Outside In Clean Content SDK
Technical Note

Microsoft PowerPoint Fingerprinting

Version 1.0 - April 22, 2008

Change History

Version 1.0 – April 22, 2009
Initial version

Overview

The purpose of this document is to provide details related to the fingerprinting feature of Clean Content as it applies to Microsoft PowerPoint documents. Fingerprinting is a feature that analyzes an input document during extraction and generates electronic fingerprints that uniquely identify specific document characteristics. This allows search technology to find documents with similar characteristics by comparing the fingerprints among a set of documents.

Fingerprints are provided as a 128 bit md5 hash code making comparison very fast while limiting the cost of storage. Clean Content optionally generates three types of fingerprints when processing Microsoft PowerPoint documents; Slide Content, Slide Appearance, and Graphic Data fingerprints. Each can be used to find similarities among Presentations in order to narrow down or expand the set of matches found during a search.

Type of Fingerprints

Slide Content Fingerprint

The Slide Content fingerprint is a unique code generated from the text and image content found on each slide. There is one Slide Content fingerprint for each slide found in the document. This fingerprint can be used to find slides that share the same content, regardless of formatting. The text and embedded graphics and objects found on the slide must be identical, and drawn in the same logical order, in order for the content fingerprint of two slides to match. All other slide attributes, including shape types, locations, and fills, are ignored when generating this fingerprint.

This fingerprint allows the calling application to detect slides that very likely originated from the same source slide even when they have undergone changes to their appearance. For example, if a presentation is modified to use a different slide master, background, or other formatting attributes, the slide content fingerprints will remain the same as in earlier versions.

Slide Appearance Fingerprint

The Slide Appearance fingerprint is an extension of the Slide Content fingerprint that additionally considers the shape type, location, fill type, and primary fill color of every shape found on the slide. The slide background and applicable master are also considered when generating this fingerprint. There is one Slide Appearance fingerprint for each slide found in the document. The text, embeddings, selected shape attributes, slide background, and the applicable master appearance fingerprint must be identical, and drawn in the same logical order, in order for the Slide Appearance fingerprint of two slides to match. There are many other slide attributes that are ignored when generating this fingerprint, including comments, speaker notes, advanced fill attributes, and transitions, to name a few. However, most reasonable changes to a slides appearance will result in a modified fingerprint.

The Slide Appearance fingerprint allows the calling application to detect slides that are extremely similar in both content and formatting. For example, when presentations or specific slides are copied and shared among a group of contributors, individual slides will maintain the same fingerprint until they undergo some level of content or formatting edits.

Graphic Data Fingerprint

The Graphic Data fingerprint is a unique code generated from the image data associated with embedded graphic content found within Microsoft Office documents, including PowerPoint. There is one Graphic Data fingerprint for each graphic image found in the document. This fingerprint is also incorporated into the Slide Content and Slide Appearance fingerprints each time the associated graphic is found on a slide.

This fingerprint allows the calling application to find all slides or documents that include a specific graphic image given its fingerprint.

Fingerprint Consistency across different versions of Office

It is not uncommon for a presentation to be modified and saved between different versions of Microsoft PowerPoint during the presentations lifecycle. When a presentation is edited using different versions of PowerPoint, the presentation is likely to undergo slight changes that are not visually obvious. The fingerprint algorithms discussed in this tech note have been designed with consideration for these changes. As a result, the Slide Content, Slide Appearance, and Graphic Data fingerprints of otherwise unmodified slides are typically maintained across saves between different versions of PowerPoint from Office 97 through Office 2003.

The introduction of Office 2007 has resulted in substantial feature and storage changes in PowerPoint documents. When a PPT document (Office 97 through 2003) is opened in 2007 in compatibility mode, and then saved back to the native PPT format, it will have undergone a conversion process that is more substantial than in earlier versions of PowerPoint. The Graphic Data fingerprint will still be maintained through such a conversion. However, the Slide Content fingerprint may sometimes change and the Slide Appearance fingerprint will almost certainly change even though if the slide has not been edited. This is due to how Microsoft Office 2007 converts between earlier versions of PowerPoint. Specifically, there are commonly slight changes to how master reference text and numerous shape attributes are stored that adversely affect the fingerprint.

Outside In Clean Content Fingerprint API

The Outside Clean Content extraction API supports the generation of the fingerprints described above through a set of Boolean options that can be enabled through the API. These options are turned off by default. The GenerateSlideContentFingerprint, GenerateSlideAppearanceFingerprint, and GenerateGraphicDataFingerprint options can be selectively turned on to cause each type of fingerprint to be generated. When a fingerprint option is enabled, the extraction process will include <fingerprint> elements in the extracted output that include an attribute named ‘type’ and an attribute named ‘value’.

The ‘type’ attribute will indicate which of the three fingerprint types is being provided.

The ‘value’ attribute provides the fingerprint as a unique, 128 bit, MD5 hash. When the extraction OutputType is set to ToXML the value is provided as a xsd:hexBinary attribute. When the extraction OutputType is set to ToHandler the value will be provided as a 16 byte array through the startFingerprint method.

The Slide Content and Slide Appearance fingerprints are always a child of a <slide> element. The Graphic Data fingerprint is always a child of an <embeddedcontent> element of type Graphic. Care should be taken to track the depth of fingerprints in the element hierarchy if recursion into embeddings is enabled.