Core Engine Product Description: Core 2

White Paper

© 2019 Glasswall Solutions Ltd ALL RIGHTS RESERVED

Information contained herein is the property of Glasswall Limited and is proprietary and confidential.

Glasswall Solutions Ltd.

(e): support@glasswallsolutions.com

Creation Date – 28/01/15

Version – 1.8

**Document History**

Table 1: Document Change History

|  |  |  |  |
| --- | --- | --- | --- |
| *Issue Date* | *Issue Number* | *Author* | *Description* |
| 28/01/15 | 0.1 | JH | Initial Draft |
| 04/02/15 | 0.2 | JH | Initial review comments |
| 12/02/15 | 0.3 | JH | SH review comments |
| 20/02/15 | 0.4 | JH | Add reference to non-conforming database |
| 02/07/15 | 0.5 | JH | Changed file size limit, took out reference to Linux, added ref to GIF 87 and interlace images |
| 02/07/15 | 0.6 | JH | VT Review comments |
| 10/08/15 | 0.7 | JH | Coorect PDF switch setting |
| 25/08/15 | 0.8 | JH | Remove Launch action type |
| 12/10/15 | 0.9 | JH | Remove Thread from all Actions |
| 29/04/16 | 1.0 | AP | Updated to include new content management switches |
| 06/07/16 | 1.1 | AP | Updated file types supported list |
| 17/01/17 | 1.2 | EBW | Updated to reflect removal of the non conforming images content management switch |
| 17/05/17 | 1.3 | JH | Rebranded and removed Section 7 |
| 19/06/17 | 1.4 | EBW | The *determine file type* description added |
| 04/08/17 | 1.5 | EBW | *Dynamic Data Exchange* content management switch added |
| 11/09/17 | 1.6 | EBW | *Content Export and Import Modes* added |
| 10/04/18 | 1.7 | EW | Issue ID retrieval information added and updated content management section |
| 28/03/19 | 1.8 | EW | Specification version number updates, file size limit updates, supported operating systems added. |

**Copyright and Contact Details**

The copyright in this work is vested in Glasswall Solutions Ltd, and the document is issued in confidence for the purpose for which it is supplied. It must not be reproduced in whole or in part or used for tendering or manufacturing purposes except under agreement or with the consent in writing of Glasswall Solutions Limited and then only on condition that this notice is included in any such reproduction. No information as to the contents or subject matter of this document or any part thereof arising directly or indirectly there from shall be given orally or in writing or communicated in any manner whatsoever to any third part being an individual firm or company or any employee thereof without the prior consent in writing of Glasswall Solutions Limited.

© Glasswall Solutions Limited 2019

If there are any questions related to this document, these should be addressed to:

Glasswall Support

Glasswall Solutions Limited

e-mail: [support@glasswallsolutions.com](mailto:support@glasswallsolutions.com)

**Contents**

[1. Introduction 6](#_Toc4770340)

[1.1 Supported Operating Systems 6](#_Toc4770341)

[1.2 Key File Processing Modes 6](#_Toc4770342)

[1.2.1 Analysis Mode 6](#_Toc4770343)

[1.2.2 Protect Mode 7](#_Toc4770344)

[1.3 Content Management 7](#_Toc4770345)

[2. Analysis Mode 8](#_Toc4770346)

[2.1 Overview 8](#_Toc4770347)

[2.2 Analysis Process 8](#_Toc4770348)

[2.3 Sample Analysis Reports 9](#_Toc4770349)

[3. Protect Mode 11](#_Toc4770350)

[3.1 Overview 11](#_Toc4770351)

[3.2 Protect Process 11](#_Toc4770352)

[3.3 Protected Files 11](#_Toc4770353)

[4. Automatic Corrections 12](#_Toc4770354)

[4.1 Overview 12](#_Toc4770355)

[4.2 Remediation 12](#_Toc4770356)

[5. Configuration Management 13](#_Toc4770357)

[5.1 Overview 13](#_Toc4770358)

[5.2 XML Reports with different Content Management Policies 13](#_Toc4770359)

[5.2.1 Allow 13](#_Toc4770360)

[5.2.2 Disallow 14](#_Toc4770361)

[5.2.3 Sanitise 14](#_Toc4770362)

[5.3 PDF 15](#_Toc4770363)

[5.4 Images 15](#_Toc4770364)

[5.5 Binary and XML Office 16](#_Toc4770365)

[6. File Types Supported 17](#_Toc4770366)

[6.1 Overview 17](#_Toc4770367)

