Action Plan Background: BIFF 8

Author: Carol C.H. Chou, FCLA

Release Date: 2/29/2008

Change History:

04/01/2008 Update Section 3.3 with the latest result from ISO regarding OOXML.

Preface

Binary Interchange File Format (BIFF, a.k.a. Microsoft Office Binary File Format) is a proprietary file format developed by Microsoft for its Excel spreadsheet software. Microsoft has developed several versions of BIFF file format for different versions of Excel. Version 8 of the BIFF file format, called BIFF8, is the native file format generated by Microsoft Excel 97, Excel 2000, Excel 2002 and Excel 2003 [11] [13] [18].

1 General Description

1.1 Format Name: Binary Interchange File Format 8 (BIFF8) Workbook

1.2 Version: 8

1.3 MIME media type name: application

1.4 MIME subtype: vnd.ms-excel, based on IANA registry [2]

1.5 Identifiers:

PRONOM: fmt/61 (http://www.nationalarchives.gov.uk/PRONOM/fmt/61), fmt/62 ((http://www.nationalarchives.gov.uk/PRONOM/fmt/62)

1.6 Short Description: BIFF8 is a binary format for storing the spreadsheet data in Excel 97 to Excel 2003.

1.7 Common Extensions: .xls (workbook)

1.8 Color depth: N/A

1.9 Color Space: N/A

1.10 Compression: In BIFF8, unicode strings may be stored in a compressed format which omits high bytes of all characters if they are all 00h [11][13].

1.11 Progressive Display: N/A

1.12 Animation: No

1.13 Byte Order: Little-endian

1.14 Magic number(s) or equivalent: BIFF8 is based on Microsoft's OLE2 Compound Document Format [5]. An OLE2 compound document essentially works like a directory that contains data streams. The first 512 bytes of an OLE document is the OLE document header which begins with the compound document file identifier: (hex) D0CF11E0A1B11AE1. The compound document header is followed by the streams that make up an Excel workbook. The first stream in an Excel workbook is a "Workbook" stream started with a BOF (Beginning of File) record. For BIFF8, the first 2 bytes of the BOF record must be (hex) 0908 [11] [13]. After skipping the next two bytes, the rest of the BOF record must be (hex) 00060500.

1.15 Specification Requirements: The following BIFF8 format requirements are extracted from Microsoft Binary File Format document [11] [13] and OpenOffice's documentation on Excel file format [1].

BIFF8 workbook is typically stored in a compound document file. A BIFF8 compound document file must contain a "Workbook" stream which is divided into several substreams to describe the workbook and the contained worksheets. A "Workbook" stream must satisfy the following requirements:

- It must contain a "Workbook Global" substream.
- The "Workbook Global" substream must be followed by at least one substream for the first worksheet, and optionally followed by additional substreams for each subsequent worksheet.
- Each substream is further divided into records that consist of a header, followed by record data. The header is composed of two 16-bit words; one describes the record type and the other specifies the length of the record data in bytes.
- Each substream, including "Workbook Global" and worksheet streams, must include a leading BOF record and a trailing EOF record.

The workbook stream may be followed by additional streams such as VBA streams if a document contains Visual Basic modules, Chart streams, PivotTables streams for PivotTable data cache, and a Summary Info stream that describes the document summary information.

The BIFF format specification specifies hundreds of different types of record that can be stored in BIFF8 format. However, the nesting requirements of the records are vaguely defined, so is the BIFF8 record order. It also appears that the record order varies among different versions of Excel. This brings a challenge for validating BIFF8 file format.

Because Excel 97 supports the feature of outputting a workbook in both BIFF5/7 and BIFF8 formats for backward compatibility [13], some Excel workbooks are stored in double

stream format that contains two complete workbook streams. The first one is in BIFF5/7 format, called a "Book" stream and the other stream is in BIFF8 format called a "Workbook" stream which is described in the previous paragraph.

Please note that the specification requirements described in this section refer to the logical structure of the BIFF8 file format. The physical structure of BIFF8 file format is based on OLE Compound Document file format.

2 Essential and Distinguishing Characteristics

BIFF8 is a specialization of Microsoft OLE compound document format. It is used to store the information in a workbook which may contain multiple worksheets. BIFF8 supports various data types, such as number, boolean, string, formula, etc. However, it does not provide a way to embed the fonts used in the spreadsheet. BIFF8 may be embedded with Charts or Pivot Tables to represent data in different perspectives, along with macros for defining custom user functions. Encryption may also be applied to workbooks, which could hinder preservation functionalities.

