Oracle® Fusion Middleware

Outside In Technlogy Clean Content Release Notes

Release 8.5

F11006-04

October 2023

Clean Content Release Notes

This Clean Content release notes list all the new features, bug fixes, and enhancements done during the current release and the past releases. See also Outside In Release Notes in this library.

For configuration and SDK related information, see Outside In Clean Content Developer's Guide.

Major changes and fixes in 8.5.7

- OIT 8.5.7 includes customer and security bug fixes on top of OIT 8.5.6.

Major changes and fixes in 8.5.6.01

- A flag-based OOXML (OfficeXML) feature has been introduced which enables you to do the following in your XML files:

  > **Note:**
  >
  > For detailed information, see `SecureOptions.OfficeXMLFeatures`.

  - Identify and/or remove of all CDATA constructs.
  - Identify and/or remove all XML comments within the XML.
  - Identify and/or remove all XML processing instructions within the XML.
  - Identify and/or remove external entity references within the XML.
  - Remove leading and trailing whitespaces within the XML.
  - Identify uncommon or unexpected XML namespaces in XML files. These namespaces can now be blacklisted using the Blacklist a namespace option. In the demo application, it can be found under **Set Scrub Option -> Additional Option**.
  - Use canonicalization of `XML.Refer XML.Refer SecureOptions.OfficeXMLCanonicalization`. For more information, see Javadoc in the CleanContent SDK.

  - Create a log file corresponding to each file being processed for removal of XML CDATA, XML Comments, XML Processing Instructions and XML External Entity within the XML.
  - Scrub unknown namespaces within the XML.
  - Rename XML namespace prefixes.
  - Whitelist known namespace prefixes.
  - Identify and scrub unused namespaces.
  - Remove bounding whitespace within text elements.
  - **CUSTOMER CODE CHANGE**: Three new JARS (slf4j-api-2.0.99.jar, slf4j-simple-1.6.99.jar and xmlsec-2.2.3.jar) added at CleanContentSDK/java/lib. Additionally, the CleanContent.jar must be added to the custom project.
  - **KNOWN ISSUES**: Canonicalicazer currently does not canonicalize all XMLs in MS Office files. It canonicalizes Content_Types.xml and all rel file for all MS office files. It also canonicalizes document.xml for .docx files, workbook.xml for .xlsx files, and the presentation.xml file for .pptx or .ppsx files. All other associated XML files such as docProps/app.xml, core.xml, and fontTable.xml, and so on will be canonicalized in a future release.

- **Other new features in this release are:**
  - Scrubbing macros from Excel files.
  - Scrubbing bad images in MS Office 2010 and newer files.
  - New option, SecureOption.ValidateEmbeddedContent, is now available to validate embedded images in MS Office files. Setting this option to true allows the extraction to report OfficeXMLPartDisclosureRisks if it exists in any files. All these masquerading files are treated as rogue elements. Rogue parts are automatically scrubbed whether this option is enabled or disabled as rogue parts serve no known valid purpose.
  - Ability to unhide comment fields.
  - Scrubbing of color obfuscated text from PDF files.
  - Extract font details from Microsoft Excel, PowerPoint, and Word files.
- **Major fixes in this release are:**
  - Enhanced ability to report and remove the color obfuscated text in doc files.
  - Enhanced stability of the processing of XLS binary format files.
  - Enhanced stability of the processing XLSX file format.
  - Improved color obfuscated text scrubbing in PowerPoint.
  - Enhanced image replacement mechanism in PDFs.
  - Clean Content can now extract font details of an MS Office binary format files and PDFs. The font is part of the string extraction of the type StringElementType.Font.
  - Clean Content can now extract font details of MS Office (2007 and above) files. The font is part of the string extraction of the type StringElementType.Font.

  - Clean Content can now unhide the hidden comment box in Excel 97 through the 2003 format when the SecureOption.Comments option is set to scrub.
  - Clean Content now successfully processes MS office files that are created online.
  - Other issues fixed in this release are: 27525264, 29349138, 29880735, 26635735, 29385624, 30570305, 31451521, 31451502, 31451506, 31451487, 27146627, 29255147, 18800498, 32147455

> **Note:**
>
> Clean Content release 8.5. requires Java 8.

Newly Supported File Formats in OutsideIn Technology Release 8.5.5

- Microsoft Word for Windows 2019
- Microsoft Excel for Windows - 2019
- Microsoft PowerPoint for Windows - 2019

New Options in OutsideIn Technology Release 8.5.5

`SecureOptions.BrokenPDFCorrection`: This option allows users to enable correction of PDFs whose internal structure is malformed. This option tries to recover PDFs if Clean Content fails to process the input PDF files. Set this parameter to `true`so that corrected PDFs are placed at `scrubbedDir/CorrectedPDFs` by default. If not at this location, then the corrected PDFs are placed at `extractDir/CorrectedPDFs` for extraction.