[6.1.1 Supported File Sizes 17](#_Toc4770368)

[6.2 PDF 17](#_Toc4770369)

[6.2.1 Processing Specific to PDF in Protect Mode 17](#_Toc4770370)

[6.2.2 Specification Validated Against 17](#_Toc4770371)

[6.3 Images 18](#_Toc4770372)

[6.3.1 JPG 18](#_Toc4770373)

[6.3.1.1 Specification Validated Against 18](#_Toc4770374)

[6.3.2 GIF 18](#_Toc4770375)

[6.3.2.1 Specification Validated Against 18](#_Toc4770376)

[6.3.3 PNG 18](#_Toc4770377)

[6.3.3.1 Specification Validated Against 18](#_Toc4770378)

[6.4 Binary Office 18](#_Toc4770379)

[6.4.1 Processing Specific to Binary Office in Protect Mode 18](#_Toc4770380)

[6.4.2 DOC 18](#_Toc4770381)

[6.4.2.1 Specifications Validated Against 18](#_Toc4770382)

[6.4.2.2 Non Supported Specifications 19](#_Toc4770383)

[6.4.3 XLS 19](#_Toc4770384)

[6.4.3.1 Specifications Validated Against 19](#_Toc4770385)

[6.4.3.2 Non Supported Specifications 19](#_Toc4770386)

[6.4.4 PPT 19](#_Toc4770387)

[6.4.4.1 Specifications Validated Against 19](#_Toc4770388)

[6.4.4.2 Non Supported Specifications 19](#_Toc4770389)

[6.5 XML Office 20](#_Toc4770390)

[6.5.1 Processing Specific to XML Office in Protect Mode 20](#_Toc4770391)

[6.5.2 Specifications Validated Against 20](#_Toc4770392)

[6.5.2.1 Office 2010 20](#_Toc4770393)

[6.5.2.2 Office 2013 20](#_Toc4770394)

[6.5.3 Non Supported Specifications 20](#_Toc4770395)

[7. Determine File Type 21](#_Toc4770396)

[8. Issue ID Information Retrieval 22](#_Toc4770397)

[9. Content Export And Import 23](#_Toc4770398)

[9.1 Exportable Content 23](#_Toc4770399)

[9.2 Importable Content 23](#_Toc4770400)

# Introduction

Glasswalls unique patented approach provides the reassurance of total document security within normally communicated documents. This is achieved principally using a patented regeneration capability whereby only known good data is allowed to be generated to a sanitised, safe file.

The engine has two broad modes of operation: an analysis mode whereby a detailed report highlighting content that is contained within a file is highlighted and a regeneration mode whereby the file is analysed and a safe, substitute copy is made. The Glasswall library operates in one of two modes: analysis or protect. The analysis mode enables the content of documents to be assessed with an XML report generated to summarise the non-conformances and potential threats. The Protect mode addresses the issues that are identified, returning an updated document with the remedies and sanitisation changes necessary to remove threats incorporated in the original document.

This document provides the specification for Glasswall’s core document security engine. This core engine contains an Application Programming Interface (API) within a Dynamic Link Library (DLL) on Windows platforms and a Shared Object (SO) on Linux platforms.

The API provides the following functionality:

* The production of XML reports detailing deep analysis of documents.
* The management of documents against a configurable content management policy.
* The configuration of the content management policy.

## Supported Operating Systems

**Windows**

The application has been successfully tested on the following Windows versions:

* Windows 7
* Windows 10
* Windows Server 2008
* Windows Server 2012

Note: The C++ 2017 Redistributable Package (x64) must be installed on each system.

**Linux**

The application has been successfully tested on the following GNU/Linux distributions:

* Red Hat Enterprise Linux 6.9
* Red Hat Enterprise Linux 7.3
* Red Hat Enterprise Linux 7.2

Note: glibc-2.12 or later and freetype 2.7 or later must be installed.

## Key File Processing Modes

### Analysis Mode

In Analysis mode an XML report is generated for each file which contains ‘Content’ Items (structures found in the file), ‘Issue’ Items (structures that do not match the specification), ‘Sanitisation’ Items (items that would be taken out as per the content management policy) and ‘Remedy items’ (structures automatically corrected back to specification).

### Protect Mode

In Protect Mode the input file is decomposed as per Analysis Mode but the file is then regenerated using the Content Items and remedied Content Items. As part of this process the Content Management rules are applied so that certain ‘Sanitisation’ Items are removed (e.g. macros) from the regenerated file so only known good items are present in the new file.

