eXtyles: Interview with Elizabeth Blake and Bruce Rosenblum

Martin Fenner, Gobbledygook

May 1, 2009

Scientific papers are submitted to a journal as word processor files, usually in Microsoft Word format. After the paper is accepted for publication, the journal takes the manuscript and translates the text into a format that is better suited for publication online and/or in print. XML and the NLM DTD – a set of XML schema modules – have evolved as the standard data format for this purpose. Files in the NLM DTD format can in turn be translated into HTML and/or PDF for publication. The NLM DTD format is also used to transfer journal articles from publishers to archives (e.g. PubMed Central) and for long-term archiving.

eXtyles is a tool that facilitate the translation between these different document formats, and in the process also help to clean up broken references and other errors in the manscript. As paper authors usually don't see much of what happens to their manuscript after submission, I thought I'd ask Elizabeth Blake and Bruce Rosenblum from Inera (the company behind eXtyles) a few questions.

1. Can you describe what eXtyles is and does?

Elizabeth Blake: Inera offers several eXtyles products, all of which are designed to clean up, structure, validate, and export scholarly content. The desktop version of eXtyles is the most widely used product; it's a plug-in to Microsoft Word that is customized according to the editorial style and production requirements of each publisher that uses it. eXtyles:

- collects and exports manuscript metadata
- cleans up extraneous or incorrect document formatting
- applies structure to the document on the paragraph and character level (using author- and editor-friendly Word styles rather than visually intrusive tags)
- automatically enforces certain editorial style requirements through large-scale, context-sensitive find and replace
- parses, restructures, links, and corrects bibliographic references using internal templates and databases as well as external sources such as PubMed

and CrossRef
- links citations and callouts to their respective references and objects
- exports high-quality XML from Word according to the NLM DTD or any other DTD required by the publisher

eXtyles is a uniquely flexible tool in that it can accommodate any editorial style and the wide variety of workflows used by different publishers, and it can also process any Word-readable, author-submitted manuscript.

2. What is the difference between eXtyles and the Microsoft Word Article Authoring Add-in ?

Bruce Rosenblum: eXtyles makes two assumptions. First, eXtyles does not rely on the author to complete document structuring tasks – in our experience, journals have not succeeded in having authors provide sufficiently accurate styling or markup in Word files (the application used by most scholarly authors) to allow for automatic creation of high-quality XML. Second, eXtyles assumes that authors make other kinds of mistakes, whether adding inappropriate formatting, using unsupported fonts for special characters, or making informational errors in reference lists, and all of these problems must be corrected as part of the publishing process.

eXtyles is designed to overcome the limitations of how authors commonly prepare manuscripts by providing the tools that publishers need to clean up, rapidly edit, and then convert manuscripts to XML. eXtyles has been carefully developed over ten years to accurately address the reality of what publishers see in author submissions.

By contrast, the Microsoft Word Article Authoring Add-in has been developed as a content creation tool. It is designed for authors to structure articles in Word 2007 as they write. In this model, authors must add structural information rather than just submit the text of their article. The Add-in assumes authors will create reference lists with the Word 2007 citation manager and will adhere to all requirements necessary to successfully save the manuscript to NLM DTD XML when they have completed writing their article. Also, the Microsoft Add-in does not provide tools to prepare an article for XML conversion if the Add-in was not used during creation of the article, which is a key feature of eXtyles.

Fundamentally the target audience differs: the Microsoft Add-in assumes use by authors, whereas eXtyles provides tools to publishing personnel that allow authors to concentrate on great scientific research rather than the technical aspects of article publication.

3. What are the most common problems in submitted manuscripts that can be fixed by eXtyles? Are there problems that have to be fixed manually?

Elizabeth Blake: The most common problems range from relatively simple issues such as extraneous formatting or misapplied styles to more complex issues such as missing data or incorrect or uncited references. Hovering in between these are violations of editorial style – for example, British versus American spelling or non-standard abbreviations for units of measure – which eXtyles can also correct.

