Outside In Clean Content SDK
Technical Note

Microsoft PowerPoint Document Assembly and Disassembly

Version 1.0 – November 3, 2008

Version 1.1 – September 8, 2009

Change History

Version 1.0 – November 3, 2008
Initial version

Version 1.1 – September 8, 2009
Updated to address support for Office 2007PowerPoint support.

Overview

The purpose of this document is to provide details related to the assembly and disassembly features of Clean Content as they apply to Microsoft PowerPoint documents. The PowerPoint document assembly/disassembly features of Clean Content are designed to allow the production of new PowerPoint documents using any set of slides referenced in any order, from existing PowerPoint documents. This process can be used to implement an automated approach to disassembling a single presentation into a set of single slide presentations and for assembling a new presentation from a selected set of source presentations.

The disassembly feature allows applications to burst a presentation into a discrete set of component slides for storage and retrieval.

The assembly feature allows a selected set of disassembled slides to be reassembled into a new presentation on demand.

The combination of disassembly, assembly, and slide fingerprinting can be leveraged to create a rich set of tools for storage, search, retrieval, and presentation building.

Both the binary (PowerPoint 97 thru 2003) and XML versions (PowerPoint 2007 and above) of PowerPoint are supported during the disassembly and assembly process. Note that the assembly process does not support mixing the binary and xml versions of PowerPoint documents.

PowerPoint Assembly/Disassembly Functional Details

The process of bursting a single presentation into a set of new single slide PowerPoint documents is implemented by effectively copying each individual slide into a new presentation file.

The process of merging a set of presentations into a single presentation is implemented by copying every slide from the source list of presentations into a new presentation file.

The Clean Content API provides several options that trigger this operation and provide the calling application with control over specific details. Sample code is provided in the SDK documentation that outlines the API usage.

Slide Content and Resources

The primary challenge involved in this process is to detect and copy the slide content, formatting, and any document wide resources leveraged by the slide to the new presentation file.

The majority of all slide content and formatting, including text boxes, shapes, word art, images, objects, colors, transitions, etc., are supported during the process.

There are numerous document wide resources that can be used across multiple slides that must be detected during this process including fonts, images, OLE Objects, sounds, video, slide masters, speaker notes, and user comments. In order to optimize the resulting file, only resources referenced by copied slides are copied into the new document.

A unique benefit of the disassembly and assembly process is that resulting documents no longer carry obsolete slides that may linger in original presentations. PowerPoint documents that have the Fast Save feature enabled are notorious for growing unexpectedly large because obsolete slides remain buried within the file.

Multiple Masters

During the disassembly process only the masters leveraged by the required slide are copied into the new single slide presentation. This is done to optimize the resulting document size.

During assembly, multiple masters are leveraged as needed in order to maintain the formatting of every slide copied into the new presentation. Each master is effectively fingerprinted in order to avoid duplicating masters that are in fact identical across a set of merged presentations.

The possibility of inheriting a single master during the assembly process has been considered but was not pursued due to the conflicts that arise when applying masters with different geometries than the original source slides. A more effective approach is to maintain the original masters during assembly and allow the final presentation to be reformatted within PowerPoint by the user; applying any desired set of masters should reformatting be required. The PowerPoint application implements robust slide reformatting when new masters are applied to a set of existing slides.

Mixed Mode Assembly Limitation

The assembly process does not allow mixing binary ppt (PowerPoint 97 thru 2003) and XML pptx (PowerPoint 2007) documents during the same assembly process. It is required that source presentations be converted to one format or the other prior to assembly. The two formats are functionally very similar but radically different at the storage level. Several techniques to allow mixed mode assembly have been considered but have not been implemented.

PowerPoint 2007 Slide Hyperlink Note

In order to support hyperlinks from one slide to another it is possible that additional slides may be included during the assembly and disassembly process of PowerPoint 2007 documents. If a slide is referenced by any slide in the source list then it too will be copied to the target presentation even if the referenced slide was not in the original source list. Linked slides that are not in the source slide list are placed at the end of the created presentation. This may result in creating larger than expected presentations during disassembly and assembly but has the benefit that linked slides will not be lost during assembly. Note that this support is different than PowerPoint 97 thru 2003 where such internal hyperlinks were lost during the process as documented below.

Known Caveats and Limitations

Certain features of PowerPoint are limited or dropped during the assembly or disassembly process as outlined below.

Encrypted documents are not supported at this time.

Embedded Visual Basic projects are not copied during the process because a document can only contain a single VB project. Copying VB projects during disassembly would result in document size drawbacks and storage duplication. Copying VB projects during assembly would result in multiple VB project conflicts that are best avoided.

Embedded Fonts are not copied during the process because embedded fonts are effectively font subsets of all characters used within the original presentation. Retaining the embedded font during disassembly has document size drawbacks and merging multiple subsets from different presentations during assembly is not supported. The effect is that generated presentations will operate as if embedded fonts were not enabled when creating the presentation. Embedded fonts may then be applied to the final presentation assembly from within the PowerPoint application if desired.

Image based bullets are replaced by the default bullet during the copying process. This is due to technical hurdles related to how image based bullets were introduced into the PowerPoint file format.

Certain types of internal presentation links may not be maintained during a disassembly/reassembly process of PowerPoint 97 thru 2003. Specifically, links to explicit slide numbers may not be maintained because the slides may be re-ordered during reassembly. The assembly/disassembly of PowerPoint 2007 documents will maintain internal slide links by including all slides that are referenced when creating the new presentation.

PowerPoint only allows a single page setup to be applied to an entire presentation. For instance, it is not possible to mix a portrait and landscape page setup across different slides within a single presentation. For this reason, assembled presentations will inherit the page setup of the first slide of the merged set of slides. This can have undesirable layout consequences on slides that were formatted with a different page setup. While not common, merging slides that were formatted with different page setups is supported but not recommended.