

# INCORPORATING SEMANTICS AND METADATA AS PART OF THE ARTICLE AUTHORIZING PROCESS

*Pablo F. Fernicola*

Microsoft Corporation, One Microsoft Way  
Redmond, WA 98052, USA  
pablofe@microsoft.com

## **Abstract**

The ongoing shift in the delivery of publications, and in the consumption of content, from print to digital presents an opportunity to streamline the publishing workflow and to optimize the authoring process with digital content as the primary output, including the capture of semantics and metadata as part of authoring and the preservation of this data to the archival copy of the document. In addition to the shift in how content is delivered and consumed, a significant development in the last few years has been the release of new versions of word processors with native file formats based on XML. The use of XML in the authoring file format, combined with extensibility in its content model, will enable a greater level of content semantics and metadata to be expressed directly by authors. The level of interoperability enabled by XML-based word processing file formats will make it possible to preserve the semantics and metadata as documents go through the submission and review process, make it through the publishing workflow and are ultimately archived, likely also in an XML based format. This article describes the design considerations and possible benefits of the Article Authoring Add-in for Word 2007 to the scholarly publishing community, in particular for workflows focused on the production of documents for digital delivery and consumption, as well as for the XML based archival of publications. The second Beta release of the add-in is available as a free download (<http://research.microsoft.com/authoring>), and it is currently being evaluated by the scholarly publishing community, with the involvement of publishers, archives, information repositories, and early adopters. In addition to facilitating the creation of structured documents, and enabling semantics and metadata to be more easily captured during authoring, the add-in provides the ability to open and save files from Word 2007 into the XML format defined by the National Center for Biotechnology Information of the National Library of Medicine. The add-in extends the file format used by Word 2007, as well as its user interface, to tailor the authoring experience for the different audiences involved in the publishing workflow. As the add-in is adopted across multiple

publications, authors will benefit from a consistent baseline experience, simplifying the authoring process and enabling a shift towards emphasising the expression of semantics over presentation by authors.

**Keywords:** semantics; metadata; NLM XML; scholarly publishing; Word.

## 1. Introduction

As the use of the digital medium in the publication of scholarly content evolves, and the discovery and consumption of this content shifts from paper to reading on electronic devices, there is an opportunity to enhance the authoring experience to optimize the content produced towards this emerging medium, making it better suited for digital delivery and consumption, by enabling authors to express more of the semantics of the content and to capture information (metadata) about different topics or subjects that appear in the document. The additional semantic content can be used to improve the consumption of documents by making the content more relevant and discoverable as part of the search process, as digital content can contain and convey more information than paper based content. For example, the digital content can include metadata about the authors and the topics being discussed in the article, encapsulate research data, or even express the semantics of the different parts of the content, all in a manner that can be analyzed by computers, with the results being available for data mining or semantic analysis. As will be discussed later, part of the shift to digital content will likely involve a reduction in the relevance of the content's presentation attributes, with elements such as font attributes (size, color) and layout (justification and spacing) no longer being relevant as part of the process of writing articles and papers by authors. Instead, there will be greater importance and reliance on capturing the structure, semantics, and metadata related to the content at the time of authoring, with the goal of capturing as much as possible knowledge directly from authors, as the domain experts on the content of the document.

One of the goals of the Article Authoring Add-in for Word 2007 is to facilitate the capture of semantics and metadata during authoring. In addition, the add-in provides bi-directional full fidelity conversion of documents to the XML based format defined by the National Library of Medicine (NLM), which is commonly used in publishing workflows and in digital archives [1].

Enabling authors to capture a document's structure and semantics (sections, statements, or references to external data, for example) as part of the writing process usually involves presenting a set of guidelines and conventions for au-

thors to follow. As authors' understanding and adherence to these guidelines is not assured, the editorial staff frequently needs to fix up documents as part of the publishing workflow. Similar challenges appear when attempting to capture information that does not explicitly appear as text in the final article, such as keywords, author information, and other article metadata. After the author's work is done, and the article has been submitted and accepted for publication, articles are processed or tagged by the editorial and production staff as part of the publishing workflow. This processing also presents the need to incorporate additional semantics and metadata, both related to the content and also to the publication where the article will be published in. In many cases, the pre-publication tagging or editing is performed after transforming the document to a format different from the one originally used by the author. The differences in format between the original material and the one used for tagging make it hard to incorporate most late changes or corrections submitted by the author. Finally, the tagging process is usually done by a third party, who may not have the same insight or background knowledge as the authors on the subject matter being discussed in the article.

