The application of Schematron schemas to word-processing documents

As traditional print-based publishing has made the transition into the digital age, a convention has developed in some quarters of capturing or even typesetting content using word‑processing applications.

These can present a convenient route to publication in the many instances where content derives (in the form of author manuscript) from the same word‑processing package. It is also a relatively cheap and efficient one, demanding the now basic and widespread skills of styling a document to achieve the desired appearance.

As a result, typesetting workflows consuming these documents still exist, template-based workflows designed to capture structured data are still in place, and for some publishers large quantities of legacy data persist in word‑processing formats only and require migration to XML to meet modern production demands.

During the long period (for some) of moving to a digital-first workflow, with publication of a single source of structured data in various renditions, it has become apparent to such publishers that the quality of their content no longer only resides in the appearance of the rendered product, but also in the quality of the data capture itself. The quality question has shifted from “Does my product look right?” to “Is my source markup sufficiently rich to service the outputs I wish to produce?” When generating XML markup from a word‑processing source, the inevitable corollary is whether the document has been styled appropriately to drive good-quality data capture.

The requirement to apply business rules to styled documents is not new. This was often done using macros to interrogate the underlying object model before Microsoft Office (OOXML)[1] and Open Office (ODF)[2] began exposing their respective file formats as XML. With the word-processing document being edited now available as XML, other, native-XML validation approaches are viable and indeed attractive.

This paper will present Schematron as a portable, standards-based alternative, demonstrating how it can be integrated into a word‑processing template to alert authors and editors directly to content problems during capture.

It will demonstrate how business rules can be applied to a word‑processing document held in one of the standard word-processing XML file formats using an ISO Schematron schema. These rules will comprise typical Schematron validation activity. They might be unexpected or missing style(s), co-occurrence constraints, or datatype errors; in fact, anything that Schematron might normally be used to identify in any other XML content. Further, it will be shown how errors found in the document and reported as SVRL can be successfully merged back *in situ* into the original document, so that an editor can address the problem so located within the originating editing environment.

Writing XPath-based Schematron rules for flat structures is reasonably tedious work: in document-based XML the structure broadly reflects the meaning of the content, whereas a word‑processing document is essentially a succession of (paragraph and character) styles, tables and other objects. Some work has been done to derive an XML-based schema for a document’s expected disposition of styles**[3]**.

The paper will further develop this theme by discussing the use and practicality of such a language to express the business rules governing a word‑processing document more declaratively, and how successfully this may be transformed to produce a Schematron schema which can then be used to perform the validation outlined above.

Finally, it will consider the feasibility and desirability of extending this approach to other kinds of office document, for instance to spreadsheets.

## References

[1] <http://www.ecma-international.org/publications/standards/Ecma-376.htm>. Retrieved 2015-03-08.

[2] <https://www.oasis-open.org/standards#opendocumentv1.2>. Retrieved 2015-03-08.

[3] Francis Cave, Francis Cave Digital Publishing: a style schema for word‑processing documents; personal communication, February 2015.