> **Note:**
>
> Considering the complex model of PDF documents and various possible erroneous structures, Clean Content may not always be able to successfully correct broken PDFs.

Other Improvements in OutsideIn Technology Release 8.5.5

- *Clean Content can now extract DDE payload and scrub them using the target LinkedObjects*: DDE (Dynamic Data Exchange) is a protocol that sends messages between applications that share data and use shared memory to exchange data between applications. In case of Microsoft Excel, DDE can be used to perform command execution through Excel sheets formulas. In the case of Microsoft Word, DDE is used in the field.
- *Clean Content now successfully scrubs commentIds from Word 2016 files completely*: When a comment is added to a Microsoft Word 2016 files, the application adds an additional XML file to the document file structure (`/word/commentsID.xml`) that previous versions do not add. Clean Content now scrubs this `commentId.xml` as a part of the target Comment.

- Improved multi-threading processing for handling documents with attachments. This improvement is without any performance overhead.
- *Temporary files are cleaned up after a PDF is processed*: Clean Content creates temporary files in the temp directory if the PDF file being processed is large. Earlier these large files used to remain in the temp directory after the processing of the PDFs. Clean Content can now remove these temporary files.
- *Clean Content now successfully extracts and scrubs sensitive hyperlinks in Microsoft PowerPoint 97 files*: Hyperlinks in PowerPoint 97 files are successfully analyzed to check whether they are sensitive. Clean Content scrubs these sensitive hyperlinks as a part of the target `SensitiveHyperlink`.

Changes in Previous Releases

- **2018.3.30**
  - Issue numbers fixed in this release: 25910751,27244381, 24390869, 23215828.
- **2018.2.28**
  - Issue numbers fixed in this release: 25340818, 27117481, 25267266.
- **2018.1.24**
  - Issue numbers fixed in this release: 25910751, 27244381.
- **2015.1.4**
  - Assembly and disassembly now operates on PDF documents as well as PowerPoint document.
  - Issue numbers fixed in this release: 22050820, 23595853, 23749554
- **2015.1.3**
  - IMPORTANT: Added the boolean option TimeoutUsingThreadStop that allows timeout of requests in tight infinite loops. See the new Timeouts section in this document for details and warnings
  - CUSTOMER CODE CHANGE: Major changes to how file are identified and the FileFormat class are coming for Clean Content 2016.1. Hundreds of new formats will be identified. The change should be mostly backwards compatible but expect FileFormat to become an Enum.
  - CUSTOMER ALERT: A display issue has been found when running the demo application under X-Windows. This will be resolved in Clean Content 2016.1.
  - Scrubbing of Color Obfuscated Text in Microsoft Word 2007 and above when the ColorObfuscatedTextRemediation option is set to ColorObfuscatedTextRemediationOption.AdjustColor has been reworked to avoid setting all text in the document to "Auto" color.
  - Added the boolean option SimulatePowerPointAnimationsDuringAssembly. This option applies to the assembly of PowerPoint 2007 and above (PPTX). When set to true, this option will cause slides that originally contained animation to be expanded into a series of slides that simulate the animations by hiding and restoring slide elements to simulate the entrance and exit of animated elements.

  - PDF files embedded in Office 2007 and above will now behave correctly after being scrubbed.
  - Issue numbers fixed in this release: 19064521, 19582411, 20663115, 20873121, 20898724, 21115231, 21305814, 21848936, 21954498, 22047140, 22293983, 22377985, 22385328, 22386086, 22390140

- **2015.1.2**

  Issue numbers fixed in this release: 20224163, 20639346, 20890404, 20906536, 21070673, 21084342, 21266459, 21385974, 21975599

- **2015.1.1**
  - Internal-only release.
  - Issue numbers fixed in this release: 20906434.

- **2015.1**
  - CUSTOMER CODE CHANGE: In extracted output, the datacell element (and event) has been removed in favor of type specific elements including numbercell, textcell, datecell and durationcell.
  - Extracted spreadsheet output now includes both the raw data value (in the value attribute) and the data as it appears formatted in the application (as text in the element). For example, a cell that was previously extracted like this, <datacell row="59" column="8" value="2238356.65"/>, is now extracted like this <numbercell row="59" column="8" value="2238356.65">$2,238,356.65</numbercell>.
  - CUSTOMER ALERT: The next major release of Clean Content will require Java 7
  - Removed Outside In Search Export integration. Customers can now use the Outside In Java and .NET APIs directly.
  - Tweaked C include file extracthandlercpp.generated.h to support newer versions of GCC
  - Added a scrub target called HybridExcel9597BookStream
  - Added support for Microsoft Office files in Strict Open XML format
  - Issue numbers fixed in this release: 8206503, 9285746, 9787417, 9787432, 13877473, 20122501, 20205999, 20402739, 20539633, 20561528

