Getting to grips with XML

XML (eXtensible Mark-up Language) is rapidly becoming the format of choice for publishers to store and propagate their data. At Extenza we can handle the entire conversion process for our customers but often publishers wish to implement their own XML solutions for reasons of cost or control. This can seem like a daunting prospect for publishers but we can help. In this article we reveal how.

From the outset we at Extenza provide in-depth advice. Comprehensive documentation—from an introduction to XML through to detailed guides to the Extenza tagging scheme—is accompanied by full support from our production and technical teams. For automatic XML document creation and validation we provide a DTD (Document Type Definition, effectively a set of rules) and schema (more sophisticated definitions), and we have experience of several XML creation tools to help publishers find the right solution for their workflows. We can arrange office visits and advise on how best to capture the essentials of the original documents.

Editors do not have to worry about the integrity of their work, as the Extenza XML provides support for many formatting styles and is fully Unicode compliant throughout—permitting the safe storage and retrieval of all international characters and unusual punctuation. Unicode is a universal character set across platform, program and language. This means we can help a customer move to an internationalised workflow at the same time as they build their XML. The full text of a document, including images, tables, references and interlinking, can also be stored faithfully within the Extenza standard.

Once a customer has produced sample XML files, we subject this metadata to rigorous quality-assurance tests. This begins with automated checking at both the article and journal issue level using tools built in-house. Next, our production team will scour the XML and proofread the content, again with the aid of home-grown tools. This combination of human and automated checking ensures maximum accuracy. Finally, we have a staging environment, a copy of the Extenza site where we can load and test the customer's data. On this internal 'website' we can make the final checks to ensure the XML has been published according to our customers' exacting standards.

At all stages of the testing process we provide feedback on the XML. If a customer might benefit from it then we will actively suggest added functionality, like issue components or complex contributor-name breakdowns. Ultimately we make sure that everyone involved in the process has full understanding of the metadata storage: only then is a consistently high level of quality truly assured.

And if the customer still wants improvements in their metadata? Automated, 'clever' data tidying and normalisation is possible, and we can build integrated conversions of customers' existing metadata formats, including SGML and CSV. Ultimately, publishers need to know their data is safe in an XML environment, and that they can move from little or no knowledge of XML to an ability to develop their own metadata. As a customer of Extenza they can be sure of it.

Why use XML?

- XML is a 'smart' format, which provides information about the content, as well as how it should look.
- XML files are portable: they carry with them their own identification and can be interpreted by anyone with a copy of the DTD.
- Capturing content in XML facilitates searching by clearly distinguishing separate elements such as authors' first names and surnames, and specifying the style for these elements.
- An XML document may capture additional information not displayed in the content, such as subject codes, pricing and rights information, to allow the communication of cataloguing information to distributors; and can hold security coding to facilitate access controls.
- Content captured in XML is media neutral, which means that it is not dependent on the software that created it. It is not tied to a specific output device or system such as word processor, page make-up software, printer, e-book format or web page. The same XML file can be adapted for all these requirements.

John-Patrick Stacey is e-Publishing Assistant, Extenza e-Publishing Services.
jstacey@extenza.com

exxenza
e-publishing services