### **1.1. XML Usage from Authoring to Archival**

The convergence of two ongoing developments will enable better capture and preservation of semantics and metadata from authoring to consumption. One development, underway for some years now, is the practice of using XML as the underlying format in publishing workflows and as the basis for archival formats. A canonical example of an archive using XML based content is the National Library of Medicine's PubMed Central. A more recent development is the growth in the use of XML as an authoring format. The introduction of Microsoft's Word 2007 brought with it the use of XML in the default file format of this popular word processor. The underlying format used by Word 2007 is called Office Open XML, and Word documents in this new format are commonly referred to as *docx* files, because of their filename extension (.docx). The use of XML by Word 2007 is automatic and transparent to authors, and does not require any changes in the document writing process. Authors are not aware that the content they write is now captured and preserved as XML tags.

It is the use of XML by Word 2007, along with the extensibility both in the document format and in the document packaging format, which enabled the development of the Article Authoring Add-in for Word 2007, with the goal of simplifying the authoring of scientific and technical articles optimized for digital distribution and consumption.

## **1.2. Key User Audiences for the Add-in**

Based on current practices and workflows, two main audiences were identified for the add-in, authors and the editorial and production staff. These two audiences have different needs, focus, and skills. One of the main differences between these two audiences is in their familiarity with the details of the format used for publishing, and in their awareness in relation to the metadata required for archiving. The editorial and production staff tends to be aware of the target format that will be used for publishing and archiving, as well as of the metadata that is required, while authors tend to be unaware of the processing that documents go through after submission or any specifics in relation to the publishing and archiving formats. It was an explicit design goal for the add-in to avoid exposing authors to the format used for publishing, or even to make them aware that XML now underlies the documents that they author. Based on this goal and the two audiences, two levels of user interface are presented by the add-in.

The user experience presented to authors by the add-in emphasizes simplicity and usability, without trying to expose authors to the publishing format or XML elements directly, while at the same time providing the editorial staff with access to all the tags and constructs that are part of the NLM format. The add-in also aims to enable authors to convey more of their domain knowledge during the authoring process, for example, by defining some parts of the article as sections in a robust manner, and by enabling authors to tag the document with keywords.

It is not the expected that the add-in will totally remove the need for the editorial and production staff to correct or annotate the document, but rather that it will simplify some of that work, first by enabling the content from authors to be more easily structured, to conform to the requirements of the different publications, and to contain more semantic information, and also by enhancing the functionality within Word to account for NLM specific content.

The use of the add-in across different publications should result in a familiar and more consistent experience for authors. The intended outcome from the use of the add-in by authors is to facilitate the creation of documents that are better suited for digital publishing and consumption, in that they capture and expose more of the author's knowledge in a structured manner to search and for semantic analysis, and to simplify the publishing workflow. The add-in should not be viewed just as an endpoint solution, as it can also serve as a platform on which additional functionality can be developed by third parties, both at the user interface level and by consuming or transforming the add-in specific information within the docx files.

### 1.3. Presentation vs. Semantics during the Authoring Phase

As the consumption and distribution of scholarly publications shift from paper to being digital, the final version of the content will increasingly be archived in an XML based format [2], with the content presented to the public for consumption in a format derived from the archived version through a transformation (with HTML and PDF being two of the most commonly used presentation formats). The NLM XML format, and proprietary variations of it, is commonly used throughout the publishing industry as part of the publishing workflows and for archival. The Authoring Add-in focuses on providing support for the Publishing, Authoring, and Book tag sets of the NLM XML format [3]. These three tag sets share a core set of elements and properties, and for the purpose of this paper will be commonly referred to as the “NLM format”.

