Format Background Document: Microsoft Word 10.0

Date: July 14, 2005
Author: Grace Carpenter

GDFR data

Canonical identifier info:gdfr/fred/f/msword
Description Microsoft Word 10.0 (DOC)

Alias
Type MIME
Value application/msword
Alias
Type Informal identifier
Value Microsoft Word XP/2002
Version 10.0
Legal or recognized owner
Name Microsoft Corporation
Organization type Commercial (for-profit) entity
Web site http://www.microsoft.com
Relationship
Type of relation May be encapsulated by target
Target format identifier
Type GDFR format identifier
Value gdfr/fred/f/ole2cdf
Target registry identifier
Type GDFR registry identifier
Value Fred
Specification
Document title Microsoft Word 97 Binary File Fomat
Document type Article
Publication date 1998
Access regime Unrestricted access
Identifier
Type URL: Uniform resource locator
Value http://www.wotsit.org/search.asp?s=text
Note This is an unofficial spec, purportedly from the MSDN Library, posted on a third-party site. However, it appears to be the only publicly available documentation on the Word format, and is also the most up-to-date.
Signature
Signature obligation Mandatory under certain conditions (see notes)
External signature type File extension
Signature value .doc
Note Standard Word file extension
Signature
Signature obligation Mandatory under certain conditions (see notes)
External signature type File extension
Signature value .dot
Note extension for file templates
Signature

Signature obligation Mandatory
Signature position Fixed position (requires offset)
Byte offset 0
Signature value 0xD0 0xCF 0x11 0xE0 0xA1 0xB1 0x1A 0xE1
Note This signature identifies an OLE Compound Document, which may or may not contain Word data

Application
Application name Microsoft Word
Application version XP/2002
Application release date 2001
Vendor
Name Microsoft Corporation
Organization type Commercial (for-profit) entity
Application's function
Process type Creates objects in the format
Application's function
Process type Renders objects in the format

Non-GDFR data

1. General

1.1 Description (long)
Microsoft Word 10.0 is the native file format of Microsoft's Word software. It is a binary file format that is always contained within a Compound Document File ([OpenOffice CompoundDoc]).

Note: I've chosen Word 10.0 for this document because a) it is the last version before it became possible to output the document in some form of XML b) although there appears to be no publicly available spec, either official or unofficial, after Word 97, there is an unoffcial spec for Excel 10, as well as one for the Microsoft Compound Document format, published by OpenOffice.org [OpenOffice CompoundDoc]. Since the Excel spec appears to be the most comprehensive and up-to-date documentation for any Microsoft Office format, I've decided to stick with that version for all Office formats. It's not clear to me how much the Word format has changed between Word 97 and Word 2003.
1.2 Content type: text

2. Category-specific (e.g., color depth, color space, progressive display, etc.)

3.General technical
3.1 Encoding Unicode
3.2 Byte order Little Endian
3.3 Encryption A Word document is encrypted with RC4 encryption if a user chooses to use password protection.
3.4 Human readable No

4. Sustainability
4.1 Proprietary Yes
4.2 Owner documentation There are no publicly available specs, official or unofficial. However, specs for Word 97 that claim to have been culled from the Microsoft website (they

appear to have since been removed from the Microsoft website) are available in a number of places, most notably wotsit.org[Wotsit]. Microsoft publishes very little information about the file format itself, but provides a great deal of information about using Word.

4.3 Other documentation There appears to be very little open and reliable information on the innards of Word. Non-Microsoft sources of information on Excel—such as the Jakarta POI project and OpenOffice.org—do not appear to provide anything like the same level of documentation about Word as they do for Excel. The Jakarta POI project [POI]has a subproject on Word (HWPF), but the project pages appear to be out of date and there is very little information about Word anywhere on the project pages.

4.4 Adoption I haven't been able to find verifiable numbers on Microsoft's share of the word processing market. I've seen estimates of Microsoft's share of the overall office application market as being at about 90%.

4.5 Competition Word's primary commercial competitor is WordPerfect. In 2003 OpenOffice.org released its OpenOffice suite, which is the main open source word processing package.

4.6 Licensing and patent claims

4.7 Other preservation issues Word documents can contain various items that complicate preservation strategies. An OLE compound document file can contain other types of streams, most notably macros in Visual Basic, and streams created by other Microsoft Office Applications.

5. Lifecycle

5.1 Version duration 1.5 years?

5.2 Version history A comprehensive list does not appear to be anywhere on the Microsoft site. I took this list from the Wikipedia ([Wikipedia]).

MS-DOS	
1	1983 (Nov)
2	1985
3	1986
4	1987
5	1989
5.5	1991

Mac	
1	1985 (Jan)
3	1987
4	1989
5	1991
6	1993
Word 98	1998
Word 2001	2000
Word v.X	2001
Word 2004	2004

Windows	
Word for Windows	1989
2	1991
6	1993
7/Word 95	1995
8/Word 97	1997

9/Word 2000 1999
10/Word XP 2001
11/Word 2003 2003

5.3 Expected new version Word 11 was released in late 2003. In the past Microsoft has generally released a new version approximately every 1.5 years, but the current date estimate for Office 12 is mid/late 2006.

6.Local use (DSpace at MIT)

6.1 Holdings as of January 2005: 11

6.2 Support level known

7. Useful URLs

[OpenOffice CompoundDoc] OpenOffice Microsoft Compound Document Format Specification
http://sc.openoffice.org/compdocfileformat.pdf

[Wotsit] Link to Microsoft Word 8/Word 97 Format spec
http://www.wotsit.org/search.asp?s=text

[POI] Jakarta POI project
http://jakarta.apache.org/poi/

[Wikipedia] Wikipedia: Microsoft Word
http://en.wikipedia.org/wiki/Microsoft_Word