2.1 Technical Metadata

Workbook Technical Metadata

Technical Metadata element (G = General file metadata, GT = General Text metadata, F = Format specific metadata	Obligation (R = Required, S = Defined in spec., D = Derived from spec., O = Optional)
Byte Order [GT]	S
The title of the workbook [F]	S
Number of worksheets in the workbook [F]	D
Number of Visual Basic modules in this workbook [F]	D
Number of charts embedded in the workbook [F]	D
The number of externally referenced workbooks, DDE links and OLE references [F]	R
Workbook security settings (password protected, read-only enforced, read-only recommended and lock for annotation) [F]	R
Whether the workbook contains a thumbnail image [F]	R

Technical Metadata element (G = General file metadata, GT = General Text metadata, F = Format specific metadata	Obligation (R = Required, S = Defined in spec., D = Derived from spec., O = Optional)
Whether the workbook uses the 1904 date system [F]	R
Whether a PivotTable is included [F]	R
Number of fonts used in the workbook [F]	D
Whether the workbook is in forced calculation mode [F]	R
Whether the workbook contains encrypted content protected by the Information Right Management (IRM) [F].	R

Worksheet Technical Metadata

Technical Metdata elemement (G = General worksheet metadata, F = Format specific metadata)	Obligation (R = Required, S = Defined in spec., D = Derived from spec., O = Optional)
The name of the worksheet [G]	R
Whether the worksheet is visible [F]	R
Number of rows in this worksheet [G]	R
Number of columns in this worksheet [G]	R
Number of imbedded images in this worksheet [F]	R
Number of hyperlinks in this worksheet [F]	R
Whether the worksheet is password-protected [F]	R
Whether the worksheet contains a filtered list [F]	R
Whether the worksheet contains formula [G]	R

3 Usefulness

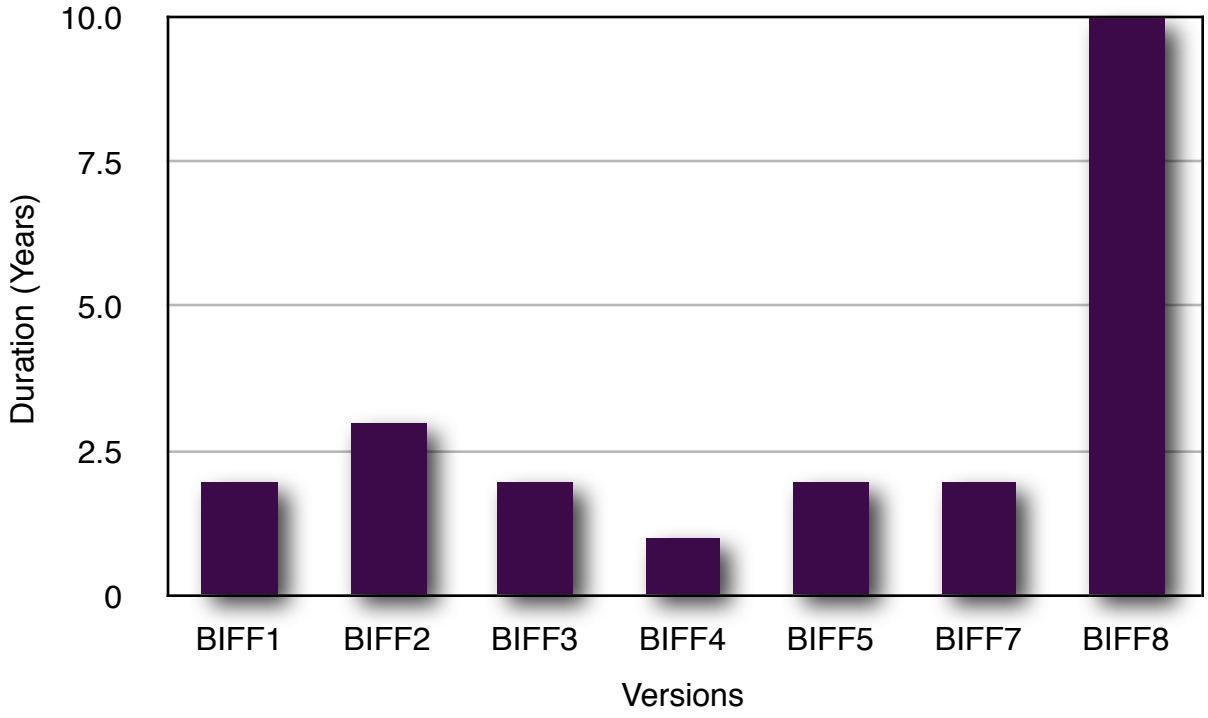
3.1 Version Duration: There is no publication date for Microsoft BIFF format specification. BIFF8 is first introduced in Excel 97 and remains to be the default file format in Excel 2000, Excel 2002 and Excel 2003 [18].