By design, the XML-based NLM format does not aim to capture presentation attributes of the content such as font size and color, or decorations for table borders. Instead, the NLM format captures the content, its semantics, and the metadata associated with the content and its authors. Presentation properties can be introduced as part of the transformation to the presentation format based on the semantics of the content, either during the archival process or on demand as the content is accessed. For example, the fact that a text string is contained in the Title element of a NLM XML file can be used to assign a specific font, color, size, and weight, such as bold, to the same text string in the HTML format, as part of the transformation of the file. The specific values for these presentation properties can be contained in style sheets, with different style sheets being created and applied for content belonging to different publications. As publishing workflows adopt this transformation based model [4], and authors move to XML enabled versions of word processors, many of the recommendations in the author’s guidelines currently in use, such as font size and color, line spacing, and paragraph justification, become unnecessary. Instead of being applied during authoring, presentation elements can be derived at a later time, based on the role (semantics) that the different parts of the content play within the document, more likely as part of the transformation from the archival format to the different presentation formats. Because it preserves the structure, semantics, and metadata, the archival format can also be used as the source material for data mining and for semantic analysis of the content. For workflows based on transformations to work well, and not result in an increase on the burden on the publishing staff, it is helpful for the document semantics and metadata to be captured during authoring, and preserved through the peer review and publishing process. The add-in facilitates this shift from presentation to semantics and richer archival content, by not only by preserving all Word content elements, such as tables and equations, when producing content in

the NLM format, but also by presenting sections in a very structured manner to authors, and preserving the structure of the article during conversion.

## 2. Methodology

Two of the core functional requirements for the add-in were to preserve as much of the content as possible when saving a document from Word 2007 into the NLM format, and to enable editing within Word 2007 of all the information that can be expressed in the NLM format, including metadata not present in the text of the document. Differences in the level of abstractions encapsulated by the NLM and Open XML formats presented a challenge in implementing the add-in. The NLM format has elements that are not found in Word 2007, such as the Speech and Acknowledgement tags, and in some cases the NLM format exposes properties that do not exist at all or in exactly the same form as in Word 2007. Also, as the NLM format is not focused on capturing presentation elements, the conversion of Word 2007 to the NLM format can be lossy in relation to presentation properties, with tables' properties such as border width and color being two examples of the information that is lost. On the other hand, the conversion from the NLM format to the Word 2007 format can retain full fidelity, as the content and user interface in Word is extended by the add-in to accommodate the NLM specific material.

From a software architecture point of view, the add-in involves three core functional areas. The primary functional area deals with the transformations between the XML formats, and it is based largely on the use of Extensible Stylesheet Language Transformations (XSLT). The second area involves extending the document and packaging formats to preserve NLM format specific information. The last functional area encompasses all of the user interface elements of the add-in. The focus of this section is describing the overall approach for expressing and preserving information in the docx file, as well as the user interaction model of the add-in.

The extensibility mechanisms provided by Word 2007 and the underlying document format enable the add-in to present all of the NLM specific information either as new content elements within the flow of the document or in new user interface elements (panels), for content that is not part of the text of the document, such as the author and article metadata. Also as part of the extensibility in the Word user interface, it is possible for the editorial staff audience to access and edit NLM specific metadata through forms, without requiring the staff to edit raw XML markup.

For its implementation, the add-in relies on or extends three key technologies: XML, Word templates, and the Word user interface. The add-in presents three user

interface components to authors: templates, a custom ribbon, and an Author Panel. Templates play a key role in conveying the requirements from the journal to the authors, and providing some level of guidance. Templates are one of the more visible manifestations of the add-in to be surfaced to authors. Opening a document created with the template automatically turns on the add-in, and information in the template can be used to automatically configure some of the user interface elements in the custom ribbon (the Ribbon is one of the key new user interface elements in the Office 2007 suite [5]). The last component of the add-in, the Author Panel, provides a simple way for authors to enter and edit metadata.

### **2.1. XML as a Key Enabling Technology**

The use of XML as the basis for the native file format for Word 2007, Office Open XML [6] [7], provides a key benefit in terms of interoperability, as it makes transformations to other XML based formats simpler, especially when compared to the work involved in transforming binary based file formats. The Office Open XML format also allows for extensibility, both as part of the document content and as part of the container of the document. Word 2007 documents follow the packaging conventions defined by the Office Open XML format and are based on the zip container format.