## Content Management

For each file type processed there is a set of content management switches that either ‘Allow’ element types which remain in the regenerated document, ‘Disallow’ element types (the document is then marked as non-conforming) or Sanitise element types which are removed from the managed document.

In many cases, supported file types are not under the control of content management options as Glasswall will process, validate and regenerate the embedded files by default. Files that are unrecognised or unsupported by Glasswall are generally under the control of content management policy settings unless deemed too risky at which point these items are removed by default.

The following file types have content management policy support:-

* PDF
* Office File formats (XML and Binary)

For file formats that do not have content management policy options, Glasswall will automatically remove any content deemed potentially malicious.

# Analysis Mode

## Overview

In Analysis Mode a file-type agnostic description of the data is logged to an XML report. The structure of the XML report is defined by an Analysis Report XSD designed to make this as easy as possible to parse and process.

The analysis report contains:

* Document Statistics.
* Document Summary – file type and size.
* Content Management Policy - the settings of content management switches that have been applied the document.
* Content Groups – the main grouping of content in the document.
* Content Items – the low-level structures within the document.
* Issue Items – content items that do not match the manufacturer’s specification.
* Sanitisation Items – the content management items that would be sanitised if the document was processed in Protect Mode.
* Remedy Items – automatic corrections applied to the document in order to bring it in line with the manufacturer’s specification.

Depending on the configuration management settings applied to the individual content management switches, the content controlled by the switches is reported either as a Sanitisation Item, Issue Item or a Content Item.

## Analysis Process

The Glasswall library receives a document through a published API and passes it through a number of process cycles. The output from each cycle becomes the input for the next hence maintaining a level of separation between processes. Analysis of the document occurs in each of the cycles. Early cycles elicit the structure of the document and the sizes of its constituent parts. The later cycles are concerned with conducting syntactic and semantic checks which identify possible sources of risk, out-of-range fields or malformed structures.

Where elements of the document are compressed, these are expanded and the results assessed, analysed and verified. This enables the analysis report published at the end of the process to give a thorough assessment of the contents and structure of the document. By stepping through the sanitization and remedying processes Glasswall is able to provide an accurate report of the actions that could be carried out by Glasswall’s regeneration functionality to make the document conformant with the specification.

During each cycle the document being processed is transformed into Glasswall’s own internal representation. This simplifies the parsing and traversing processes and helps provide isolation. As the analysis process navigates through the Glasswall structures the detailed checks are not only made on individual components but also at a higher level between the components of the documents. These higher level checks enable the semantic structure and consistency of the document to be properly verified.

The analysis aspects of the Glasswall functionality provide two forms of reporting. The principle output of the analysis process is the analysis report. This is an XML document that enables the detailed information generated by Glasswall to be interrogated and interpreted by third party applications. The secondary output provided is an engineering report which is a very technical report providing detailed information of the analysis process in an ASCII log format.

## Sample Analysis Reports

This section contains an abridged Glasswall Analysis Report containing all the principle elements of a typical report. This particular example is based on a PDF file.

Each XML report starts with a document summary show below:

<?xml version="1.0" encoding="UTF-8"?>

<gw:GWallInfo xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://glasswall.com/namespace/GWallInfo.xsd" xmlns:gw="http://glasswall.com/namespace">

<gw:DocumentStatistics>

<gw:DocumentSummary>

<gw:TotalSizeInBytes>2293803</gw:TotalSizeInBytes>

<gw:FileType>pdf</gw:FileType>

<gw:Version>PDF-1.4</gw:Version>

</gw:DocumentSummary>

The content management policies that were used on the file are then listed, note all policies for all file types are listed in each report. Some of the PDF content management switch settings are shown below:

<gw:ContentManagementPolicy>

<gw:Camera cameraName = "pdfConfig">

<gw:ContentSwitch>

<gw:ContentName>javascript</gw:ContentName>

<gw:ContentValue>sanitise</gw:ContentValue>

</gw:ContentSwitch>

<gw:ContentSwitch>

<gw:ContentName>acroform</gw:ContentName>

<gw:ContentValue>sanitise</gw:ContentValue>

</gw:ContentSwitch>

<gw:ContentSwitch>

<gw:ContentName>embedded\_files</gw:ContentName>

<gw:ContentValue>sanitise</gw:ContentValue>

</gw:ContentSwitch>

The number of different content groups found in the file (16) along with an example of content item from the first group is shown below:

<gw:ContentGroups groupCount = "16">

<gw:ContentGroup>

<gw:BriefDescription>PDF document has Basic File Section structure instances</gw:BriefDescription>

