Google Search Appliance

Indexable File Formats

Google Search Appliance software version 7.2 and later

Google

Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
www.google.com

GSA-IFF_200.03
March 2015

Contents

Indexable File Formats

This document lists the file formats that the Google Search Appliance can crawl, index, and search.

Overview

The following sections list word processing, spreadsheet, database, presentation, and other formats that the Google Search Appliance can crawl, index, and search. Please note the following:

- The Google Search Appliance can directly crawl the file formats listed in this document. Other file formats can be indexed and searched by using a content feed or a connector, including, but not limited to VSD; multimedia content types such as WAV, MP3, and MP4; and dynamic content types such as CGI, PHP, ASP, and ASPX.
- Text embedded in graphics is not indexed.

  The Google Search Appliance cannot index text contained in graphic file formats, such a JPEG, GIF, or TIFF. When a file in a graphic format is submitted for indexing, text embedded in the graphic is not indexed. However, the file name is indexed. If any metadata is associated with the graphic in HTML meta tags, that metadata is indexed. If a JPEG file has Exchangeable Image Format (EXIF) data, the data is indexed as metadata.
- Encrypted, viewable PDF documents are converted to HTML for indexing, but the cached HTML is not displayed.
- Encrypted Excel spreadsheets (xls format) cannot be indexed or searched. If the search appliance attempts to crawl and index an encrypted Excel spreadsheet, you see the following error on the Crawl Diagnostics page:

  ```
  Crawled with empty body: Conversion error
  ```

  To make Excel spreadsheets indexable, disable encryption on the Excel **Tools > Options > Security** tab and resave any affected spreadsheets.
- PDF files created by scanning with optical character recognition (OCR) software are supported.
- If you are using the Google Search Appliance, metadata can be fed from a database and then indexed.
- Files in XML format are crawled and indexed as plain text. Links are not extracted or followed and XML tags are converted to escaped HTML counterparts.

- The search appliance supports extraction and indexing of compressed content from the formats listed in “Indexable Archive Formats” on page 16.
- Microsoft Office 2007 files, which have file extensions of .docx, .pptx, .xlsx, and so on, consist of ZIP archives of many XML files. The search appliance indexes most Microsoft Office 2007 files correctly. However, if the uncompressed file size is larger than 30 MB, the search appliance cannot index the file. In these cases, you see a `Conversion error` message on the **Index > Diagnostics > Index Diagnostics** page.
- The search appliance attempts to determine the type of file it is crawling by first examining the Content-Type HTTP header and then by examining the file extension. Provided that the Content-Type header is present at crawl time, the search appliance crawls and indexes files where the content type does not match the file extension. For example, an HTML file saved with a PDF extension is correctly crawled and indexed as an HTML file.
- If the search appliance crawls or is fed a document with a mime type that it does not know how to interpret, it associates that document with text/other and indexes it. In this case, when the document is searched for, the search appliance guesses the actual mime type from the file extension. If it cannot determine the mime type from the file extension, then the mime type returned is application/octet-stream.

How the Google Search Appliance Determines the Document Title

The Google Search Appliance analyzes documents during the indexing process to determine which text is the document title and which is the body text. How the search appliance makes the determination varies by the document type.

If you want titles extracted from document metadata, do not use a value for the title metadata that is the same as the file name.

The search appliance ignores the title tag in a web page if it has only one character.

PDF Documents

The search appliance uses the PDF document title property as the title in the search index. If the document title is the same as the file name, the search appliances uses the first text it discovers in a large font within the document itself. In all cases, the values of the metadata fields are indexed as part of the document content.

Only documents without copyright protection (documents with printing, copying, and editing enabled) will show cached versions and document previews.

XLS Documents

The search appliance uses the **Properties > Title** property as the title in the search index. If the search appliance is unable to do this, it uses the name of the first worksheet.

Extracted document properties become metatags in the HTML representation of an XLS document. For example:

```
<meta http-equiv="Content-Type" content="text/html; charset=Latin1">
<meta name="Producer" content="Acrobat Distiller 4.05 for Windows">
<meta name="ModDate" content="D:20011129112148-06'00'">
<meta name="Author" content="Charles Dickens">
<meta name="CreationDate" content="D:20011129112114">
<meta name="Creator" content="Microsoft Word 9.0">
```

Text Documents

Text documents do not have titles associated with the document. The search appliances uses the first 70 bytes of the document as the title when it serves search results.

Indexable Word Processing Formats

The following table lists supported word processing formats.

