Metadata in Microsoft Word documents

Martin Fenner, Gobbledygook

March 20, 2015

Metadata such as author, title, journal or persistent identifier are essential for scholarly documents, and some of us are spending a significant part of our time adding or fixing metadata. Unfortunately we sometimes don't pay enough attention to the flow of metadata, i.e. we ignore already existing metadata, or reinvent the wheel in how we describe or store them.

Storing metadata in text-based formats is usually straightforward. This blog post is written in markdown with a YAML header - think of YAML as the more human-readable version of JSON - at the beginning of the document:

```
---
title: Metadata in Microsoft Word documents
---
```

This is then translated into this HTML when the blog post is published:

```
<meta property="dc:title" content="Metadata in Microsoft Word documents" />
```

XML is of course a very natural format for metadata, here for example JATS used for scholarly articles:

```
<article-title>Metadata in Microsoft Word documents</article-title>
```

Many scholarly documents start out as Microsoft Word documents. And while the `docx` format introduced by Microsoft in Microsoft Office 2007 is XML-based, few users are aware of this fact. And probably even fewer users (including myself) ever go to the `Properties...` settings of a `docx` document and add a `title`, `keywords` or other metadata (the `author` is usually set automatically).

This is very unfortunate, as these metadata are very often required, e.g. in a journal article submission, and then need to be collected again, usually either by asking the author to fill out a web form, and/or by extracting the metadata (e.g. title) from the document.

The best place for metadata is with the document (not *in* the document), and if the file format (`docx` in this case) supports it, we should take advantage of this. The main benefit: metadata stay with the text when the document is sent to co-authors via email, or put on a file server, or into Dropbox.

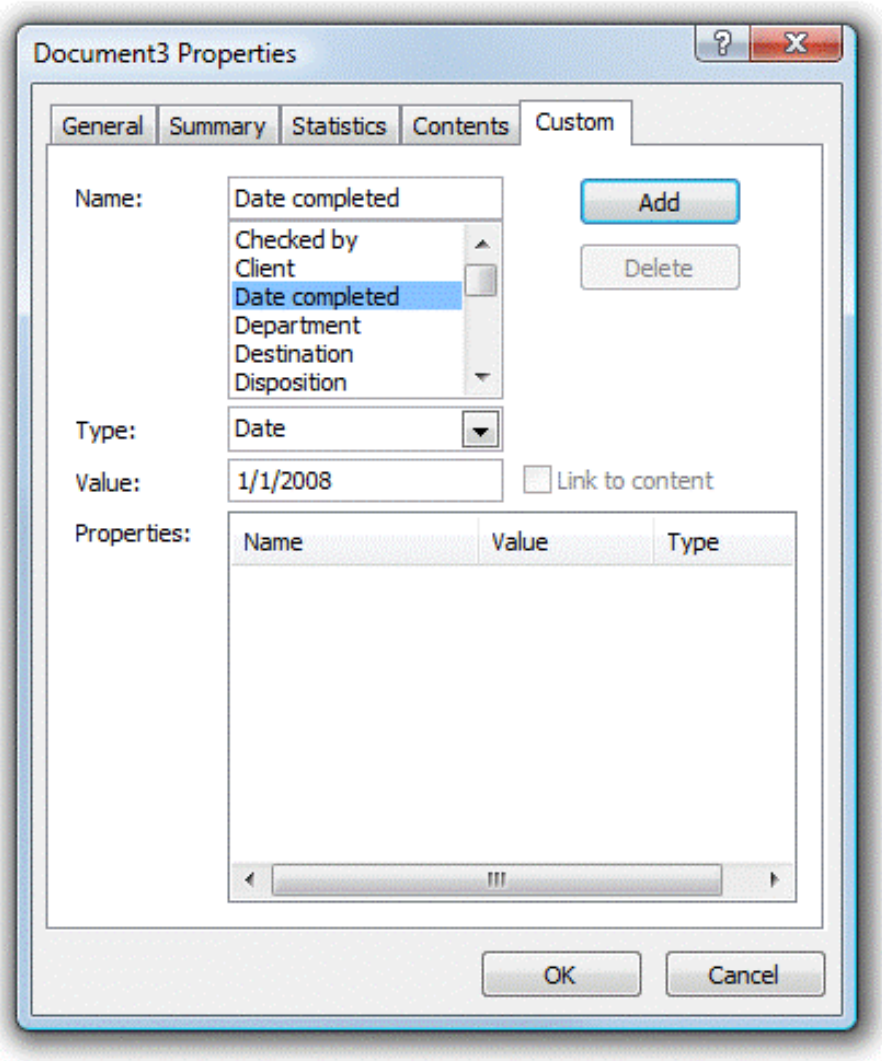

Figure 1: Microsoft Word 2007 Properties. Image from Microsoft Developer Network

In the case of `docx`, the metadata support is actually pretty good, using the standard Dublin Core, and storing the metadata in a separate file called `core.xml`. You can see this file if you unzip your `docx` file (e.g. after giving it a `zip` extension). The `core.xml` file for this blog post (after converting the markdown file to `docx` using Pandoc) looks like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<cp:coreProperties xmlns:cp="http://schemas.openxmlformats.org/package/2006/metadata/core-pr
```

Because `docx` is XML, we can read/write this file not only in Microsoft Word, e.g. using macros, but also outside of Microsoft Word, e.g. in workflows that converts `docx` documents into other formats, or tools that check `docx` files for required metadata (e.g. by using rakali that I wrote last year). So please encourage authors to use the Microsoft Word `Properties...` settings, and update existing tools to take advantage of the Dublin Core metadata stored in every `docx` file.