<gw:ContentItems itemCount = "5">

<gw:ContentItem>

<gw:TechnicalDescription>PDF Header Instances</gw:TechnicalDescription>

<gw:InstanceCount>1</gw:InstanceCount>

<gw:TotalSizeInBytes>15</gw:TotalSizeInBytes>

<gw:AverageSizeInBytes>15</gw:AverageSizeInBytes>

<gw:MinSizeInBytes>15</gw:MinSizeInBytes>

<gw:MaxSizeInBytes>15</gw:MaxSizeInBytes>

</gw:ContentItem>

As the GWFILE\_MANAGE\_REMOVE\_EMBEDDED\_FILES switch has been set to Sanitise a dictionary is shown as tagged for removal. See Section 4 Configuration Management for details on content management switches

<gw:SanitisationItems itemCount = "1">

<gw:SanitisationItem>

<gw:TechnicalDescription>Document information dictionary detected in a document trailer dictionary.</gw:TechnicalDescription>

<gw:InstanceCount>1</gw:InstanceCount>

<gw:TotalSizeInBytes>0</gw:TotalSizeInBytes>

<gw:AverageSizeInBytes>0</gw:AverageSizeInBytes>

<gw:MinSizeInBytes>0</gw:MinSizeInBytes>

<gw:MaxSizeInBytes>0</gw:MaxSizeInBytes>

</gw:SanitisationItem>

Not all XML reports include Remedies as these are automatic corrections made to bring any regenerated file in line with the file specification but one is shown in this case.

<gw:RemedyItems itemCount = "1">

<gw:RemedyItem>

<gw:TechnicalDescription>PDF Stream is missing an End-Of-Line before the &apos;EndStream&apos; marker.</gw:TechnicalDescription>

<gw:InstanceCount>7</gw:InstanceCount>

</gw:RemedyItem>

Very few files have an issue as this means they are not just non-conformant with the file specification but Glasswall has been unable to Remedy the issue back to the specification. In Protect Mode a file with an issue can not be regenerated.

gw:IssueItems itemCount="1"> -

<gw:IssueItem>

<gw:TechnicalDescription>/Info dictionary contained an unexpected key (/GTS\_PDFXConformance).</gw:TechnicalDescription>

<gw:IssueId>46</gw:IssueId>

<gw:InstanceCount>1</gw:InstanceCount>

<gw:RiskLevel>Medium</gw:RiskLevel>

</gw:IssueItem>

Each Sanitisation Item, Remedy item or Issue Item has a unique numeric id associated with it that never changes so the item can be uniquely identified by other aplications that may wish to process the XML Reports.

Each Issue and its corresponding numeric id is stored in a database along with a reference to the file specification which, when the file was analysed against it, generated the issue. The file specification can then be searched for more information as to why the file is considered non-conforming.

# Protect Mode

## Overview

In Protect Mode Content Management Policies allow control of various document element types such as file attachments, executable code, interactive form content and a number of actions (e.g. external links or the execution of Javascripts). These document element types are known to be common attack vectors and when they are encountered within a document the Content Management Policy will define how Glasswall should process them. Each document type has its own Content Management Policy.

The active Content Management Policy can be updated on a document by document basis, but it must be set prior to Glasswall being able to process any document. In the event of the Content Management Policy not being set before processing documents, documents will be classified as 'non-conforming'.

## Protect Process

In Protect mode (as per in Analysis mode) an input file is read in and the manufacturer’s specification is used to validate each byte as it is processed. This allows all the data structures (content items) within the file to be syntactically validated and if they pass syntax validation they are subjected to further semantic checks. Those data structures that have passed both syntax and semantic validation are then written out to the new version of the input file that is regenerated on a data structure by data structure basis.

If a data structure does not pass validation an attempt is made to force the structure back to the manufacturer’s specification (the content item is remedied) before it is written to the regenerate file. If this cannot be done then a Content Issue is reported as the file cannot safely be regenerated and the input file has to be quarantined.

Whilst the input file is being read in, the Content Management policies are used to decide if key document element types, for example, macros or embedded files are required in the regenerated file. If these items are not required they are not written to the regenerated files as they are sanitised out.

## Protected Files

Files that have been through Protect mode and have been regenerated are fully conformant with the manufacturer’s specification and depending on the content management policies applied may have had a number of data items removed from the file. If all the content management policies were applied and set to Sanitise then if the regenerated file was to be reprocessed by Glasswall in Analysis mode there would be no issues items, no remedies items and no sanitisation items reported in the associated analysis report.