Format	Extension	Versions Supported
Adobe FrameMaker	mif	Versions 3.0–6.0
Adobe Illustrator Postscript	ppd	Level 2
Ami	sam	
Ami Pro for OS2	sam	
Ami Pro for Windows	sam	Versions 2.0, 3.0
ANSI Text (7 & 8 bit)	ans	All versions
ASCII Text (7 & 8 bit)	txt	All versions
DEC DX	dx	Versions through 4.0
DEC DX Plus	wpl	Versions 4.0, 4.1
DisplayWrite	rft, dca	Versions 2.0–5.0
DOS character set		
EBCDIC		
Enable	wpf	Versions 3.0–4.5
First Choice	pfc	Versions 1.0, 3.0
Framework	net	Version 3.0
Hangul	hwp	Versions 97–2007
HTML	html, htm	Versions 1.0–4.0 (some limitations)
IBM DCA/FFT	fft	All versions
IBM DCA/Revisable Form Text	rft	All versions
IBM Writing Assistant	iwa	Version 1.01

Format	Extension	Versions Supported
JustSystems Ichitaro	jaw, jbw, jtd	Versions 5.0, 6.0, 8.0–13.0, 2004, and 2010
JustWrite	jw	Versions through 3.0
Kingsoft WPS Writer	wps	Version 2010
Legacy	leg	Version 1.1
Lotus Manuscript	doc	Versions through 2.0
Lotus WordPro	lwp	Versions 9.7, 96–Millennium 9.6
Lotus WordPro (non Win32)	lwp	Versions 97–Millennium 9.6
Macintosh character set		
MacWrite II	mcw, mw, mwii	Version 1.1
MASS11	m11	Versions through 8.0
Microsoft Publisher (file ID only)	pub	Versions 2003–2007
Microsoft Rich Text Format	rtf	All versions
Microsoft Word for DOS	doc	Versions 4.0–6.0
Microsoft Word for Macintosh	doc	Versions 4.0–6.0, 98–2008
Microsoft Word for Windows	doc	Versions 1.0–2010
Microsoft Word for Windows	doc	Version 2003 XML (text only via XML filter)
Microsoft Word for Windows	doc	Version 98-J
Microsoft WordPad	rtf, doc	All versions
Microsoft Works for DOS	wks, wps	Version 2.0
Microsoft Works for Macintosh	wks, wps	Version 2.0
Microsoft Works for Windows	wks, wpf	Versions 3.0, 4.0
Microsoft Write for Windows	wri	Versions 1.0–3.0
MultiMate	dox	Versions through 4.0
MultiMate Advantage	dox	Version 2.0
Navy DIF	dif	All versions
Nota Bene	nb	Version 3.0
Novell PerfectWorks	wpw	Version 2.0
Novell WordPerfect for DOS	wpd	Version 4.2
Novell WordPerfect for Mac	wpd	Versions 1.02–3.1
Novell WordPerfect for Windows	wpd	Versions 5.1–X4
Office Writer	ow4	Version 4.0–6.0
OpenOffice Writer	odt, ott	Versions 1.1–3.0
Oracle Open Office Writer	odt, ott, sxw, stw	Versions 3.x
PC File Doc		Version 5.0

Format	Extension	Versions Supported
PFS: Write	pfb	Versions A, B
Professional Write for DOS	pw	Versions 1.0, 2.0
Professional Write Plus for Windows	pw, pwp	Version 1.0
Q&A Write for Windows	dtf	Versions 2.0, 3.0
Samna Word IV	sam, sm	Versions 1.0–3.0
Samna Word IV+	sam, sm	
Samsung Jungum Global (file ID only)	gul	
Signature	sig	Version 1.0
SmartWare II	smt	Version 1.02
Sprint	spr	Version 1.0
StarOffice Writer	sxw, odt	Versions 5.2–9.0
Total Word	tw	Version 1.2
Unicode Text	txt	Versions 3.0, 4.0
UTF-8	utf	
Volkswriter 3 & 4	vw	Versions through 1.0
Wang IWP	iwp	Versions through 2.6
Wireless Markup Language	wml	All versions
WordMarc	wmc	Versions through Composer Plus
WordPerfect for DOS	wpd	Version 4.2
WordPerfect for Macintosh	wpd	Versions 1.02–3.1
WordPerfect for Windows	wpd	Versions 5.1–X.4 (recheck)
WordStar 2000 for DOS	ws1, ws2, ws3	Versions 1.0–3.0
WordStar for DOS	ws	Versions 3.0–7.0
WordStar for Windows	ws, wst, wsd	Version 1.0
XML (text only)	xml	
XHTML (file ID only)	xhtml	Version 1.0
XyWrite	xy3, xyp, xyw	Versions through III Plus

Indexable Spreadsheet Formats

The following table lists supported spreadsheet formats.