3.2 History of Prior Versions Duration:

Due to the lack of revision history in the current Microsoft BIFF specification and the limited documentation for early BIFF versions, it is a challenge to establish the exact timeline for the prior releases of BIFF formats. The historical information about BIFF formats listed below are compiled from the documents related to Excel file format on www.OpenOffice.org [1] and the article about Microsoft Excel on wikipedia [3].

- **1985** The first version of Excel was released on Mac. Till today, this version of Excel file format remains undocumented.
- **1987** **BIFF2** - The native file format used by Excel 2.x. It uses a simple stream file to store the entire worksheet stream.
- **1990** **BIFF3** - The native file format used by Excel 3.0. It introduces the concept of workspace stream and store it in a simple stream file.
- **1992** **BIFF4** - The native file format used by Excel 4.0
- **1993** **BIFF5** - The native file format used by Excel 5.0 and Excel 7.0 (Excel 95). BIFF5 initiates the use of OLE2 compound document format to store the entire workbook. Additionally, it allows the encoding of Visual Basic for Applications (VBA) in the worksheet to automate tasks and define user functions.
- **1995** **BIFF7** - The native file format used by Excel 7.0 (Excel 95).
- **1997** **BIFF8** - The native file format used by Excel 8.0 (Excel 97) and Excel 9.0 (Excel 2000). In BIFF8, strings are stored using the UTF-16 encoding.

Figure 1: Duration in Years of BIFF versions



3.3 Expected Newer Versions:

The latest version of Microsoft Excel (that is Excel 2007) uses an XML-based Office Open XML file format (OOXML) as its native file format. The Office Open XML file format uses a ZIP container to store files that are used to build up the documents such as spreadsheets, presentations and word documents. OOXML has a similar design as the earlier Microsoft Office XML format (a.k.a Office 2003 XML format), but the data in the document (such as charts, images, document metadata, etc) are stored as separate components with the file rather than embedded within the file. OOXML is designed as a successor to the legacy Office Binary Format, it provides a replacement counterpart for every feature in the Office Binary Format [17]. OOXML includes a set of specialized markup languages to be used by the Microsoft Office. One of the markup languages, SpreadsheetML, is the markup language used to encode spreadsheets in OOXML.

In December 2006, OOXML was published as an ECMA International standard (ECMA 376) [17]. The specification can be download directly from ECMA website. In January 2007, ECMA submitted ECMA 376 to ISO/IEC for fast-track standardization process (DIS 29500). The fast-tracking of OOXML was voted down in August 2007 and was then undergoing the ballot resolution process. A Ballot Resolution Meeting (BRM) for OOXML was held in Geneva in March 2008 which eventually approved OOXML as an ISO/IEC standard.

3.4 Existence of Publicly Available Complete Specifications:

The latest Microsoft BIFF format specification for Excel, that is "Microsoft Office Excel 97-2007 Binary File Format (.xls) specification", is available for a direct download from the Microsoft website [13]. The specification is not dated and does not contain any revision history. It covers the implementation details for BIFF5, BIFF7 and BIFF8 formats for Excel 97-2007. This specification is authoritative and is the most complete one at this moment. However, many areas in the specification can be perplexing to readers as it requires prior background knowledge about Excels. The specifications for BIFF 4 and earlier are still not available from the Microsoft website.

Until recently, Microsoft has kept the BIFF file format specification for its internal use and is available only under a restrictive license. OpenOffice.org has reverse-engineered and published a specification about BIFF formats generated by Microsoft Excel software. The document is called, "OpenOffice.org's Documentation of the Microsoft Excel File format: Excel Version 2, 3, 4, 5, 95, 97, 2000, XP, 2003" [1]. The specification is not authoritative but has served as the most complete BIFF document during the time when the availability of BIFF format specification was restricted by Microsoft.