# Automatic Corrections

## Overview

Automatic corrections back to the file specification is performed on file regeneration. The purpose of this is to enable Glasswall to take out threats that are hidden within the file structure as well preventing the possibility of activating exploits via the misuse of structural components in the file.

In analysis mode, all automatic corrections made to a file are reported as remedy items.

## Remediation

There are various types of remedies that are performed across all supported file types.

* Unrecognised objects that are hidden within the file structure and are not defined in the official specification are removed automatically.
* Where possible, components of a file that deviate from the manufacturers specification are automatically corrected back to the standards set in the specification, otherwise an issue is reported.
* As mentioned in the Configuration Management section, content management is not configurable on image files therefore content such as metadata is automatically removed from images in addition to the points above.

# Configuration Management

## Overview

Content Management Policies are used to specify the set of content management switches that should be applied to a particular document type.

The content management switch is used to identify a document element type and associated action.

The content management setting is used to specify what actions carried out by Glasswall for a particular content management switch. Each content management switch can be set to one of three settings:

* Allow - Glasswall processes any associated document element types and they remain in the managed document and is logged in the Analysis report as a Content Item.
* Disallow - If any of the associated document element types are identified in a document, Glasswall identifies this document as being non-conforming, the presence of this document element type is logged in the Analysis report as an Issue Item.
* Sanitise - If any of the associated document element types are identified in a document, Glasswall removes them from the managed document, the removal is logged in the Analysis report as a Sanitisation Item.

## XML Reports with different Content Management Policies

The following sections show how the same item of data that is under control of a content management switch is represented differently in the XML report, depending on the switch setting.

### Allow

This is an extract from the XML report for a doc file which has metadata inside it with the content management switch for metadata set to Allow.

<gw:Camera cameraName = "wordConfig">

gw:ContentSwitch>

<gw:ContentName>metadata</gw:ContentName>

<gw:ContentValue>allow</gw:ContentValue>

</gw:ContentSwitch>

<gw:ContentItem>

<gw:TechnicalDescription> Metadata detected in #05SummaryInformation</gw:TechnicalDescription>

<gw:InstanceCount>1</gw:InstanceCount>

<gw:TotalSizeInBytes>4096</gw:TotalSizeInBytes>

<gw:AverageSizeInBytes>4096</gw:AverageSizeInBytes>

<gw:MinSizeInBytes>4096</gw:MinSizeInBytes>

<gw:MaxSizeInBytes>4096</gw:MaxSizeInBytes>

</gw:ContentItem>

### Disallow

This is an extract from the XML report for a doc file which has metadata inside it with the content management switch for metadata set to Disallow. In Protect mode this would cause the file to be Non-conforming.

<gw:Camera cameraName = "wordConfig">

<gw:ContentSwitch>

<gw:ContentName>metadata</gw:ContentName>

<gw:ContentValue>disallow</gw:ContentValue>

</gw:ContentSwitch>

<gw:IssueItem>

<gw:TechnicalDescription> Metadata detected in #05SummaryInformation</gw:TechnicalDescription>

<gw:IssueId>96</gw:IssueId>

<gw:InstanceCount>1</gw:InstanceCount>

<gw:RiskLevel>Medium</gw:RiskLevel>

</gw:IssueItem>

### Sanitise

This is an extract from the XML report for a doc file which has metadata inside it with the content management switch for metadata set to Sanitise. In Protect mode this would cause the metadata to be removed from the regenerated file.

<gw:Camera cameraName = "wordConfig">

<gw:ContentSwitch>

<gw:ContentName>metadata</gw:ContentName>

<gw:ContentValue>sanitise</gw:ContentValue>

</gw:ContentSwitch>

<gw:SanitisationItem>

<gw:TechnicalDescription>Metadata detected in #05SummaryInformation</gw:TechnicalDescription>

<gw:InstanceCount>1</gw:InstanceCount>

<gw:TotalSizeInBytes>4096</gw:TotalSizeInBytes>

<gw:AverageSizeInBytes>4096</gw:AverageSizeInBytes>

<gw:MinSizeInBytes>4096</gw:MinSizeInBytes>

<gw:MaxSizeInBytes>4096</gw:MaxSizeInBytes>

</gw:SanitisationItem>

## PDF

The PDF content management switches are shown in the table below:

|  |  |
| --- | --- |
| Content Management Flag | XML Report Issue Technical Descriptions |
| javascript | Action of type /JavaScript  Javascript content |
| acroform | Interactive form (Acroform) content  Widget annotation  Action of type /SubmitForm  Action of type /ResetForm  Action of type /ImportData |
| external\_hyperlinks | Action of type /URI |
| embedded\_files | Embedded file content |
| metadata | Document information dictionary detected  XML Metadata detected |
| actions\_all | Action of type /[action type]  *Where action type =[Rendition|Sound|*  *Movie|Hide|SetOCGState|*  *GoTo3DView* |
| internal\_hyperlinks | Action of type /GoTo |
| embedded\_images | Embedded image present in [store] |

