Outside In Clean Content SDK
Technical Note

Microsoft Office Open XML Support

Version 3.0 – November 11, 2013

Change History

Version 1.0 – September 7, 2010
Initial version

Version 2.0 – March 24, 2011
Added information on Part Validation feature

Version 3.0 – November 11, 2013
Revised Markup Compatibility section based on new behavior in Clean Content 2013.1

Overview

The introduction of an XML-based file format in Microsoft Office 2007 and the subsequent "standardization" of this format in ECMA 376 and later in ISO 29500 has produced many issues related to correctly processing this file format. This document touches on the major issues as they relate to Clean Content extraction, analysis and scrubbing. Some familiarity with the standards involved is assumed.

ECMA 376 vs. ISO 29500

As of version 2010.2 Clean Content is now basing it's processing of Office 2007/2010 documents on a modified version of the ISO 29500 Transitional schema. These modification have been made either to support known instance documents that don't conform or based on recommendations from Microsoft on OpenXMLDeveloper.org or MSDN.

Markup Compatibility

Clean Content correctly processes AlternateContent blocks and Ignorable attributes which constitute the core of the *ISO 29500-3 Markup Compatibility and Extensibility* specification. As of Clean Content 2013.1, all Choice and Fallback blocks are processed for both scrubbing and extraction.

Effect on extraction and analysis

In cases where AlternateContent blocks are present in the document, customers will see two, often different, versions of the information in Clean Content's XML output.

Effect on scrubbing

All Choice and Fallback blocks will be preserved through scrubbing.

Part Validation

As of version 2011.2, Clean Content now supports additional deep validation of the part structure of Office 2007/2010 documents as defined in *ISO 29500-2 Open Packaging Conventions*. This feature is enabled by setting the `OpenXMLPartValidation` option to `true` (it defaults to `false`). When this feature is enabled Clean Content provides the following additional behaviors.

- All known references, all parts and all `.rels` file entries will be tracked.

- XML parts in schemas that Clean Content understands will be parsed even if they are not used for extraction. Word's `settings.xml` file is a good example of this.
- Non-XML parts that Clean Content understands will be processed as usual. Embedded images are a good example of this.
- All other parts fall into four categories. There is now a new analysis target for each of these categories.
    - **Unanalyzed** parts
      These are parts that are correctly referenced (explicitly or implicitly) but for which Clean Content does not have a parser. Examples are `PrinterSettings.bin`, `ActiveX*.bin`, etc. Many of these are scrubbable under certain targets, for example `PrinterSettings.bin` is scrubbed under the PrinterInformation target. If this is the case the isScrubbable attribute of the part's officexmlpart element (see below) will be set to true. Parts of this type that are not scrubbed should be of concern. The new analysis target for this type of part is OfficeXMLUnanalyzedParts.
    - **Unexpected** parts
      These are parts that have a valid `.rels` entries but are either not referenced or referenced in a context that Clean Content does not understand. An example of this is certain custom XML parts added by Microsoft SharePoint. It is expected that over time Oracle will support additional undocumented or obscure references thereby reducing the already small number of unexpected parts in any set of documents. Parts of this type should always be a concern. The new analysis target for this type of part is OfficeXMLUnexpectedParts.
    - **Alternate Content** parts
      These are parts referenced from the Choice sections of an AlternateContent elements and not from anywhere else. An example of this is when Office Mac inserts a PDF as an image. The Choice section of the AlternateContent element has a reference to the PDF but the Fallback section (the one Clean Content uses and keeps when scrubbing) has a reference to a PNG image. These parts will be scrubbed and so should not be a concern in the scrubbed document. The new analysis target for this type of part is OfficeXMLAlternateContentParts.
    - **Rogue** parts
      These are parts that exist in the ZIP container but do not have valid `.rels` entries. Example include parts in the [Trash] folder. These parts will be scrubbed but any document with a rogue part that isn't in the [Trash] folder might be of interest. The new analysis target for this type of part is OfficeXMLRogueParts.
- All Alternate Content parts and Rogue parts will be scrubbed.
- Extracted output will include a new `collection` element of type `OfficeXMLPartDiscloserRisks`. This collection will contain `officexmlpart` elements for all the parts in the categories listed above. Under each `officexmlpart` element will be the content of the part itself as if it were an embedding. This allows the customer to extract (if Clean Content has a parser for the format) and/or export any parts they find questionable.

Part validation is expensive from a processing standpoint and some loss of performance will be associated with turning this feature on. It is expected that only customers in high security environments where intentional disclosure of information is a real concern will make use of this feature.