- **2013.1.6**
  - Issue numbers fixed in this release: 18461820, 19307501, 19504465, 19508023, 19516940, 19637681, 19773754, 19807436, 19815782

- **2013.1.5**
  - Fixed memory leak on Linux when using C/C++ or .NET APIs. The leak was most noticeable when using the BFSetFileOption function. For each call to this function a number of bytes equal to 2 times the length of the file name were being leaked.
  - Issue numbers fixed in this release: 17866823, 18536642, 18697470, 18808601, 18808807, 18967956, 18995179, 18995293, 19074672, 19165003, 19180563

- **2013.1.4**
  - Added new BFStartupEx function to C API. This allows the path to CleanContent.jar, the path to the JRE, and the JRE options to all be set programmatically. For more information, see **C/C++ Guideline**s shipped as part of the application.
  - Added a new version of SecureHelper.Startup to the .NET API which mirrors the new BEStartupEx function in the C API. For more information, see **.NET Guidelines**s shipped as part of the application.
  - Issue numbers fixed in this release: 18614206, 18553874, 18116944
- **2013.1.3**
  - Issue numbers fixed in this release: 18420341, 18551449, 18298913, 18507762
- **2013.1.2**
  - Issue numbers fixed in this release: 18070491, 18298913
- **2013.1**
  - Customers will see a roughly 70% performance improvement in the analysis of XML-based Microsoft Office 2007-1013 documents, and significant increases in scrubbing and extraction performance.
  - The following scrub or analysis **Targets** were added for this release: AppsForOffice, ExcelDataModel, UnknownXML, PDFAlternateImages , PDFDeprecatedPostscriptObjects, PDFAlternatePresentations, PDFWebCaptureInformation, PDFLegalAttestation, PDFDigitalSignatures, PDFThumbnailImages, PDFAnnotations, SensitiveContentLinks, XMPMetadataStreams, and GPSData.
  - The following options were added for this release: PDFMinimumImageDimensionRequiredToProcess. See API for details.
  - The option SensitiveHyperlinksRegex was renamed to SensitiveLinksRegex and now applies to both the SensitiveHyperlinks and SensitiveContentLinks targets.
  - Extraction and scrubbing of the Open Office XML format (Microsoft Office) now includes processing of all choices in Alternate Content elements.
  - Added support for XMP embedded in PDF documents.
  - Added support for identification of XMP, RDF, WAV and AIFF file formats.
  - CUSTOMER ALERT: The last public update of Java 6 was Feb 2013. An upcoming version of Clean Content will drop support for Java 6 and require Java 7. Please contact your Oracle sales representative if this change will create an issue for your product.
  - Issue numbers fixed in this release: 9463630, 13377241, 13425435, 13475829, 13582954, 13956841, 14239187, 14509582, 14763715, 15893487, 16403744, 16561941, 17050811, 17337929, 17471338, 17633976
- **2012.1.2**
  - Added clarification on custom property modification to this document.

  - Issue numbers fixed in this release: 16895155, 16914757

- **2013.1.6**

  Major changes and fixes for **2012.1.1**

  - Verified against Microsoft Office 2013. Verification means that Office 2013 documents are correctly extracted and that scrubbed documents of any supported Microsoft Office version will open correctly in Office 2013. Support for Office 2013 schema changes, additional scrub targets, etc. will be available in a later version of Clean Content.
  - Clarified the validity of output in exception and error cases.
  - Scrubbing of Unknown Fields in Word 2007-2013 files now works as documented
  - The demo application now uses net.java.awt.Desktop class to open browser windows
  - Issue numbers fixed in this release: 11677587, 16284559, 16297194, 16396281, 16515824, 16516432, 16601120, 16673293

- **2013.1.6**

  Major changes and fixes for **2012.1**

  - Scrubbing of PDF documents as well as many PDF related targets have been added including ClippedText, PDFActions (and it's sub-targets like PDFJavaScriptActions), and PDFPrivateApplicationData.
  - Integration with Oracle's Outside In Search Export has been added. Customers who choose to purchase that product will be able to extract text, metadata and structure from over 500 additional file formats through the Clean Content API. For more information on the API, see APIs listed as part of the application..
  - XML processing is now more forgiving of XML that does not follow its XML Schema. An InvalidXML target and logged warnings have been added to alert developers to invalid XML elements in case they require strict schema conformance
  - PDF seach hit highlighting support has been removed from the demo application. See related Customer Alert under 2011.2 changes
  - Clean Content now requires Java 6 or higher
  - The XML schema used for extraction has been extended with many additional elements and attributes in the message, archive and database areas to support the Outside In Search Exportintegration
  - Fixes were made to how XML output is produced allowing use of Saxon and other third-party XSLT processors that plug-in to the Java JAXP infrastructure.
  - Added UninitializedDocfileData target
  - CUSTOMER ALERT: In an upcoming revision the Clean Content options that are enumerations will become real Java Enums in the Java API. This may end up being transparent to the customer's code but please expect this change.

