DELTAXML
EXPERTS IN INFORMATION CHANGE

White Paper

XML for Publishers – solving common challenges for Documentation Managers

Change is one of the dynamics of the publishing world – a consequence of industrial or scientific advance, of shifting fashions, of political whim or of the editor's red pencil. But the casualties of change should not be seen as a waste product, for the old can be as meaningful as the new and the difference, a thing of real value.

DeltaXML is a powerful tool for documentation managers that streamlines the management of change in the most complex documentation and unlocks its potential.

Management Summary

Structured documents have transformed the world of publishing. The majority of structured documents, in general use today, are based on XML. Instead of a simple sequence of words and objects, an XML document is now more like a database – containing additional information and instructions - and capable of an infinite number of outputs.

XML documents are now used almost everywhere – from bank statements to aircraft instruction manuals and from pharmaceuticals to education. For publishers, the flexibility that XML brings to the party is of huge value. They can now easily re-use content at will. They can slice it and dice it, customise and internationalise and create multiple products from a single source.

But the versatility of XML brings its own challenges. Structured documentation demands sophisticated management, not just to facilitate the creation and editing of documents but to manage the process of change.

Change is not just important in the creation or authoring process. Some documents are in a state of continual change that may last many years. Customers, regulators and stakeholders are often more interested in the history of successive changes that led to the final document than the detail of the current version. It is often the case that the amount of effort involved in the ongoing management of a document is far greater, and involves many more people, than was involved in the original creation. The situation for the document manager may be made more complicated in applications where the people and managers who edit or approve a document have no understanding of XML.

While change tracking and document comparison tools do exist for structured documents, not all are fully "XML-aware". As a result they suffer from a number of limitations that create challenges and problems for document managers. For example, they may:

- create additional, often unnecessary, work by presenting changes where none in fact exist
- present results in ways that introduce confusion - rather than clarity and comprehension
- introduce ambiguities that would be avoided if the editing tools were fully "XML-aware"
- be unable to record in detail the sequence of changes that transform one version of a document to the next
- be unable to identify clearly changes in parts of the document, such as complex tables, where the need for accuracy may be of paramount importance and manual checking will be both laborious and error-prone
- present results that then still rely on the judgement and intervention of an editor
- deliver inconsistent results – for example in a corporate environment where different teams use different editing tools
- deliver output that is not sufficiently accurate and consistent (nor in the right format) to enable subsequent process automation.

The result is that the process of editing and maintaining structured documents consumes more time and effort, negating many of the potential benefits of standardisation and process automation. Worse still, poor management of document change can also present real business risks such as failure to comply with regulatory or legal conditions.

This paper identifies some of the challenges that result from adopting XML and explains how a fully XML aware comparison tool such as DeltaXML can improve accuracy, consistency and quality while at the same time **reducing review time by up to 50%,** and significantly reducing "total edit time" – thereby delivering dramatic cost savings for XML document managers.

Benefits of structured documents in publishing

Structured documentation has transformed the world of publishing. The paradigm shift occurred in the 1960s, when early computer boffins realised the need for a standard way to describe an electronic document. They broke the tie to the printed page and described a document not in terms of its appearance in one media but in terms of its attributes and hierarchy. Instead of an ordered string of words and objects, a structured document could now be treated like a database capable of an infinite number of outputs.

And it is this flexibility that has led to the broad acceptance of XML, now the most widely used derivative of the industry standard for structured documentation, SGML (Standard Generalized Markup Language). Publishers can reuse content at will to meet the demands of different audiences and the technical requirements of different media. Localisation costs can be dramatically reduced; they can now publish for print, for e-books and for the web from a single source; and indexes, cross-references, footnotes, anthologies and bibliographies can be generated on demand to deliver a richer experience for the user or to create new content for publishing or re-sale.

For example, from a single source, a medical publisher can produce complex print and digital publications as well as subsets for syndication or to meet publishing opportunities for specific populations or diseases as well as generating information for specific groups such as doctors, nurses and patients.

The opportunity for managing change in structured documents

Structured documentation has transformed the way in which publishers can create, maintain and publish documents; but the

management of change in those same documents has failed to keep pace. While change tracking and document comparison tools do exist for structured documents, they don't match the capability of desktop systems and are generally limited to producing a view in an editor or, in some cases, a redline document in PDF or HTML.

This paper aims to show how it is possible to extend the advantages of structured documentation to the management of change by providing rich and easily processable information about changes that can add real value to the management process.

Embracing change

Change is going to happen anyway, especially if you are in the publishing business, so rather than treating it as a necessary evil, embrace it. Consider change an important and integral part of the publishing process, a valuable by-product not a waste product, and one that can deliver significant competitive advantage. The first version of a document is rarely the final version; a first draft will not be the last.