The add-in extends the document content with a number of elements that are part of the NLM format and are not found in Word. These new elements appear within the document content, and are expressed in the file format as new XML elements, referred to as “Custom XML” elements. The add-in extends the document package by storing the template used to create the document, as well as metadata associated with the authors, the article, and the journal. In order to assist with interoperability, all data stored by the add-in is expressed using XML (re-using elements from the NLM format), which should enable documents authored using the add-in to be easily transformed and processed as part of publishing workflows, without loss of semantics or metadata.

A benefit of the document packaging being based on the zip format, and of the document content being in XML, is that the document can be transformed to other XML formats on any software platform, only requiring software libraries to interact with the zip format (to access the files inside of a docx file) and to parse, edit, or extract information from the XML files [8]. Given this level of access to the article’s content and metadata, publishing systems will be able to validate and extract relevant metadata from articles in an automated fashion, perhaps even as part of the article intake process, and alleviate the need for authors to fill out forms as part of the online submission process.

### 2.3. Templates as the Foundation for Structured Content

The add-in enhances Word templates (.dotx files) with the ability to present authors with required and optional sections, pre-populate keyword suggestions and minimum keyword requirements, as well as to provide guidance to authors on information such as commonly used grant numbers and funding agencies, which are useful for internal reports or grant proposals. The presence of this additional information in the templates is optional. In response to this information being present in template files, the Authoring add-in is enabled automatically and it configures required and optional sections in the user interface, providing greater guidance to authors, and simplifying the instructions that need to be presented to authors.

At the level of sections, templates can contain rules as to the minimum and maximum number of words required, restrict the sections available to authors to a pre-defined set, specify whether sections are required or optional, and also enable arbitrarily titled (custom) sections. At the document level, the template can encode keyword requirements (as well as expose a set vocabulary of keywords for author tagging), and require taxonomy classification of the content by the author (mapping to the Subject/Subject Group concept in the NLM format).

The information added to Word templates for use by the add-in is encapsulated as an NLM article, stripped of content, just capturing the structure of the sections, as well as keywords and subject groups if desired, and augmented with a set of custom properties, prefixed by the namespace "ms". It is this set of custom properties which are used, for example, to express the minimum and maximum number of words required for sections. The custom properties in the template are only used internally to the add-in, and are not included in the resulting NLM file. As a practical example of how templates and their properties can be used, the author guidelines for this paper specified that the papers should contain an Abstract, and that the Abstract should not exceed 450 words. The example below shows a NLM section element with the custom properties applied to express these author guidelines:

```
<sec sec-type="Abstract" ms:RequiredSection="True"  
ms:Instances="One" ms:MaximumWordCount="450" />
```

The guidelines encoded in templates can be automatically validated by the add-in, before submission of the article, enabling the author to correct any issues as part of the authoring process, thereby reducing the need for the editorial staff to send the article back for corrections.

The add-in installs a number of sample templates which can be used as a starting point to create custom templates. It is envisioned that journals and other publi-

cations will create their own templates and make these custom templates available for download by authors. Given that articles are many times rejected by one journal and subsequently submitted to others, the add-in assists in the conversion of documents from one template to another, adding any required sections from the new template not found in the original template to the document, and informing the author of any sections already present in the document which may not be allowed by the new template. As part of the process of applying a new template, the add-in also updates any requirements in relation to the presence and number of keywords, or other metadata presented by the add-in to authors.

## 2.4. User Interface Configuration

The default user interface presented by the add-in is focused on the Author audience and it is kept as simple as possible. It is a goal for authors not to have to learn new commands and for authors to be able to rely on the native Word user interface as much as possible. To the author audience, the add-in presents three user interface elements:

- A custom ribbon (named Insert Sections)
- A side panel (Author Panel)
- Custom XML elements.