- Issue numbers fixed in this release: 13636289, 13940798, 13987464, 14038681, 14051875, 14064291, 14120406, 14188819, 14193959, 14198083, 14251214, 14304898, 14527636, 14539010

- **2013.1.6**

  Major changes and fixes for **2011.2.7:**

  - Tweaked the XML structure of scrubbed Word documents (.docx) to work around a bug in Apple's iPhone/iPad email attachment preview feature
  - Normalized the definition and implementation of the Weak Protections target
  - Added more descriptive reporting of invalid XML elements
  - Resolved significant issue in PowerPoint .pptx disassembly
  - Issue numbers fixed in this release: 13068853, 13796206, 13972531, 14019099, 14051609, 14078741, 14128837

- **2011.2.6**

  - Scrubbing of ImplicitRef, Formula and Unknown fields now supported in Word
  - Added Weak Protection support in Excel 2007/2010. See the Weak Protections target description for more information.
  - Added Custom XML support for Word 97-2003. See the Custom XML target description for more information.
  - Tracked Changes in Word 2007/2010 that indicate moves are now supported for scrubbing
  - Issue numbers fixed in this release: 13041497, 13257351, 13388342, 13425420, 13576875, 13641062, 13642206, 13682781, 13682823, 13465121, 13776088

- **2013.1.6**

  Major changes and fixes for **2011.2.5:**

  - CUSTOMER ALERT: The next major version of Clean Content will require Java 6 or above, Java 5 will no longer be supported
  - Fields without structured data (PRIVATE, RD, TA, TC and XE) will be scrubbed correctly in Microsoft Word 97, 2000, XP & 2003 documents
  - Unknown fields (that is fields that don't have a structured data type and don't match any field name) will be scrubbed in Microsoft Word 97, 2000, XP & 2003 documents based on the new SecureOptions.Fields.Unknown field type
  - Added missing GetEnumResult method to BFSecureResponse class in the C/C++ API
  - Added missing GetResult(EnumOption) method to the SecureResponse class in the .NET API
  - Fixed declaration of .NET enumeration option values from 'static' to 'const' so they can be used in a switch statement
  - Updated C/C++ and .NET sample code in the **Response** section of this document to use the new ProcessingStatus option

  - In previous versions the SecureResponse class had references to other large objects trees causing the collection and later processing of a large number of these objects to be very memory intensive. The code has be refactored to produce a very light SecureResponse object allowing collection of a large number of responses without huge memory overhead.
  - In version 2011.2.4 the UNKNOWN FileFormat was accidentally moved to be a sibling of the ALL FileFormat instead of its child, this has been corrected
  - Issue numbers fixed in this release: 13345226, 13332027, 13325655, 13253127, 13103312, 13087267, 13082906, 13042424, 13361562, 13345226
- **2011.2.4**
  - CUSTOMER ALERT: The behavior of property modification has been altered so that the actions None and Scrub when specified for individual properties will propogate down into embedded documents. The Replace and AddOrReplace actions will behave as before in that they will not affect properties in embedded documents.
  - Changed processing of Word and PowerPoint XML element types to allow certain unusual documents to be processed correctly.
  - Changed processing of Microsoft Office 2007-2010 documents to recover from malformed VML
  - Fixed the C function BFHelperGetDateString
  - Fixed infinite loop cases in certain malformed Microsoft Office documents
- **2013.1.6**

  Major changes and fixes for **2011.2.3**:
  - Changed methodology for scrubbing Comments in Word 97-2003 documents. The earlier method could, in exceptional circumstances, produce documents unreadable by specific versions of Word. The new method fixes this issue while still passing Word 2010's Office File Validation test.
  - Fixed small issues in property categorization and scrubbing
- **2011.2.2**
  - Added array allocation checks in generated code to avoid OutOfMemoryError being thrown in cases where malformed documents drive the code with unexpected data
  - Removed all use of the String object's intern method to avoid filling PermGen space and causing OutOfMemoryError
- **2011.2.1**
  - Small, document specific fixes to the PowerPoint 97-2003, PowerPoint 2007/2010, Word 2007/2010 and PDF transforms
  - BMP is now a valid image replacement format in Office 2007/2010 documents
- **2011.2**
  - CUSTOMER ALERT: Two fixes have been made to the C API (and therefore the C++ and .NET APIs built on top) to solve issues where certain documents processed through these API's could cause untrapped segmentation faults in

the Java Virtual Machine. All customers using the C, C++ or .NET APIs on Clean Content 2010.1 or above should upgrade ASAP to 2011.2. A patch to Clean Content 2011.1 for this issue is also available on My Oracle Support at Doc ID 1307556.1. This is not an issue in the Java API.