Change is not just important in the production or authoring process. In many cases a published document is likely to have future versions and this is where change itself has immense value. A user who has read one version of a 50-page document is not going to be very happy if he is presented with a new version with no indication of what has changed. If the publisher can provide a view of the changes since the last version or, better still, the changes since a user last viewed the document, the reader will be much happier and this richer experience can add real bottom-line value for the publisher.

Embracing change can take you further. Once a publisher understands how its products are being used, it may well find that its customers are more interested in change than in the information in the document itself. An engineer, for example, might be familiar with a user manual but is only interested in changes since he last reviewed it or changes between a manual for the latest revision of a product compared to the version being replaced.

The publication of standards is one area where change itself is fast becoming an essential product. Traditionally, standards have been static documents that will go through continuous iterations from interim versions to the final adopted document. This process of change can make life very difficult for consumers of standards if they cannot see where the changes have been made. Standards bodies, including ISO are now addressing this and looking to publish a view of changes to standards.

Tracking and Comparing

There are two basic methods used to measure change in documents; changes can be recorded and tracked as they are made, or the new and old documents can be compared. Many people are familiar with the tools provided for these functions in widely-used office automation software.

But, on its own, tracking has a number of significant disadvantages.

- there is no way of being certain that tracking has been turned on all the time
- not all changes are tracked, for example, changes to tables
- unnecessary changes are often included – for example, changes which are made and then discarded
- it is not possible to show selected changes, for example, only those made since a user's last review

Comparison is almost always used alongside tracking, especially with documents that need to change during their lifetime. Understanding the differences between different versions of a document is absolutely vital and, in some industries, can even be a legal requirement.

What is difficult about changes in structured documents?

Regular comparison tools simply do not work for structured documents. Most of today's word processors incorporates powerful review and compare tools, but they are not able to work with XML and so cannot be used to compare structured documents. It is possible to review a changed document by generating a PDF but it would not be possible to accept or reject changes or to process the changes in any other way.

Text comparison tools, such as those used in source code control systems, generally compare on a line-by-line basis and, because they have no understanding of XML, they will invariably detect too much difference and the end result delivers no advantage over a visual comparison of the two documents.

In order to get good comparison results for structured documents, the comparator must fully understand not only the structure of XML, but all the syntactic details. For example, XML attributes need to be handled in a special way, they cannot be handled as text; it is not possible to handle whitespace correctly without being able to read-in the schema and understand certain XML attributes. XML also has certain methods for representing standard text items, and unless these are understood by the comparator, they cannot be handled correctly. The same applies to change identification at word level, especially if there are references to external content.

How can a structured change be represented in XML?

Understanding the XML in order to perform a sensible comparison is only the first step. The next step is probably even more important. Once the changes have been found, they need to be represented, and the ideal would be to represent them in XML in a way that could be easily processed. The DeltaXML delta file does this in a unique way as it can represent not only any change between any two documents, but any change between any number of documents.

Once a change has been represented in this delta format, then all the advantages of processing the original XML document are immediately available to process the changes. The examples in the figure below show how the changes would be rendered in HTML and as a redline display.

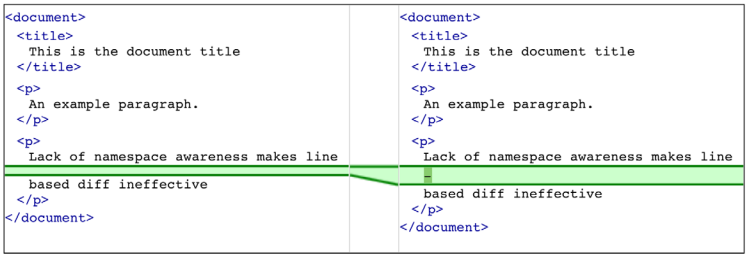

Figure 1: Side-by-side rendering of changes

Figure 2: Redline rendering of changes

Because the result produced by DeltaXML contains all of the content and structure of the input documents, it is possible to generate a wide range of outputs, from simply highlighting the added content to a complete merge of the two. This flexibility can lead to new opportunities for richer published material and to highly targeted and bespoke change representation for specific purposes.

Issues affecting readability of changes

Identifying changes in structured documents is only a part of the problem. If the representation of those changes is difficult to read the whole point of identifying the changes in the first place is lost. As is so often the case, the simplest looking and most intuitive results can take a lot of complex processing to achieve.