## Images

Content management is not configurable on image files, therefore only content items, issues and remedies can appear in an XML report for image file. Sanitisation items cannot be present in their XML reports. Automatic corrections are applied to a file in order to bring it in line with the manufacturer’s specification and/or remove unrecognised or unwanted data. See section [4.1](#_Overview_1).

## Binary and XML Office

Common

There are a number of content management switches common to each of the Office file types

|  |  |
| --- | --- |
| Content Management Flag | XML Report Issue Technical Descriptions |
| macros | Macros present |
| external\_hyperlinks | External Hyperlinks present in [links]  *links = [HFD|VtHyperlinks|\*]* |
| embedded\_files | Embedded File present in [store]  *store = [Object pool|Data Stream|\*]* |
| metadata | Metadata detected in [meta type].  *meta type = [SttbfRMark|Document Summary Stream|Summary Information Stream|\*]* |
| review\_comments | Review comments present in [comment meta type]  *comment meta type = [Content Stream|GrpXstAtnOwners|CommentTree|aATRDPre10|Comment10TextAtom|TxO|\*]* |
| internal\_hyperlinks | Internal Hyperlinks present in [links]  *Link = [location | CT\_Hyperlink | \*]* |
| embedded\_images | Embedded image present in [store] |

File type Specific

The following content management switches are file type specific.

|  |  |  |
| --- | --- | --- |
| **Content Management Flag** | **XML Report Issue Technical Descriptions** | **File Type(s)** |
| dynamic\_data\_exchange | DDE links present in present in [links]  *Link = [CT\_DdeItems|CT\_CalcChain|CT\_Cell| SupBook|\*]* | Excel Binary, Excel XML, Word Binary and Word XML |

Note:

* doc and docx are controlled by the same switch setting(s)
* xls and xlsx are controlled by the same switch setting(s)
* ppt and pptx are controlled by the same switch setting(s)

# File Types Supported

## Overview

The table below lists the document formats supported by Glasswall

|  |  |  |
| --- | --- | --- |
| **File Extension** | **Other Extensions** | **Document Type** |
| .pdf |  | PDF documents |
| .jpg |  | JPEG image files |
| .gif |  | GIF image files |
| .png |  | PNG image files |
| .emf |  | EMF image files |
| .wmf |  | WMF image files |
| .doc | .dot | Word Binary File Format files |
| .xls | .xlt | Excel Binary File Format files |
| .ppt | .pot | PowerPoint Binary File Format files |
| .docx | .docm, .dotx, .dotm | Office Open XML Document files |
| .xlsx | .xlsm, .xltx, .xltm | Office Open XML Workbook files |
| .pptx | .pptm, .ppsx, .ppam, .potm, .ppsm | Office Open XML Presentation files |

### Supported File Sizes

Glasswall provides a 64-bit version of a Dynamic-link Library (.DLL) for windows and/or a Shared Object (.SO) for Linux operating systems that can analyse and protect **files up to the size of 2GiB** (230 bytes) provided that other internal limits (such as recursion depth) are satisfied.

## PDF

### Processing Specific to PDF in Protect Mode

PDF is the only file type that allows for an optional watermark to be added to the regenerated file. The watermark is a maximum of 20 (ASCII) characters and is added to the top right hand corner of each page of the regenerated file.

### Specification Validated Against

* PDF Reference (third edition) Adobe Portable Document Format Version 1.7

## Images

There are no content management switches for any of the image file formats. Images embedded in a document which do not conform to the specification would result in the original document being quarantined.

### JPG

#### Specification Validated Against

* JPEG File Interchange Format, Version 1.02 September 1, 1992
* JP2 for ISO/IEC 15444-1 (JPEG 2000)
* JPX for ISO/IEC 15444-2 (JPEG 2000)

### GIF

#### Specification Validated Against

* Graphics Interchange Format, Version GIF 87
* Graphics Interchange Format, Version GIF 89a

### PNG

#### Specification Validated Against

* Portable Network Graphics (PNG): Functional specification. ISO/IEC 15948:2003 (E)

## Binary Office

### Processing Specific to Binary Office in Protect Mode

The content management switch GWFILE\_MANAGE\_REMOVE\_MACROS when set to sanitise in Protect mode removes the VBA script from the regenerated file but copies over the macro container. This can give the impression that the macros have not been deleted.