  - CUSTOMER ALERT: In Acrobat X, Adobe Systems has chosen to remove the "hit highlighting" feature Clean Content can leverage to highlight search terms in PDF documents. This feature was available but disabled in Acrobat 9. These changes call into question the usefulness of the GenerateAcrobatHighlightPositions option in Clean Content.
  - Added OfficeXMLPartValidation option and associated analysis targets: OfficeXMLUnanalyzedParts, OfficeXMLUnexpectedParts, OfficeXMLAlternateContentParts and OfficeXMLRogueParts. Details may be found in the **Microsoft Office Open XML Support** in the Technical Note as part of the application.
  - CUSTOMER CODE CHANGE: The little used getCharBuffer method in the ElementHandler interface has been modified. It now takes a single parameter specifying the minimum allowable size of the returned character buffer.
  - In Microsoft Word and PowerPoint 2007/2010 formats the line break element is now being handled correctly, avoiding cases where text separated by line breaks world run together in the extracted output.
  - Several enhancements have been made to PDF extraction in the areas of vertical text and space inference.
  - Added ability to scrub of PRIVATE and BIDIOUTLINE fields in all supported versions of Microsoft Word
  - Fixed issue where images and embeddings with the hidden character property were not being scrubbed when the HiddenText target is set to SCRUB in Microsoft Word 2007/2010.

- **2013.1.6**

  Major changes and fixes for **2011.1**:

  - Added support for encrypted Microsoft Office and PDF documents and associated PasswordList and DecryptionStatus options. Details may be found in the **Encrypted Document Support** in the Technical Note as part of the application.
  - Added ToTextEncoding option allowing either UTF-16 or UTF-8 encoding when extracting text with OutputType set to ToText. In addition, some small improvements were made to the extracted text itself.
  - Added ProcessingProblem response to make response handling easier to write. The WasProcessed, WasIdentified, WasSupported, WasTimeout and WasException responses are still available but deprecated.
    The **Response** section has been updated to reflect these changes.
  - Added support for field modification in Office 2007/2010 documents.
  - Added ability to change the starting page number of a Microsoft Word 2007/2010 document through the new ChangeStartingPageNumber (boolean) and StartingPageNumber (integer) options.

  - Added analysis and scrubbing of document variables in Microsoft Word 2007/2010 under the DocumentVariables target.
  - Moved analysis and scrubbing of document variables in Microsoft Word 97-2003 from the MacrosAndCode target to the DocumentVariables target.
  - CUSTOMER ALERT: Microsoft Office 2010 added a security feature called **Office File Validation**. This feature (enabled by default) adds numerous document integrity tests when opening pre-Office 2007 binary formats for Word, Excel and PowerPoint. This is an effort by Microsoft to prevent specially crafted pre-Office 2007 documents from compromising system security. Unfortunately this feature will also flag some documents that have been scrubbing by earlier versions of Clean Content causing the Office 2010 application to enter Protected View and warn the user. Note that such documents open correctly and with no warning in all previous versions of Microsoft Office. The Clean Content scrubbing process has been modified so that pre-Office 2007 documents scrubbed by this version will not receive this warning in Office 2010.
  - CUSTOMER CODE CHANGE: Customers are able to apply an XSLT transform to the extracted XML output using the TransformResult (boolean) and ResultTransform (file) options. Up until this version the XSLT provided was required to treat XML attributes in the extracted output as "qualified" even through the untransformed XML output shows them as "unqualified". This issue has been fixed and XSLT passed to the ResultTransform option should be modified to treat XML attributes as "unqualified".
- **2010.2**
  - Full support for Office 2010. Details may be found in the **Microsoft Office Open XML Support** in the Technical Note as part of the application.
  - Changed the way the Comments target is scrubbed out of Microsoft Word 97-2003 documents to avoid a warning when opening such scrubbed documents in Microsoft Word 2010.
- **2010.1**
  - Added C/C++ API support for 64 bit versions of Windows and Linux on the x86-64 architecture.
  - Added .NET API support for 64 bit versions of Windows on the x86-64 architecture.
  - Numerous PDF extraction and analysis features have been added with this release. These include the detection of JavaScript (Macros and Code) and Incremental Updates (Fast Save Data) and the extraction of Document Outlines, Thumbnail Images, Article Thread Information, Interactive Forms, and XFA Forms. Enhancements have also been made for AES Encrypted documents, PDF Portfolio support, and Malformed PDF document parsing. Additionally, numerous heuristic enhancements have been made to the line detection and character mapping algorithms that further improve extraction from poorly crafted PDF documents. Details may be found in the **PDF Extraction and Analysis Support** in the Technical Note as part of the application.
  - Processing of PDF documents with AES 256 bit encryption requires the Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy Files.