According to [7], Microsoft once included the file format documentation in their software development kit for Excel 97 and on their MSDN website under a restrictive license. However, the documentation was later removed from the MSDN web site in 1999 and become only available for "use that is complementary to Office". Information in [7] also reflected by an article in the MSDN website dated in 2001 [10] which states the following:

"The Microsoft Excel Binary File Format (BIFF) information is documented in the *Excel 97 Developer's Kit* (ISBN 1-57231-498-2). ... Portions of the Microsoft Excel 97 Developer's Kit are contained in the online MSDN library. However, the sections that involve Excel BIFF are not included in the online MSDN library; additionally, the *Microsoft Excel 97 Developer's Kit* is no longer in print. ..."

As part of Microsoft's effort to promote Office Open XML file format and to encourage users to convert from legacy Microsoft office binary file format, the Office Binary Format documentation is available again on a royalty-free basis in 2006 [6]. Users may receive a copy of the specification with an email request and a signed license agreement. In addition, to fulfill the agreement with ECMA TC45, Microsoft makes the specification available for direct download under its Open Specification Promise starting from February 15, 2008 [15]. To ease the concern about the long-term availability of these specifications, Microsoft has made an agreement with the British Library to provide an independent archive and access for the Microsoft Office Binary file format specifications [18].

3.5 Specifications-controlling Body:

Microsoft Corporation.

3.6 Related Legal Issues:

Microsoft owns the copyright to the BIFF formats. The latest format specification for BIFF file formats is available from Microsoft under its Open Specification Promise [9]. Microsoft Open Specification Promise indicates that "Microsoft irrevocably promises not to assert any Microsoft Necessary Claims against you for making, using, selling, offering for sale,

importing or distributing any implementation to the extent it conforms to a Covered Specification". Compared to the earlier license agreement that only allows a royalty-free use of Microsoft BIFF format specification, the Open Specification Promise frees up third-party BIFF implementors from potential patent-infringement lawsuits from Microsoft.

3.7 Application and Platform Support:

There are many applications that can render and generate BIFF8, including Microsoft Excel 97 and up, the Numbers application in Apple iWork (the Office equivalence on Mac), Open Office Calc, Gnumeric, KSpread, etc. Many of these applications are cross-platform, allowing BIFF8 to be accessible from many different platforms, such as Microsoft Windows, Mac OS, Linux, etc.

Microsoft provides C/C++ API, called Windows Structured Storage API, for Windows developers to manipulate BIFF8 files [8]. Additionally, there are many open source software for reading and writing BIFF8 files, such as Apache POI [14], JACOB project and Java Excel API. [15] lists fourteen Java libraries for reading and writing Excel .xls files.

To encourage users to migrate to Open XML from legacy BIFF, Microsoft has recently launched a Source Forge project, named "Office Binary (doc, xls, ppt) Translator to Open XML" [16]. According to the project documentation, a translator will be first available for Word documents around summer 2008 while translators for PowerPoint and Excel will follow after that. The translators will be available under a BSD license.

3.8 Limitations and Issues:

In BIFF8, each record has a limit of 8228 bytes in length. If a record is longer than the maximum length, it may be appended by one or more CONTINUE records.

One well-known BIFF8 limitation is the handling of floating-point numbers. BIFF8 uses the IEEE 754 specification to store the floating-point numbers. Hence, it suffers the same limitations as IEEE 754. For example, positive floating points are limited to the value between 2.2250738585072E-308 and 1.79769313486232E+308. Floating-point numbers are also limited to 15-digit precision which could result in inaccurate arithmetic results for very large or very small numbers [21].

In addition, Excel versions prior to Excel 2007 have a limit of 65536 (2^16) total number of rows and 256 (2^8) total number of columns. This is perhaps due to the 1 byte limit in column reference fields and 2 bytes limit for row reference fields in BIFF8. Other limitations include the 1GB maximum memory that Excel can use, 1000 maximum characters for formula data, maximum 65536 rows by 256 columns for a Pivot Table, etc [12]. Some of the limitations seem to be restricted by Excel software, not by the BIFF8 file format.

Another well-known issue with Excel lies in its handling of date system. Excel uses two kinds of date system, one with epoch of 1/1/1900 (1900 date system) and the other with epoch of 1/1/1904 (1904 date system). To be compatible with other spreadsheet software like Lotus 1-2-3, Excel for Windows by default uses the 1900 date system. Unfortunately, the 1900 date system contains a leap-year bug where 1900 is mistakenly treated as a leap year. The 1904 date system is used as the default date format by Excel for Macintosh [20]. Due to the use of two date systems, there may be compatibility issues when copying dates among workbooks.

