technical


Online Browsing Platform (OBP) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Online Browsing Platform (OBP)
https://www.iso.org/obp/ui/#iso:std:iso:7304:-2:ed-1:v1:en
Schematron
Online Browsing Platform (OBP)
Sign in
Language
Help
Search
ISO
Search
Search results
ISO 7304-2:2008(en)
Buy
More information
Table of contents
Available in: en fr
Figures
Tables
Parts

ISO 7304-2:2008(en) Alimentary pasta produced from durum wheat semolina — Estimation of cooking quality by sensory analysis — Part 2: Routine method

Foreword
1 Scope
2 Normative references
3 Terms and definitions
4 Principle
5 Reagents
5.1 Tap water.
6 Apparatus
7 Sampling
8 Cooking procedure
8.1 Determination of optimum cookir
8.2 Sample preparation
9 Evaluation procedure
9.1 General
9.2 Liveliness (only for long, solid stra

and crushed with the tip of the thumb

3.4
optimum cooking time
t
time after which the continuous white line visible at the centre of a strand of pasta during cooking disappears, as determined by crushing using a crushing plate (6.11) in the case of long, solid strands of pasta (e.g. spaghetti) or by cutting the strand at right angles with a blade (6.12) in the case of short, hollow strands of pasta (e.g. macaroni).

Note 1 to entry: By convention, the white line is considered to have disappeared when it is visible only as a row of dots (see Annex A).

This part of the standard is not accessible to you. To view the full content, you will need to purchase the standard by clicking on the "Buy" button.

Bibliography

[1] ISO 24333, *Cereals and cereal products — Sampling* 1)

Figure 2: A view into ISO 7304.

and business requirements, or hiring an expert consulting firm to guide you through the process. Finally, design your new workflow based on business goals, not technical goals; run a pilot project; evaluate the results; fine tune the process; and then move into production. None of this will happen overnight, and all of it will require hard work. But the results can be a much improved publication process that can rapidly deliver the new products your stakeholders want.

Bruce Rosenblum, CEO of Inera, has thirty years of experience in design and development of electronic publishing solutions. He is a founding member of the JATS Working Group (ANSI/NISO Z39.96–2012), and he served on the NISO Board of Directors from 2005 to 2013. Inera is a vendor to ISO; however the opinions stated in this article are solely those of the author.

References

1 www.w3.org/TR/REC-xml/
2 www.sgmlsource.com/8879/
3 www.unicode.org/
4 www.w3.org/Math/
5 www.iso.org/iso/home/standards_development/resources-for-technical-work/iso_templates.htm
6 www.inera.com
7 www.typefi.com
8 www.marklogic.com
9 www.tei-c.org/index.xml
10 www.docbook.org/
11 www.oasis-open.org/committees/tc_home.php?wg_abbrev=dita
12 http://jats.nlm.nih.gov/
13 www.iso.org/schema/isosts/isosts-v1.1.zip
14 http://standards.iso.org/ittf/PubliclyAvailableStandards/c040833_ISO_IEC_19757–3_2006(E).zip
15 https://www.iso.org/obp/ui/

Words of Wisdom

"The modern age has a false sense of superiority because of the great mass of data at its disposal, but the valid criterion of distinction is rather the extent to which man knows how to form and master the material at his command" … Johann Wolfgang von Goethe (1810).