Outside In Clean Content SDK
Technical Note

Microsoft Office 2007 XML Document Properties

Version 1.0 – March 24, 2008

Change History

Version 1.0 – March 24, 2008
Initial version

Purpose

Microsoft Office documents contain three groups of metadata. Microsoft Office always creates two document property files, and a third when necessary. This document will explain how and when these three document property files are used by Microsoft, usable by third parties, and processed by Clean Content.

Background

There are three groups of metadata. Each group of metadata is stored in a different part[1]. The parts' relationships and their namespaces can be found in the master relationships file (_rels\.rels) and the content types file (\[Content Types].xml).

Core Properties

Core properties are properties that are common across all Microsoft documents. These properties are typically stored in "\docProps\core.xml" using the namespace "http://schemas.openxmlformats.org/package/2006/relationships/metadata/core-properties".

This file makes use of Dublin Core[2] for many of its properties. The fields displayed when choosing "Properties" from the "Prepare" menu from any of the Office applications are core properties. These properties are: Author, Title, Subject, Keywords, Category, Status and Comments.

The "Title", "Subject", "Creator", "Keywords", "Description", "Category" and "Content Status" properties are treated as Summary scrub targets by Clean Content. The "Revision", "Created" and "Modified" properties are Statistic scrub targets. Office will automatically repopulate these if the corresponding XML elements are missing. When these properties are scrubbed, Clean Content replaces the values with empty strings to prevent this behavior.

If an Office document does not have a docProps/core.xml stream, Clean Content will create one. Clean Content will create XML elements for only the properties that have an Add or AddOrReplace action.

Application Properties

Microsoft Word, PowerPoint and Excel each have application specific properties. These properties are typically stored in "\docProps\app.xml" using the namespace "http://schemas.openxmlformats.org/officeDocument/2006/relationshiops/metadata/extended-properties".

[1] "Parts" and other terms are defined in Microsoft's Office XML documentation: http://msdn.microsoft.com/en-us/library/aa338205.aspx#office2007aboutnewfileformat_introduction
[2] Information on the Dublin Core Metadata Initiative can be found here: http://dublincore.org/

This file will use elements from the namespace "http://schemas.openxmlformats.org/officeDocument/2006/docPropsVTypes" which includes definitions for vectors and variants.

Most of these document properties may be found in a dialog in the Office applications by choosing the "Advanced Properties…" drop down from the Document Properties panel. The application properties can be found in the Summary, Statistics and Contents tabs.

The Summary tab shows the Core properties along with "Hyperlink base" and "Template" properties. These two properties are Summary targets and when scrubbed will be replaced with empty strings rather than removing their XML elements.

The Statistics tab's properties are all Statistic targets.

The Contents tab shows data that are treated as Content Properties by Clean Content. These correspond to the HeadingPairs and TitlesOfParts elements. PowerPoint puts fonts, slide titles and themes here. Excel puts sheet names here. Word uses these for headings and titles.

Some documents have been found that contain an "HLinks" element. This element is a vector of variants. The variants that contain text are treated as content properties. When scrubbed, they are replaced with empty strings.

Custom Properties

Authors and third party applications may add custom properties. These properties have a Name, Type, and Value. The Microsoft Office user interface allows Text, Date, Number or Boolean property values. The XML specification allows for other types, but Office will present a warning and fail to parse the entire XML stream. SharePoint and Oracle Universal Content Management use custom properties to manage metadata.

If there is no custom property stream in the source document, and if any custom properties are being added, then Clean Content will create the stream in the resulting document.

Thumbnail

A thumbnail image of the document may be stored, and is typically stored in "\docProps\thumbnail.jpeg". This image is treated as the Thumbnail scrub target.

Non-valid Office documents

It is possible for third parties to create Office documents that properly conform to the various schemas, but are not opened by the Office applications. Clean Content will scrub and analyze such documents without throwing any exceptions.

Documents that do not conform to the schemas will throw exceptions when scrubbed or analyzed. When document property XML files are valid but not parsed by Office, the Office application will ignore the document property stream after providing an error dialog asking whether Office should attempt to recover the rest of the document.