3.9 Perceived Popularity:

Excel has been the most predominant spreadsheet application since Excel 5.0, especially on Microsoft Windows and Mac platforms [3]. Since Excel 5.0, Excel has been bundled as part of Microsoft Office suite. Because BIFF8 is the default file format for Excel 97-2003, it is probably the most popular BIFF format. In the Florida Digital Archive, over ninety percentage of Excel files are in the BIFF8 format. In terms of Microsoft Office Binary File formats, Tom Ngo of ECMA TC 45 estimates that "more than 400 million users generate documents in the binary formats, with estimates exceeding 40 billion documents and billions more being created each year" [17].

One way of comparing the popularity of BIFF8 with other spreadsheet formats is perhaps via the use of Google Trend. The following graph (Figure 2) shows that the usage of BIFF(.xls) has maintained pretty steady for the last four years. It has been searched more in Google than the rest of spreadsheet formats such as CSV, the OpenDocument Spreadsheet format (.ods), and the new OOXML spreadsheet (.xlsx) format. It also seems that OOXML spreadsheets started to gain some momentum around the end of year 2007. Please note that the result of Google Trend is performed by querying a portion of google search database. It may not be completely accurate but is used here as supplementary information for analyzing Excel popularity.

Figure 2: A comparison of search hits in the last 4 years among several spreadsheet formats, including Excel (.xls), Office Open XML Spreadsheet (.xlsx), Open Document Spreadsheet (.ods) and Comma Separated Value (.csv). The result is gathered by using Google Trend on February 25, 2008.

4 Related Formats

4.1 Specification Variations:

Though Microsoft indicates that BIFF8 has been used as the default file format for Excel 97 to 2003 [18]; later versions of Excel, like Excel 2000, Excel 2002 and Excel 2003, have extended the BIFF8 file format to introduce new records. Additionally, there are some data structures applicable to only certain versions of Excel. These extensions and differences may cause compatibility issues among different spreadsheet applications. Besides, some elements in BIFF8 data structure are not documented, such as the stream format used by the hyperlink record. The undocumented features could prohibit proper rendering of the spreadsheets by third-party implementers.

- **Excel Workspace (.xlw):** Workspace files store the display information about the workbook such as the window sizes, print areas, and various display settings. It does not contain workbook data. The Microsoft binary format specification provides very limited information about Excel workspace formats. A complete format specification for Excel Workspace is currently not available.

- **Excel Template (.xlt):** Template files are used as the basis for Excel workbooks. The file format for Excel template is currently not documented.

- **Binary Workbook Format(.xlsb):** Excel 2007 introduces the use of a new binary file format, called Binary Workbook Format, a.k.a. BIFF12. [4]. Similar to BIFF8, the binary workbook format consists of a list of records, each starting with a header consisting of a record type (1 or 2 bytes) and a record size (1 to 4 bytes). Different to BIFF8, the record type and record size are stored using a variable length scheme. In addition, the records stored in the binary workbook format are completely different from those used in earlier BIFF versions. The binary workbook format is designed to provide better performance than the OOXML spreadsheet format, especially for very large spreadsheets. However, it does not provide full fidelity to the Excel 2007 features as OOXML does.

5 Summary and Conclusion

BIFF8 has been in widely use since 1997. It is perhaps the most popular BIFF format that is still in use today. It is a complex format; the complexity comes from the added extensions for MS Excel (version 2002 to 2007), the backward compatibility support, and its rich set of features, such as Charts, PivotTable, VBA, Information Right Management, etc. Because BIFF8 uses OLE compound document format as its underlying architecture along with several other not well-documented formats such as Document Property Stream format, it is strongly advised to use existing libraries such as ApachPOI to write the format parser.

BIFF8 is currently still supported by MS Excel 2007 and other spreadsheet software. However, Microsoft has superseded BIFF8 with its OOXML spreadsheet format. It appears that

BIFF8 may be on the path of becoming obsolete especially if major spreadsheet software stop supporting BIFF8. The Florida Digital Archive (FDA) will evaluate the feasibility for migrating BIFF8 to OOXML spreadsheet format when FDA supports OOXML spreadsheet format and the project "Office Binary Translator to Open XML" becomes available for spreadsheets.