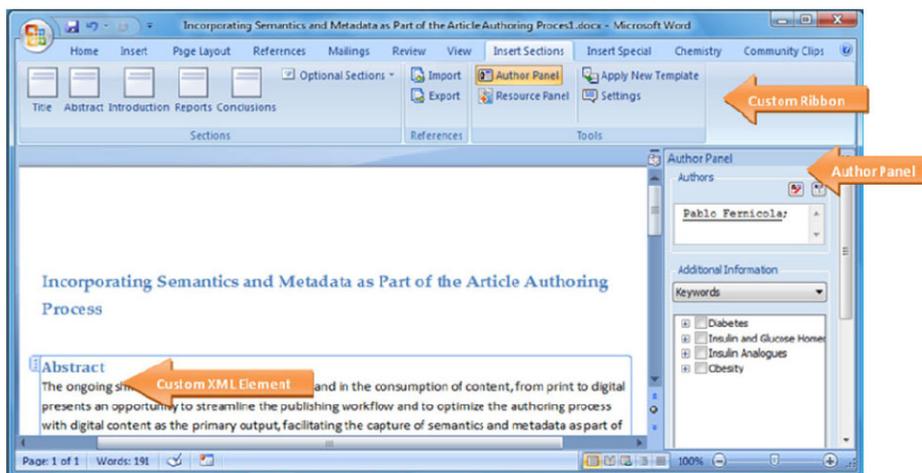


Figure 1: Author User Interface Elements.

From its default configuration, the user interface presented by the add-in can be expanded to present functionality that targets the editorial and production staff. This additional functionality manifests itself as the Insert Special custom ribbon, enabling access to elements that are specific to the NLM format, an additional

panel for article notes, as well as a Document Information Panel (DIP), which provides access to all of the metadata, expressed in the NLM format and presented as a form, using the Microsoft InfoPath technology.

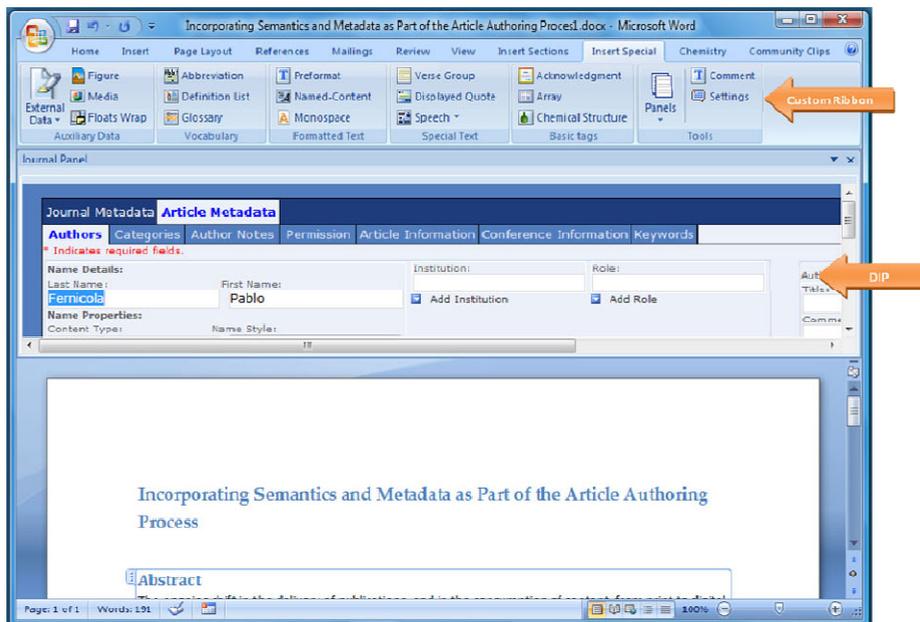


Figure 2: Editorial Staff User Interface Elements.

## 2.5. Client Side Validation

The enhancements to Word's template capabilities introduced by the add-in enable journals to surface elements such as the required sections of an article to authors when new documents are created based on a template. Further enhancing the authoring experience, optional sections can be presented in the user interface, under the Insert Section ribbon, providing greater guidance to authors, and simplifying the author instructions. Perhaps more valuable, templates can contain rules for sections, keyword requirements, and require taxonomical classification of the content by the author. Currently, validation is only performed automatically when the document is saved to the NLM format or when it is uploaded to a repository or submission system through the add-in's SWORD submission functionality, with the author being informed of any requirements that were specified in the template which are not met by the document.