The content management switch GWFILE\_MANAGE\_REMOVE\_EMBEDDED\_FILES when set to sanitise in Protect mode do one of two options.

* If the embedded file is file that Glasswall can be process it will process that file with its own configuration management polices as if it were a standalone file. If the file can be regenerated the regenerated file will be re-embedded, it if cannot the file it is part of will be marked as non-conforming.
* If the embedded file is file that Glasswall cannot process like a .mp4 or .dat file then the embedded file is removed.

Embedded files are validated to five levels deep.

### DOC

#### Specifications Validated Against

The following specifications are used to validate a doc file type:

* From *[MS-DOC]: Word (.doc) Binary File Format - Microsoft Corporation, 2011, v20121003* to *[MS-DOC]: Word (.doc) Binary File Format - Corporation, 2017, v20170112*
* Office Common Data Types and Objects Structure Specification- Microsoft Corporation, 2011, v20110608
* Office Document Cryptography Structure Specification- Microsoft Corporation, 2011, v20110608
* Office VBA File Format Structure Specification- Microsoft Corporation, 2011, v20110608

The following files extensions are supported .doc, .dotm and .dot.

#### Non Supported Specifications

Word 95 and earlier versions (Word 1 to Word 6) are not supported and if these file types are processed they will cause an Issue to be reported in the XML report in Analysis mode and the file not to be regenerated in Protect mode.

Word files containing Apple Mac or LibreOffice specific file features will be marked as non conformant.

### XLS

#### Specifications Validated Against

The following specifications are used to validate an xls file type:

* From *[MS-XLS]: Excel Binary File Format (.xls) Structure Specification - Microsoft Corporation, v20141018* to *[MS-XLS]: Excel Binary File Format (.xls) Structure Specification - Microsoft Corporation, v20161017.*
* Spreadsheet Data Model File Format,
* Office Graph Binary File Format v20141019

The following files extensions are supported .xls, .xla, .xlt and .xlsb.

#### Non Supported Specifications

Excel 95 and earlier versions are not supported and these file types are processed cause an Issue to be reported in the XML report in Analysis mode and the file not to be regenerated in Protect mode.

Excel files containing Apple Mac or LibreOffice specific file features will be marked as non conformant.

### PPT

#### Specifications Validated Against

The following specifications are used to validate a doc file type against:

* From *[MS-PPT]: PowerPoint (.ppt) Binary File Format - Microsoft Corporation, v20121003* to [MS-PPT]: PowerPoint (.ppt) Binary File Format - Microsoft Corporation, *v20160922*.

The following files extensions are supported .ppt, .ppa, .pps, and .pot.

#### Non Supported Specifications

PowerPoint 95 and earlier versions are not supported and if these file types are processed it will cause an Issue to be reported in the XML report in Analysis mode and the file not to be regenerated in Protect mode.

PowerPoint files containing Apple Mac or LibreOffice specific file features will be marked as non conformant.

## XML Office

### Processing Specific to XML Office in Protect Mode

If a XML Office file has another XML Office file embedded inside it and the GWFILE\_MANAGE\_REMOVE\_EMBEDDED\_FILES is set to sanitise, the embedded file will be removed when a new file is regenerated.

Office XML files are stored as zipped archive files and although the zipped file may be less than 10Mb when the streams are extracted and uncompressed an individual stream may be greater than 10MB. When using the 32-bit DLL, streams that are larger than that 10 MB cause the file to be marked as non–conforming and are not processed any further. When using the 64-bit DLL, streams of up to 50MB are processed.