The other possible alternative is to migrate BIFF8 to OpenOffice.org XML spreadsheet format, the Open Document Spreadsheet (ODS). ODS is part of the Open Document Format (ODF) which is already an ISO standard (ISO/IEC 26300:2006). It supports the implementation of OpenFormula, a draft OASIS standard for exchanging recalculated formula among spreadsheets. Nevertheless, migrating BIFF8 to ODS may incur more features lost than converting to the OOXML spreadsheet format. Thus, ODS appears to be more suitable as a normalization format for BIFF8. With the heated competitions between OOXML and ODF where each with their strong support bases, it appears that both formats will coexist at least for some time. The Florida Digital Archive would normalize BIFF8 to ODS format, either by a direct conversion to ODS or by migrating BIFF8 to OOXML spreadsheets which will be normalized to ODS.

6 References

[1] Daniel Rentz, "OpenOffice.org's Documentation of the Microsoft Excel File Format: Excel Version 2, 3, 4, 5, 95, 97, 2000, XP, 2003", January 10, 2008, http://sc.openoffice.org/excelfileformat.pdf

[2] Sukvinder S. Gill, "email conversation on Microsoft Media Types for registration", http://www.iana.org/assignments/media-types/application/vnd.ms-excel

[3] Wikipedia article about Microsoft Excel, http://en.wikipedia.org/wiki/Microsoft_Excel

[4] Microsoft, "Microsoft Office Excel 2007 Binary File Format Specification [*.xlsb], downloaded from http://www.microsoft.com/interop/docs/OfficeBinaryFormats.mspx on February 20, 2008

[5] Daniel Rentz, OpenOffice organization, "OpenOffice.org's Documentation of the Microsoft Compound Document File Format", Aug 07, 2007, http://sc.openoffice.org/compdocfileformat.pdf

[6]Brian Jones, Microsoft, "Mapping documents in the binary format (.doc; .xls; .ppt) to the Open XML format", January 16, 2008, http://blogs.msdn.com/brian_jones/archive/2008/01/16/mapping-documents-in-the-binary-format-doc-xls-ppt-to-the-open-xml-format.aspx

[7] Rob Weir, "A File Format Timeline", June 24, 2007 http://www.robweir.com/blog/2007/06/file-format-timeline.html

[8] Microsoft, "How to extract information from Office files by using Office file formats and schemas", http://support.microsoft.com/kb/840817/en-us

[9] Microsoft Open Specification Promise, September 12, 2006, http://www.microsoft.com/interop/osp/default.mspxs

[10] Microsoft, "Microsoft Office Development with Visual Studio", http://msdn2.microsoft.com/en-us/library/aa188489(office.10).aspx#vsofficedev_topic19

[11] Microsoft Corporation, "Microsoft Office Excel 97-2003, Binary File format Update for Office Excel 2007 [*.xls (97-2003) format]", received from Microsoft on February 15, 2008, through email requests.

[12] The team blog for Microsoft Excel and Excel Services, "Some other numbers", http://blogs.msdn.com/excel/archive/2005/09/26/474258.aspx

[13] Microsoft "Microsoft Office Excel 97-1007 Binary File Format Specification [*.xls (97-2007) format]", downloaded from http://www.microsoft.com/interop/docs/OfficeBinaryFormats.mspx on February 20, 2008

[14] Apache, "Apache POI - Java API to Access Microsoft Format Files", http://poi.apache.org/

[15] Microsoft, "Microsoft Office Binary (doc, xls, ppt) File Formats, http://www.microsoft.com/interop/docs/OfficeBinaryFormats.mspx

[16] Office Binary (doc, xls, ppt) Translator to Open XML project, http://b2xtranslator.sourceforge.net/

[17] Standard ECMA-376, "Office Open XML File Formats", December 2006, http://www.ecma-international.org/publications/standards/Ecma-376.htm

[18] Microsoft, "Using Excel 2003 with earlier versions of Excel" http://office.microsoft.com/en-us/excel/HP051985111033.aspx

[19] British Library, "Digital Object Formats", http://www.bl.uk/dp/formats

[20] Microsoft, "Description of the differences between the 1900 date system and the 1904 date system in Excel", http://support.microsoft.com/kb/214330

[21] Microsoft, "Floating-point arithmetic may give in accurate results in Excel", http://support.microsoft.com/kb/78113/EN-US/