### 3. Conclusions

Currently, the second version of the Article Authoring Add-in is in its late Beta stage (<http://research.microsoft.com/authoring>). In disciplines where the use of Word is prevalent, the add-in and its templates have the potential to enable documents to remain in the original format used by the author further into the publishing workflow, likely postponing the conversion to the archival format to the final step in the publishing pipeline. Also, the use of the add-in should enable the editorial staff to use the same word processor as the authors for editing the document or adding metadata, without having to deal directly with the raw XML file. Finally, the add-in and its template functionality should simplify the conformance of authors to authoring guidelines, as well as enable the capture of more semantic elements and metadata during authoring.

Publishers will be able to develop additional Word add-ins, building on the functionality exposed by the Article Authoring Add-in. As part of the publishing workflow, content in the document can be linked to databases, either by extracting data from the article and storing it in a database, by validating metadata against a database, or by bringing in data from external sources and creating the required XML tags in a document programmatically.

The add-in will not solve all issues related to the authoring, but it should help in improving the structure, semantics, and metadata of articles, as well as provide some level of validation prior to the article submission process, enabling the author to correct problems earlier in the publishing cycle. Perhaps the greatest benefit from the use of the add-in will be in the preservation of semantics and metadata entered by authors as the article goes through the publishing workflow, reducing the need for additional tagging and re-work, and resulting in articles with richer semantic content being exposed by search engines and available for semantic analysis.

### Acknowledgements

The author wishes to acknowledge the support, guidance, and inspiration for this project given by Dr. Tony Hey, Corporate Vice President of the External Research Division, Microsoft Corporation, and Jean Paoli, General Manager of Interoperability Strategy, Microsoft Corporation.

### Notes and References

- [1] *XML for e-journal archiving*. STEPHEN L. ABRAMS, BRUCE ROSENBLUM 4, s.l.: MCB UP Ltd, 2003, Vol. 19. 1065-075X.

- [2] DAY, MICHAEL Preserving the outputs of scholarly communication for the long term: a review of recent developments in digital preservation for electronic journal content. [book auth.] Wayne Jones. *E-Journals Access and Management*. s.l.: Routledge, 2008.
- [3] National Center for Biotechnology Information of the National Library of Medicine. Journal Archiving and Interchange Tag Suite. [Online] [Cited: April 5, 2009.] <http://dtd.nlm.nih.gov/>.
- [4] NATHAN C. HULSE, B.S., ROBERTO A. ROCHA, M.D., PHD, RICHARD BRADSHAW, M.S., GUILHERME DEL FIOL, M.D., M.S., and LORRIE ROEMER, R.N. Application of an XML-based Document Framework to Knowledge Content Authoring and Clinical Information System Development. *AMIA Annual Symposium Proceedings*. 2003, Vol. 2003.
- [5] Microsoft Corporation. The Microsoft Office Fluent user interface overview. *Microsoft Office Online*. [Online] [Cited: April 5, 2009.] <http://office.microsoft.com/en-us/products/HA101679411033.aspx?pid=CL100796341033>.
- [6] ECMA International. Standard ECMA-376 Office Open XML File Formats. *ECMA International*. [Online] [Cited: April 5, 2009.] <http://www.ecma-international.org/publications/standards/Ecma-376.htm>.
- [7] ISO. Office Open XML File Formats - Part 1: Fundamentals and Markup Language Reference. *International Organization for Standardization*. [Online] [Cited: April 5, 2009.] [http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=51463](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51463).
- [8] OpenXML Developer. *OpenXML Developer.org*. [Online] Microsoft Corporation. [Cited: April 5, 2009.] <http://openxmldeveloper.org/>.

June 2009  
Printed on demand  
by "*Nuova Cultura*"  
[www.nuovacultura.it](http://www.nuovacultura.it)

Book orders: [ordini@nuovacultura.it](mailto:ordini@nuovacultura.it)