Seeing change on a word-by-word basis

It is clearly more helpful to see changes on a word-by-word basis rather than having whole sections deleted and added. But achieving this is not as trivial as it sounds. For example, punctuation must be correctly identified as a word separator, yet some characters form word boundaries in some situations but not in others. A full stop can also be used as a decimal point in numbers and a comma also serves as a numeric separator or indeed as a decimal point in European languages.

Intelligent, language-dependent lexical analysis of the document is necessary in order to get a good result. This is even more evident in Eastern languages, which do not use spaces and punctuation as separators.

Handling formatting structure

XML mark-up is not only used to define the structure of a document in terms of section or paragraphs but also to determine the formatting of particular words or phrases. However, in some situations a change to the formatting can result in false identification of changes to the text and it requires special processing to ensure that these formatting elements are handled correctly. This ability to identify these elements accurately brings the added advantage that changes to formatting can be shown or suppressed in the results at will.

Making results readable

The comparison process will always try to identify the best match, typically a mathematically optimal result, between two paragraphs that are in the same place in the two documents being compared. But this can sometimes lead to results that are either very difficult to read or not intuitive. For example, in a block of text where most of the words have been changed, if two or three words are the same and are each individually separated by changes, it can be very difficult to read. In this situation, it is better that these orphaned words are merged into the added or deleted text to make it easier to read.

```
<para>A pangram uses all the letters of the alphabet. It and is often used
        to test typewriters or computer keyboards, for example:
        TheA quick brown fox jumps overmovement of the lazy dogenemy
          will jeopardize six gunboats.</para>
```

Figure 3: Separated changes are difficult to read

```
<para>A pangram uses all the letters of the alphabet. It and is often used
        to test typewriters or computer keyboards, for example: The quick
        brown fox jumps over the lazy dogA quick movement of the enemy
          will jeopardize six gunboats.</para>
```

Figure 4: Grouping changes significantly improves readability

Similarly, where two paragraphs have only a small percentage of text that is the same, it may be better to identify them as different paragraphs and show one as deleted and the other as added. Again, special processing is needed to ensure that this happens by setting a threshold that has to be reached to consider two paragraphs as being the 'same'.

IETF RFC 2119
IETF (Internet Engineering Task Force). *RFC 2119: Key words for use in RFCs to Indicate Requirement Levels*. Scott Bradner, 1997. (See http://www.ietf.org/rfc/rfc2119.txt.)
IETF RFC 23963066
IETF (Internet Engineering Task Force). *RFC 2396: Uniform Resource Identifiers (URI): Generic SyntaxRFC 3066: Tags for the Identification of Languages*. T. Berners-Lee, R. Fielding, L. Masinter. 1998, ed. H. Alvestrand. 2001. (See http://www.ietf.org/rfc/rfc2396rfc3066.txt.)

Figure 5: Incorrect alignment of paragraphs causes confusion

IETF RFC 2119
IETF (Internet Engineering Task Force). *RFC 2119: Key words for use in RFCs to Indicate Requirement Levels*. Scott Bradner, 1997. (See http://www.ietf.org/rfc/rfc2119.txt.)
IETF RFC 23963066
IETF (Internet Engineering Task Force). *RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax*. T. Berners-Lee, R. Fielding, L. Masinter. 1998. (See http://www.ietf.org/rfc/rfc2396.txt.)
IETF (Internet Engineering Task Force). *RFC 3066: Tags for the Identification of Languages*, ed. H. Alvestrand. 2001. (See http://www.ietf.org/rfc/rfc3066.txt.)

Figure 6: Correct separation of different paragraphs improves readability

Turning the Tables

Tables can be particularly challenging, especially in legal and financial documents, where they can be large and the requirement for accuracy in any changes to the content can be critical. Identifying changes to the content of individual cells is relatively easy, but the addition or deletion of columns and rows, or the merging of cells is far trickier. Some comparison programs avoid these difficulties by simply showing the old and the new tables without identifying what has changed. In a large and complex table, this is hardly helpful to the reviewer.

DeltaXML

DeltaXML's roots date back to 1991 when the company specialised in the management of change in complex structured datasets, such those used to develop electronic circuits, an area demanding exceptional accuracy. With the advent of XML in 1998, the company began to represent those changes in the new mark-up language and soon realised the potential for applying its comparison technology to structured documents. A period of extensive development followed to enable accurate handling of syntax, tables, formatting and language as well as the readability of the output.

The first product, DeltaXML Core was launched in 2001 and is now used by some of the world's largest blue chip companies, government departments and by OEMs in many different markets. DeltaXML is recognised as the leading product in XML document and data comparison and is now helping over 200 class-leading companies and tens of thousands of users worldwide to manage change in a wide range of structured documentation environments.