    See the **PDF Extraction and Analysis Support** in the Technical Note as part of the application.

  - CUSTOMER CODE CHANGE: The C interface (and by extension the C++ and .NET interfaces built on top) has been largely rewritten to be thinner, faster and more portable. As part of this rewrite several small changes have been made to the C API itself, these include...

    * Access to actions and new values for scrubbing and changing Properties and Word Fields has been changed. The former API defined a structure like SecureOptions_Properties_Author which contained option id's for the action and newValue on the Author Property. The new API simply defines static constants for the two option ids. So all a developer needs to do in their code is change things like SecureOptions_Properties_Author.newValue to SecureOptions_Properties_Author_newValue.

    * Arrays listing the options of various types used to be directly available through arrays like AllScrubOptions or AllStringOptions. These arrays and the count of their elements are now available only through function calls like BFGetAllScrubOptions and BFGetAllStringOptions.

    * The two changes above (and other changes not visible in the API) were undertaken to allow multiple source files to include secureapi.h without the cumbersome #define BFDONTDEFINE method used in earlier versions.

  - The C interface library is now buildable by the OEM customer in order to support the C/C++ API on any reasonable unix-like platform with a Java 1.5 JDK and a recent GNU toolchain. This also allows the customers using Linux to relink libCleanContentAPI.so with standard library versions of their choice.

  - This release ships static (libCleanContentAPI.a) versions of the C interface library for Linux as well as the shared (libCleanContentAPI.so) versions.

  - As of Sun JVM version 1.5 the requirement on Linux to add the Java library paths (like /usr/java/j2re1.4.2_05/lib/i386 and /usr/java/j2re1.4.2_05/lib/i386/client) to an applications library search path (either using -rpath at link time or LD_LIBRARY_PATH at run time) has been removed.

  - The granularity of the RequestTimeout option has been improved to roughly half a second. Previous versions had a timeout resolution of about 10 seconds.

- **2009.1.1**

  - Fixed problem with certain PDF documents were never timing out

  - Fixed .NET API issue where dates out of a certain range cause an untrapped exception

  - Fixed issue where the Content Properties target was not being set for Microsoft Office 2007 files containing content properties

  - Fixed issue where the Headers and Footers target was not being set for Microsoft Word 2007 documents containing headers or footers

  - Fixed issue where Open Office 3.1 would not open Microsoft Office 2007 documents scrubbed by Clean Content

  - Updated the Targets section of this guide to more clearly list the formats that can contain each target
  - CUSTOMER CODE CHANGE: The ExportMaximumReplacementSize option retrieved during
    the startEmbeddedContent and processEmbeddedContent methods may now be 0 (zero) indicating that any size replacement is allowable.
  - Fixed issues with C, C++ and .NET documentation
- **2009.1**
  - Full support for extraction, analysis and scrubbing of Office 2007 Word (.docx, .docm, .dotx, .dotm), Excel (.xlsx, .xlsm, .xltx, .xltm, .xlam), PowerPoint (.pptx, .pptm, .potx, .potm, .ppsx, .ppsm, .ppam) documents.
  - Full support for Microsoft PowerPoint 2007 assembly and disassembly.
  - Full support for extraction and extensive analysis support of Office 2007 Excel Binary (.xlsb) documents. Property and Macro scrubbing is also included for this format.
  - CUSTOMER CODE CHANGE: Added ScrubbedFormat option and new code examples in the **Response** section. It is critical that existing customers that would like to correctly scrub Office 2007 formats read this new information and update their code accordingly. In short, Office 2007 files may need to have their extensions changed when scrubbed. For example, when scrubbing macros from a .docm file it must be changed to a .docx file or it will not open in Word. The ScrubbedFormat option allows the developer to understand when this is required and change the name accordingly. See the **Response** section for details.
  - Added the CustomXML and StructuredDocumentTags scrub targets.
  - CUSTOMER CODE CHANGE: The hyperlinkbegin and hyperlinkend elements in the extraction schema have been changed as follows. The url and sensitive attributes have been removed in favor of always using a contentref or linkedcontent child. This normalizes how hyperlinking and external references are accomplished in the schema.
  - CUSTOMER CODE CHANGE: The signatures of the StartEmbeddedContent and ProcessEmbeddedContent methods in BFBaseElementHandler have changed. In both cases the exportOptions parameter is now part of the BFEmbeddedContentElement structure instead of being an additional parameter. In the case of C++ users extending BFBaseElementHandler the exportOptions handle will now need to be used with the C option functions or used to create a BFOptionSet. See the **Embedding Replacement** sample code for an example of this.
  - Java 1.4 is no longer supported. Clean Content now supports only Java 1.5 and above.
  - The demo application now includes an improved single file processing UI and allows setting of nearly all options available in the API.
  - In a change in behavior, SecureRequest will now delete the newly created scrubbed document when processing fails. This avoids leaving partially constructed, corrupt documents floating around. This is only the case where