Format	Extension	Versions Supported
Enable	300, wpf, ssf, dbf	Versions 3.0–4.5
First Choice	ss, fol	Versions through 3.0
Framework	fw3	Version 3.0
Kingsoft WPS Spreadsheets	wps	Version 2010
Lotus 1-2-3	wku, wk1, wk2, wk3, wk4, wk5, wki, wks	Versions through Millenium 9.6
Lotus 1-2-3 Charts (DOS & Windows)	wku, wk1, wk2, wk3, wk4, wk5, wki, wks	Versions through 5.0
Lotus 1-2-3 (OS/2)	wku, wk1, wk2	Versions through 2.0
Lotus Symphony	wr1	Versions 1.x through 2.0
Microsoft Excel Charts	xlc	Versions 2.x through 7.0
Microsoft Excel for Macintosh	xls	Versions 98–2008
Microsoft Excel for Windows	xls, xlw	Versions 3.0 through 2010 (2007 with extensions xlsx and xlsm)
Microsoft Excel for Windows	xlsb	Versions 2007–2010 (binary)
Microsoft Excel for Windows	xml	Version 2003 XML (text only via XML filter)
Microsoft Works (DOS)	wps, wks, wdb, wcm	Version 2.0
Microsoft Works (Windows)	wps, wks	Versions 3.0, 4.0
Microsoft Works (Macintosh)	wps, wks, wdb, wcm	Version 2.0
Multiplan	col, cod, mod	Version 4.0
Novell Perfect Works	wpw	Version 2.0
OpenOffice Calc	odc, sdc	Versions 1.1–3.0
Oracle OpenOffice Calc	ods, ots, sxc, stc	Versions 3.x
PFS: Plan	tid	Version 1.0
QuattroPro (DOS)	wkq, wq1	Versions through 5.0
QuattroPro (Windows)	wb1, wb2, wk3	Versions through X4
SmartWare II	ws	Version 1.02
SmartWare Spreadsheet	ws	
StarOffice Calc (Windows and UNIX)	sdc, sxc, ods, ots	StarOffice versions 5.2–9.0, and OpenOffice version 1.1 (Text only)
SuperCalc 5	cal	Version 5.0
VP-Planner	np	Version 1.0

Indexable Database Formats

The following table lists supported database formats.

Format	Extension	Versions Supported
DataEase	dba, dbm, dql	Version 4.x
DBASE	dbf	Versions III, IV, V
First Choice	pfc	Version 3.0
Framework	fwk, fw, fw2, fw3	Version 3.0
Microsoft Access	mdb	Versions 1.0, 2.0
Microsoft Access Report Snapshot (file ID only)	mdb	Versions 2000–2003
Microsoft Works (DOS)	wdb, wks	Versions 1.0, 2.0
Microsoft Works (Macintosh)	wdb, wks	Version 2. 0
Microsoft Works (Windows)	wdb, wks, dbf	Versions 3.0, 4.0
Paradox (DOS)	fsl, db, px	Versions 2.0–4.0
Paradox (Windows)	fsl, db, px	Version 1.0
Q&A	qa, qw, dtf	Versions through 2.0
R:Base 5000	rbf, dbf	R:Base 5000
R:Base System V	rbf	R:Base System V
Reflex	r2d	Version 2.0
SmartWare II	db	Version 1.02

Indexable Graphics Formats

The following table lists supported graphics formats. Note that text that is part of a graphic is not indexed. Only file names and metadata are indexed.

Format	Extension	Versions Supported
Adobe FrameMaker Graphics	fmv	Vector/raster 3.0–5.0
Adobe Illustrator File Format	ai	Versions 4.0–7.0, 9.0
Adobe Illustrator	xmp	Versions 11–13 (CS 1–3)
Adobe InDesign	xmp	Versions 3–5 (CS 1–3)
Adobe InDesign Interchange	xmp	
Adobe Photoshop File Format	psd	Version 8.0–10.0 (CS 1–3)
Adobe Photoshop	psd	Version 4.0
Adobe Portable Document Format	pdf	Versions 1.0–1.7 (Acrobat Versions 1–9, including Japanese PDF)