Some problems do have to be fixed manually. eXtyles does not take the place of an editor; rather, it automates as much of the low-level copy editing as can be reliably automated while drawing the editor's attention to issues that require follow up. An example would be eXtyles flagging a callout to a table that is missing from the manuscript; this obviously requires human intervention, but the warning saves time and flags problems early in the workflow when they are easier to resolve.

On a larger scale, the problem eXtyles is designed to solve is getting the accepted author manuscript published as quickly and accurately as possible. The fully eXtyled file can be flowed into a typesetting system such as InDesign with the paragraph styles aligned to the composition template, saving a lot of labor during the typesetting stage, or the Word file can be converted directly to NLM XML with the push of a button, allowing users to create rich, valid XML without any XML knowledge or expertise.

4. eXtyles helps with editorial tasks such as document cleanup and citation checking. Why shouldn't these tasks be left to the authors and checked when a manuscript is submitted?

Elizabeth Blake: Theoretically these tasks are the authors' responsibility! In practice it is an extremely rare manuscript that doesn't have errors, particularly reference errors. Our goal is not to discourage authors from submitting clean and accurate manuscripts; our goal is to facilitate the process when that doesn't happen, which is most of the time. The eXtyles reference-processing tools, in particular the tools that link and correct references with data retrieved from PubMed and CrossRef, go a step beyond what even a very thorough copy editor is typically able to flush out given the time constraints of a deadline-driven workflow.

As for performing these tasks at submission, the eXtyles integration with Editorial Manager does just that, providing an informative reference quality check early in the workflow so that responsibility for fixing the references can, if the publisher prefers, be pushed back on the authors.

5. If eXtyles validates and corrects references, why do most journals insist on bibliographies formatted in a specific house style?

Elizabeth Blake: There are two ways in which eXtyles corrects references: one is by correcting the data (e.g., PubMed reports that the first author in reference 2 is incorrect) and the other is by correcting the format according to the publisher's house style. Ensuring correct data in a reference, which eXtyles does with Automatic Reference Correction, is part of ensuring accuracy in the scientific record. As for the varieties of reference styles, many developed as necessary elements of print publication – for instance, some abbreviated house styles were devised as paper-saving measures. So long as print is not dead, these concerns are still relevant. However, in an electronic world, formatting of references may be less important, though a good editor will always have an argument for why their style is preferable (I say that as a former editor). We at Inera remain neutral on the topic since eXtyles can reformat references according to any preferred editorial style if a publisher requires it.

6. What is the NLM DTD? What was your part in developing it?

Bruce Rosenblum: The NLM DTD is a family of tag sets designed for full-text XML markup of scholarly articles. The intent of the NLM DTD Suite is to mark up and preserve the intellectual content of journals independent of the form in which the content is delivered. The suite can be used to publish and archive journal content, and to facilitate content interchange between organizations. And despite the NLM moniker, the suite was designed from day-one for full-text markup of journal content in any discipline.

Our work on the NLM DTD started in 2001 when the Harvard University Library, under a Mellon Foundation grant, was asked to study formats for long-term archiving of eJournal content. Harvard decided that PDF was not a viable archive format, and discovered that every publisher had a proprietary DTD. Harvard then approached Inera, because of our experience working with many DTDs including ISO 12083, to study the feasibility of developing a single DTD into which the content from all publishers could be converted for the purposes of long-term archiving. The e-Journal Archjive DTD Feasibility Study we wrote describes the requirements for such a DTD.

At the same time, NLM was making major revisions to the PubMed Central 1.0 DTD. NLM, Harvard, and Mellon decided to combine resources on a single project co-developed by NLM, Mulberry Technologies, and Inera. Version 1.0 of the NLM DTD was released in April 2003. The scope of use was originally focused on the needs of PubMed Central, and a long-term archive (now Portico) seed-funded by the Mellon Foundation.