the ScrubbedDocument option is set using a File object in Java, a path in C/C++ or a FileInfo object in .NET.

  - FEATURE CHANGE: The ScrubInPlace option has been removed. Unlike the binary Microsoft Office file formats, the XML Microsoft Office file formats need to be completely rewritten in order to be correctly scrubbed. This requirement makes ScrubInPlace impossible or at least highly inefficient to implement. This, along with other more technical issues, have led us to remove this feature from the API.

- **2008.1.5**
  - Fixed PowerPoint 2000 assembly issue
- **2008.1.4**
  - Added image extraction from Adobe Acrobat PDFs
  - Various fixes and enhancements for Adobe Acrobat PDF .
  - CUSTOMER CODE FLASH: The **2008.2** release will operate only on Java 1.5 and above. Java 1.4 will no longer be supported.
- **2008.1.3**
  - Various fixes for PowerPoint assembly/disassembly
- **2008.1.2**
  - Added property extraction support for Office 2007 XML (MSOOX) files
  - Miscellaneous fixes for a number of file formats including several for PDF
  - CUSTOMER CODE CHANGE: This release includes an update to SimpleChannel and its C, C++ derivatives to include a truncate method. Any developer providing or consuming data through a SimpleChannel (or its C, C++ derivatives) will need to update their implementation and recompile to work with the 2008.1.2 release. The equivalent functionality in C# is provided through the .NET Stream class and C# developers should need no changes to their code.
- **2008.1.1 (Oracle internal only)**
  - Added property modification support for Office 2007 XML (MSOOX) files. Currently only supports addition/deletion/modification of custom properties
  - Added PowerPoint slide fingerprinting support (see technical note at end of this document)
  - Added extended support for identifying Office documents with incorrect CLSIDs
  - Demo application now supports additional PDF, PowerPoint and Excel options
  - CUSTOMER CODE FLASH: The 2008.2 release will update SimpleChannel and its C, C++ and C# derivatives to include a truncate method. Any developer providing or consuming data through a SimpleChannel (or its C, C++ or C# derivatives) will need to update their implementation and recompile to work with the 2008.2 release.
- **2008.1**
  - Added PDF extraction with hit highlighting support

  - Added logging support in Java
  - Published XML schema for extracted XML and ElementHandler events
  - CUSTOMER CODE CHANGE: Addition of a published XML Schema has lead us to do some cleanup on the XML output (and consequently the ElementHandler events). These changes should not affect most customers but anyone post-processing the XML or coding for specific series of ElementHandler events should review the new XML Schema and modify their code, xslt, etc. accordingly. In the future we may refine and tweak the XML Schema but no drastic changes are envisioned in the short to medium term.
  - Fast Save data in Word can now be exported for processing
  - Added PDF technical notes

- **2013.1.6**

  Major changes and fixes for **2007.2.1**:

  - Added PowerPoint Assembly/Disassembly

- **2013.1.6**

  Major changes and fixes for **2007.2**:

  - Added ExtremeCells, ExtremeIndenting, ExtremeObjects and OverlappedObjects targets
  - Added regex-based Header/Footer removal and modification
  - Added .NET API
  - Property modification and scrubbing will now work on any Microsoft Docfile format document not just the formats (Word, Excel and PowerPoint) where more comprehensive scrubbing occurs. This mean that property modification can now occur on formats like Microsoft Visio and others that use the DocFile container format.
  - Reworked documentation and extended sample code