### Specifications Validated Against

#### Office 2010

The following specifications are used to validate docx/xlsx/pptx file types:

* [ECMA-376 4th edition Part 1](http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-376,%20Fourth%20Edition,%20Part%201%20-%20Fundamentals%20And%20Markup%20Language%20Reference.zip)
* [ECMA-376 4th edition Part 2](http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-376,%20Fourth%20Edition,%20Part%202%20-%20Open%20Packaging%20Conventions.zip)
* [ECMA-376 4th edition Part 3](http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-376,%20Fourth%20Edition,%20Part%203%20-%20Markup%20Compatibility%20and%20Extensibility.zip)
* [ECMA-376 4th edition Part 4](http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-376,%20Fourth%20Edition,%20Part%204%20-%20Transitional%20Migration%20Features.zip)

#### Office 2013

The following specifications are used to validate docx/xlsx/pptx file types:

* Word (.docx) Extensions to the Office Open XML SpreadsheetML File Format
* Excel (.xlsx) Extensions to the Office Open XML SpreadsheetML File Format
* PowerPoint (.pptx) Extensions to the Office Open XML SpreadsheetML File Format
* Office Drawing Extensions to the Office Open XML Structure

The following files extensions are supported:

* .docx, .docm, .dotx
* .xlsx, .xlsm, .xltm, .xltx
* .pptx, .pptm, .ppsm, .ppam, .ppsx

### Non Supported Specifications

Office 2010 or Office 2013 files containing Apple Mac or LibreOffice specific file features may be marked as non conformant or the components may be removed from the file.

# Determine File Format

Glasswall has the ability to determine the file type for a given file provided that the file format is supported by the Glasswall engine. This enables the user to determine and supply the Glasswall processing API functions with the correct file type for the file that is to be processed, regardless of the presence of a file extension or an incorrect extension. The API that determines the file format will specify how the file format was deduced.

# Issue ID Information Retrieval

There are two API functions responsible for retrieving information related to issue ID’s. Issue IDs are generated when files processed in analysis mode and can be found in the output analysis report. Descriptions of the API functions are as follows:-

* The first of the two API functions is responsible for retrieving an issue description for a provided issue ID.
* The second API function is responsible for retrieving issue descriptions for all issue ID ranges. The output XML file lists the ranges and the descriptions for each range.

# Content Export and Import

Glasswall provides the ability to export and import certain content items from **PDF** and **MS**-**Office documents**.

This allows internal components of processed files to be made available to external programs for additional processing outside of the Glasswall domain, enabling the user to validate the components externally before recomposing the files to include those externally modified components.

In order to enable the user to carry out additional analysis on components within files, the files must be processed by the engine twice: once to extract a package containing a processed file, report any extracted components, and a second pass to reintegrate the externally analysed and/or modified components back into the document. The files are re-validated and regenerated for each pass to ensure file integrity.

## Exportable Content

The following bullet points list the exportable content supported in Export Mode.

* Images
  + PNG
  + JPEG
  + GIF
  + EMF
  + WMF
* Text - This covers all text stored in a document.

## Importable Content

The following bullet points list the importable content supported in Import Mode.

* Images
  + PNG
  + JPEG
  + GIF
  + EMF
  + WMF

# Prohibited Word Tagging

Glasswall can search the metadata for forbidden words in these file formats:

* PDF
* Microsoft Binary Office
* Office Open XML files
* PNG
* JPEG
* GIF
* EMF
* WMF
* BMP and DIB
* RTF
* Audio types (MP3, WAV)
* Video types
  + MPEG Program Stream (MPG, MPEG, M2P, PS)
  + MPEG Transport Stream (TS, TSV, TSA)
  + MP4 (MP4, M4A, M4P, M4B, M4R, M4V
* TIFF
* GeoTIFF
* Binary (ELF, Mach-O, PE)
* Archive Types (7z, RAR, GZ, BZ2, ZIP, TAR, UNIX Compress)

Additionally, Glasswall can detect words classified as forbidden within the content of PDF, Microsoft Binary Office files, and Office Open XML files.

The forbidden words will be provided in a list and associated with user-defined tags which will define the type of action to be applied to the matching text. Supported actions are redaction, removal, or replacement of the matched words and will define the scope of the action as a word, sentence, or page.

Text export will report the location of text within files, the location of the text, and its encoding.

The APIs for forbidden word detection support string and character based matching, and regular expression matching.

# Watermarking and Security Tagging

Glasswall can embed data into files to watermark them and provide security information and an audit trail. Provision to log the following data will be provided: last modified date, the IP address of the machine of origin, the MAC address of the machine of origin, and various date and timestamps indicating different types of edit. Provision to log the users associated with various types of edit will be provided. Where possible, existing or default values will be used. The user will be able to supply a string to be embedded.

The locations of storage of the metadata will be in different places according to the file format. A toolset will be provided to verify a file contains the metadata and to detect if the file has been tampered with. Confirmation of malicious changes may not be possible.

Hidden files with encoded signature data can be embedded in Microsoft Office Binary File Format files (OLE compound file) for watermark purposes.

Content Management settings are available to enable or disable watermarking.