The NLM DTD was quickly adopted by others when they discovered the coverage and flexibility. And as publishers switched from SGML to XML, they found that adopting an off-the-shelf public DTD was far less expensive than converting their existing SGML DTDs to XML. As use has expanded, support has grown

so that many off-the-shelf solutions are now available for working with the DTD. Inera continues to serve on the NLM DTD Advisory Board.

7. How does eXtyles use the NLM DTD?

Bruce Rosenblum: eXtyles users, with no knowledge of XML, can create high-quality XML according to the NLM DTD as a simple one-button action after using eXtyles to easily complete editorial preparation of a manuscript. In other words, eXtyles XML creation is a natural by-product of normal manuscript preparation for publication, and it requires no specialized user knowledge.

Because eXtyles was developed as an open platform that can be customized to the editorial and production needs of any publisher, it is internally DTD agnostic. In this regard, eXtyles can be used to convert content to any of the DTDs in the NLM DTD Suite (Journal Publishing, Archive and Interchange, or Book).

Furthermore, the NLM DTD suite actually allows for a wide range of interpretation, and eXtyles can easily meet any of those interpretations. For example, the NLM DTD supports two table models, XHTML and CALS. eXtyles can produce content in either table model, depending on the workflow requirements of the publisher. A wide range of other configuration options are available to meet the needs of specific publisher requirements.

8. Can the NLM DTD also used by authors to submit their papers to journals or repositories? If not, what are the limitations?

Bruce Rosenblum: In theory authors could submit their papers to journals or repositories in XML. However, in practice this has not occurred and we do not consider it likely to occur in the near future. There are a few key obstacles.

First, the primary job of researchers is to conduct research and report the results. Because technical knowledge of publishing formats is not a required part of the typical researcher's job, the majority of researchers do not submit papers to journals or repositories in XML when they can much more easily do so in PDF or Word.

By extension of this point, most publishers are interested in high-quality research. One publisher told us, "If a manuscript with great science is submitted on birch bark, we'll find a way to publish it." We believe that in a world where journals compete for the best papers, publishers will continue to put research quality ahead of submission formats.

Second, creation of high-quality XML is just not that simple. Even with the tools that have been developed in the past ten years, publishers with experienced editorial and production teams still have problems producing consistently high-quality XML; if you doubt this point, please look at a typical first round PubMed Central publisher validation report. If publishers with experienced production

teams do not find this easy, then we believe that production of consistently high-quality XML by authors is not likely to occur in the next few years.

10. What are your responsibilities at Inera?

Bruce Rosenblum: As CEO, I am involved in all aspects of Inera. My background is originally in software development, and I use this expertise to guide the development and quality assurance of eXtyles. I am also responsible for managing the business side of Inera. But probably my most important role is coordinating the incredibly creative team of people who work on eXtyles.

Elizabeth Blake: My role is primarily customer facing, from marketing to configuration and training up through support. My goal is to ensure that potential customers get all the information they need about eXtyles to make the right decision and to ensure that, once they become customers, they continue to be happy with that decision. We rely heavily on customer feedback when planning enhancements to eXtyles, and I tend to focus on the user experience when we work through the details of any new project.

11. What did you do before starting to work on eXtyles?

Bruce Rosenblum: Before eXtyles, Inera spent years providing SGML and XML consulting and software development services to publishers. Prior to that, I spent 15 years developing commercial software products for companies like Microsoft, Word Perfect, Houghton Mifflin, and Broderbund. Going way back, I've always had an interest in problems of working with text – my first professional software project was building a word processor for Chinese on an Apple II in 1980.

Elizabeth Blake: I started in scholarly publishing as a copy editor for the Cell Press journal Neuron and then worked as the managing editor of Neuron for several years before moving to The New England Journal of Medicine. At Inera, my years of real-world editorial and production experience have been very valuable and help to ensure that eXtyles development is centered on solving the problems that publishers face every day.