Format	Extension	Versions Supported
Adobe Portable Document Format Package, Portfolio	pdf	Version 1.7 (Acrobat Versions 8–9)
Ami Draw	sdw	
AutoCAD Drawing	dwg	Versions 2.5, 2.6
AutoCAD Drawing	dwg	Versions 9.0–14.0
AutoCAD Drawing	dwg	Versions 2000i–2010
AutoShade Rendering	rnd	Version 2
CALS Raster Format	gp4	Type I and Type II
Computer Graphics Metafile	cgm	ANSI, CALS NIST Versions 3.0
Corel Draw	cdr	Versions 2.0–9.0
Corel Draw Clipart	cmx	Versions 5.0, 7.0
Encapsulated PostScript	eps	tiff header only
Enhanced Metafile	emf	
Escher graphics		
GEM File (vector)	gem	
GEM Image (bitmap)	img	No specific version
Graphics Environment Manager	gem	Bitmap and vector
Graphics Interface Format	gif	No specific version
Hewlett Packard Graphics Language	hpgl	Version 2
IBM Graphics Data Format	gdf	Version 1.0
IBM Picture Interchange Format	pif	Version 1.0
IGES Drawing	igs	Versions 5.1–5.3
JBIG2	jb2	(Graphic embeddings in PDF)
JFIF (jpeg not in tiff format)	jfif	All Versions
JPEG (including EXIF)	jpeg	All versions
JPEG 2000	jpeg	JP2
Kodak flash pix	fpx	
Kodak Photo CD	pcd	Version 1.0
Lotus PIC	pic	All versions
Lotus Snapshot	snp	All versions
Macintosh PICT and PICT2	pict	Bitmap only
MacPaint	pntg	No specific version
Micrografx Designer	drw	Versions through 3.1
Micrografx Designer	dsf	Version 6.0
Micrografx Draw	drw	Versions through 4.0

Format	Extension	Versions Supported
Microsoft Windows Bitmap	bmp	
Microsoft Windows Cursor	cur	
Microsoft Windows Icon	ico	
Microsoft XPS (text only)	xps	
Novell PerfectWorks	draw	Version 2.0
OpenOffice Draw	sda, odg, otg	Versions 1.1–3.0
Oracle Open Office Draw	odg, otg, sxd, std	Versions 3.x
OS/2 Bitmap	bmp, ico, ptr	
OS/2 Warp Bitmap	bmp	
Paint Shop Pro 6 (Win32)	psp	Version 5.0, 6.0
PC Paintbrush	pcx, dcx	All versions
Portable Bitmap	pbm	All versions
Portable Graymap	pgm	No specific version
Portable Network Graphics	png	Version 1.0
Portable Pixmap	ppm	No specific version
PostScript	ps	Level 2
Progressive JPEG	jpeg	No specific version
StarOffice Draw	sxd	Versions 6.x–9.0
Sun Raster	srs	No specific version
TIFF Group 5 & 6	tiff	Versions through 6
TIFF CCITT Group 3 & 4	tiff	Versions through 6
TrueVision TGA	targa	Version 2.0
Visio (Page Preview mode)	wmf, emf	Version 4
Visio	vsd	Versions 5.0–2007
Visio (file ID only)	xml, vsx	Version 2007
WBMP wireless graphics format	wbmp	No specific version
Windows Enhanced Metafile	emf	No specific version
Windows Metafile	wmf	No specific version
WordPerfect Graphics	wpg, wpg2	Versions 1.0, 2.0–10.0
X-Windows Bitmap	xbm	x10 compatible
X-Windows Dump	xdm	x10 compatible
X-Windows Pixmap	xpm	x10 compatible

Indexable Presentation Formats

The following table lists supported presentation formats.

Format	Extension	Versions Supported
Harvard Graphics Chart (DOS)	hgs, cht, ch3, prs	Versions 2.0–3.0
Harvard Graphics (Windows)	hgs, cht, ch3, prs	Windows versions
IBM Lotus Symphony Presentations	odp	Version 1.x
Kingsoft WPS Presentation	wps	Version 2010
Lotus Freelance	pre	Version 1.0–Millenium 9.6
Lotus Freelance for OS/3	pre	Version 2
Lotus Freelance (Windows)	flw, shw, drw, pre	Versions 95, 97
Microsoft PowerPoint for Windows	pptm, pptx	Versions 3.0–2010
Microsoft PowerPoint for Macintosh	ppt, pptx	Versions 4.0–2008
Microsoft PowerPoint for Windows Slideshow	pps, ppsx	Versions 2007–2010
Novell Presentations	shw	Versions 3.0, 7.0
OpenOffice Impress	odp	Versions 1.1, 3.0
Oracle Open Office Impress	odp, odg, otp, sxi	Version 3.x
StarOffice Impress (Windows and UNIX)	text only	StarOffice versions 5.2–9.0 and OpenOffice version 1.1 (text only)
WordPerfect Presentations	wpd	Versions 5.1–X4

Indexable Email Formats

The following table lists supported email formats.