DeltaXML is uniquely able to represent changes using standard XML mark-up, so that the changed document can take advantage of the

same easy processing as the original. It can then be output in a variety of forms for different purposes from simple redline documents to selective indexes of changes, statistics and even to enhance online web content by highlighting changes.

And because DeltaXML understands the structure of XML and is able to handle changes to attributes as well as being aware of word breaks and punctuation, the results are more intuitive, saving time particularly for costly resources such as subject matter experts.

There are specialist applications that work in conjunction with DeltaXML Core to handle specific standard formats such as HTML5, DITA and DocBook. These applications understand the particular formatting and table elements that are used in these standards and are able to provide results that can be used in applications such as the submission of changes to regulators. And because it has been designed for customisation, DeltaXML can even be used where organisations have developed their own document schemas and need a versatile tool kit that can be customised for their particular use.

DeltaXML Core has a comprehensive Java API and a .NET API for seamless integration. And because it was designed for large and complex XML environments it will scale for large systems and for big data, protecting investment by avoiding the need to change tools as complexity grows.

DeltaXML Publishing Applications

DeltaXML has been used extensively by publishers not just to manage change, but to add value to their businesses. These are some of the applications.

Identifying changes for review

Review is an essential part of any publishing process and, whether it is review by peers, a management review or a subject matter expert review, the processes are very similar but the requirements in each case can vary. Change tracking in XML documents will be a standard process internally, but with external reviews, those controls are not always present and changes can be made that are not tracked. DeltaXML's powerful comparison tools will identify the changes, removing the need to develop workarounds, for example having to transform documents back into Word just so that they can be compared.

Whether it's more convenient to have a PDF document for review or to have tracked changes within an XML editor so that each change can be accepted or rejected, either way, DeltaXML is able to provide a solution.

Submission of Financial Publications to Regulators

The management of change in financial publications can be extremely demanding. The documents themselves often have to meet strict legal requirements, demanding a high degree of accuracy; they are frequently long, running to hundreds of pages, and can contain many complex tables. DeltaXML will always identify every change. Some comparison systems are unable to process tables intelligently and will simply show the old and the new without any indication of the detailed changes that have been made. DeltaXML has addressed this and will compare even the most complex tables, indicating not just changes in individual cells, but showing where rows have been added or removed, or cells merged.

Because DeltaXML represents changes in standard XML, the output can be tailored to meet exact requirements each time. For example, when submitting financial documents to a regulator, the requirement is to show only the exact changes between the new version and the version previously reviewed and not any intermediate changes that may have been made.

Online publishing – identifying changed content

Online publishers are always looking for ways to enhance their readers' experience, and make their content richer. One online publisher of court case information is using DeltaXML to provide its readers with personalised updates and changes to this information. By using XML documents containing all the change information, the publisher is able to let readers see updates relative to any selected time base.

There are many similar applications for technical documentation, standards, patents and other areas where users are able to see what has changed since a specific date, version or previous review.

Legislation

Legislative documents are highly structured and subject to very formal change control processes. DeltaXML is being used in a number of states in the USA to facilitate the process of formal legislative review and update and even to provide public access so that citizens can see what is being proposed in any change to the law.

Translation

The European Commission has to manage the daily challenge of translating legislation and changes in the law into more than 20 languages. DeltaXML is not only able to meet the accuracy required, but is also helping to reduce costs significantly by identifying only the changed sections for translation.

Software testing

Software testing may not seem like an application of publishing, but most publishers of structured documents have some fairly complex underlying software systems to manage and control the publication process. As this software is updated, it is necessary to make sure that it works as intended. DeltaXML is ideally suited to software regression testing for any software system that is generating or consuming XML documents. It can perform accurate comparisons quickly and, if there are changes, these can be notified to a user in a summary report, eliminating the need for time spent trying to identify whether changes are as expected or not.

Conclusion

This white paper has shown how change, when harnessed and managed, can add real value to a publishing business. The widespread use of XML structured documentation systems, and the deployment of XML-aware change management tools like DeltaXML, has made it possible to bring all of the advantages of structured information to the publishing of change in that same information.

The ability to process and present that information in many different ways, for different audiences and through different media, is leading to new opportunities to enrich user experience and to streamline documentation processes.

Head Office (Sales and Support)

DeltaXML Ltd
Malvern Hills Science Park
Malvern, Worcestershire
WR14 3SZ UK

deltaxml.com
info@deltaxml.com
+44 1684 532 130

DELTAXML EXPERTS IN INFORMATION CHANGE

© DeltaXML Ltd 2014 WP.XFP.10.14