- **2007.1**

  - CUSTOMER CODE CHANGE: Property and Field scrubbing/modification API has been reworked to avoid specialized methods/functions and decrease the API footprint. If you use Property or Field scrubbing/modification you will need to change your code to conform to the new API. This is true of both the C/C++ and Java APIs.
  - JAVA CUSTOMER CODE CHANGE: All trapped and many untrapped exceptions are now being caught and reported through a TransformException (a subclass of IOException). As a result, developers now need only catch IOException on the request's execute method. Existing code will work if it followed the earlier guidelines but many of the catch blocks will never be reached.
  - Added Size Obfuscated Text scrub target
  - Added Color Obfuscated Text scrub target
  - Added table in table support in Microsoft Word

- Removed Scenario Comments scrub target in favor of new Scenarios scrub target
- The WasIdentified, WasSupported, WasException and WasTimeout boolean results have been added to improve error reporting.
- Demo application now has GUI support for all Clean Content options.
- Demo application now provides more statistics about groups of documents being processed.
- Demo application now provides better reporting of exceptions and other errors.
- A timeout mechanism has been added to interrupt very long or looping requests. See the RequestTimeout option for details.
- The C/C++ API now has a setting that allows easier debugging of the Java startup process that occurs during BFStartup function or BFSecureRequest.Startup method.
- In Microsoft Word the bit that indicates a document is Fast Save is now cleared when Fast Save Data is scrubbed.
- Rebranded to Oracle
- SDKs now ship with the Sun's 1.4.2_13 JRE which resolved some vulnerabilities in earlier JREs. See http://www.us-cert.gov/cas/techalerts/TA07-022A.html

- **2006.2**
  - Added ability to remove, modify and add document properties in Microsoft Office documents.
  - Added ability to remove hyperlinks based regular expression matching.
  - Rebranded to Stellent
- **2006.1**
  - CUSTOMER BUILD CHANGE: Rebranded to Clean Content SDK. JAR, DLL and SO file names have changed. Customers should inspect and update their build process.
  - Performance on all formats has roughly doubled.
  - Added the ability to extract the text and structure from documents as they are processed.
  - Added the ability to export embedded objects for further processing or display.
  - Added the ability to recur into embedded objects for scrubbing, analysis, extraction and/or export purposes.
  - Added the ability to replace images and embedded objects.
  - Added option to extract only properties in order to increase performance for those customers needing only document properties.
  - Extracted XML now includes better delineation and referencing of items like headers, footers, footnotes, etc.

  - Several new scrub targets have been added including: Alternative Text, DocumentVariables, Fields (including individual field type removal and text replacement) and Meeting Minutes.
  - Added the ability to 'unhide' hidden cells in Excel.
  - Added the ability to define a global default scrub behavior and changed the default value of all scrub target options to use this default.
  - The Secure sample application can now extract to XML as well as produce an HTML report.
  - JAVA CUSTOMER CODE CHANGE: In order to avoid difficulty in writing Java code all options available in Secure are now available in the SecureOptions class. The ExtractOptions and SharedOptions classes still exist but Stellent recommends using SecureOptions class instead.
  - C/C++ CUSTOMER CODE CHANGE: In order to support both Secure's and Extract's options in the C/C++ interface the option names which were formerly things like SourceDocument, AuthorHistory and OutputType have now been namspaced by prefixing with 'SecureOptions_'. So the above are now called SecureOptions_SourceDocument, SecureOptions_AuthorHistory, SecureOptions_OutputType. In addition, a C++ cover class on these options has also been added so C++ customers can use BFSecureOptions::SourceDocument, BFSecureOptions::AuthorHistory and BFSecureOptions::OutputType. Customer's C/C++ code will need to be modified to reflect this change.
- **2005.2**
  - Added C/C++ API.
  - Added several new ways to provide the SourceDocment and ResultDocument options including as a block of memory, as a FileChannel and as a SimpleChannel (defined in this SDK) which allows arbitrary IO redirection.
  - Many small fixes to analysis, scrubbing and reporting behavior.
- **2005.1**
  - The information content and look of the reports has been improved substantially.
  - Many small fixes to analysis, scrubbing and reporting behavior.
  - Many small fixes to the BitformSecureSDK application.
  - Performance on Microsoft Word documents has been almost doubled.
  - The TransformReport option now has its documented effect.

Documentation Accessibility

For information about Oracle's commitment to accessibility, visit the Oracle Accessibility Program website at `http://www.oracle.com/pls/topic/lookup?ctx=acc&id=docacc`.

Access to Oracle Support

Oracle customers that have purchased support have access to electronic support through My Oracle Support. For information, visit `http://www.oracle.com/pls/topic/lookup?ctx=acc&id=info` or visit `http://www.oracle.com/pls/topic/lookup?ctx=acc&id=trs` if you are hearing impaired.

Oracle Fusion Middleware Outside In Technlogy Clean Content Release Notes, Release 8.5
F11006-04