Format	Extension	Versions Supported
Apple Mail Message	emlx	Version 2.0
Encoded mail messages	mht, multipart (alternative, digest, mixed, newsgroup, signed), tnef	
IBM Lotus Notes Domino XML Language DXL	dxl	Version 8.5
IBM Lotus Notes NSF (file ID only)	nsf	Versions 7.x, 8.x
IBM Lotus Notes NSF (Windows, Linux x86-32 and Oracle Solaris 32-bit only with Notes Client or Domino Server	nsf	Version 8.x
MBOX Mailbox	mbox	RFC 822
Microsoft Outlook Message (MSG)	msg	Versions 97–2007
Microsoft Outlook Express (EML)	eml	
Microsoft Outlook Forms Template (OFT)	oft	Versions 97–2007
Microsoft Outlook OST	ost	Versions 97–2007

Indexable Multimedia Formats

The following table lists supported multimedia formats.

Format	Extension	Versions Supported
AVI (Metadata extraction only)	avi	
Flash (text extraction only)	swf	Versions 6.x, 7.x, Lite
Flash (file ID only)	swf	Versions 9, 10
Real Media (file ID only)	rm	
MP3 (ID3 metadata only)	id3	
MPEG-1 Audio layer 3 V ID3 (file ID only)	mp3	Versions 1, 2
MPEG-1 Video (file ID only)	mpg	Versions 2, 3
MPEG-2 Audio (file ID only)	mpg	
MPEG-4 (metadata extraction only)	mp4	
MPEG-7 (metadata extraction only)	mp7	
Quicktime (metadata extraction only)	mov, qt	
Windows Media ASF (metadata extraction only)	wma, wmv	
Windows Media DVR-MS (metadata extraction only)	dvr-ms	
Windows Media Audio WMA (metadata extraction only)	wma	
Windows Media Playlist (file ID only)	wpl	
Windows Media Video WMV (metadata extraction only)	wmv	
WAV (metadata extraction only)	wav	

Indexable Archive Formats

The following table lists supported archive formats.

Note that the search appliance only indexes file names and plain text files inside the archive.

Format	Extension	Versions Supported
7z (BZIP2 and split archives not supported)	7Z	
7z Self Extracting .exe (BZIP2 and split archives not supported)	exe	
LZA Self Extracting Compress		
LZH Compress	lzh	
Microsoft Office Binder	obt	Versions 95-97
Microsoft Cabinet	cab	
RAR	rar	Versions 1.5, 2.0, 2.9
Self-extracting .exe	exe	
UNIX Compress	z	
UNIX GZip	gz tgz	
UNIX tar	tar	
Uuencode	uue	
Zip	zip	PKZip
Zip	zip	WinZip

To enable the search appliance to crawl these types of compressed files, comment out these file types under **Do Not Follow Patterns** on the **Content Sources > Web Crawl > Start and Block URLs** page.

Other Indexable Formats

The following table lists other supported formats.

Format	Extension	Versions Supported
AOL Messenger (file ID only)	aim	Version 7.3
Microsoft InfoPath (file ID only)	xsn	Version 2007
Microsoft Live Messenger (via XML filter)	eml	Version 2010
Microsoft OneNote (file ID only)	one	Version 2007
Microsoft Outlook Message	msg	97 through 2007
Microsoft Project (table view only)	mpp	Versions 98–2003, 2007, 2010
Microsoft Windows Compiled Help (file ID only)	chm	
Microsoft DLL	dll	
Microsoft Executable	exe	
Microsoft Windows Explorer Command (file ID only)	scf	
Microsoft Windows Help (file ID only)	hlp	
Microsoft Windows Shortcut (file ID only)	lnk, url	
Trillian Text Log File (via text filter)	txt	Version 4.2
Trillian Text Log File (file ID only)	txt	Version 4.2
TrueType Font (file ID only)	ttf, ttc	
vCalendar	vcs	Version 2.1
vCard	vcf, vcard	Version 2.1
Yahoo Messenger	log	Versions 6.x–8