Preservation Options for Excel 10.0/XP/2002 (BIFF8X)

Date: July 26, 2005
Author: Grace Carpenter

Introduction

Because Excel is used across many disciplines for many purposes, it would be almost impossible for users to agree on what is the preservable "essence" of a spreadsheet. For some users it may function solely as a data repository, in which case its essence can be probably be preserved as a .csv file. For others it may function as a presentation tool, and the information in it could be transmitted equally well as a .pdf or .html file. For yet others the analytical functions contained in the cells may be as crucial as the visible information, and only the spreadsheet itself can do justice to the multiple layers of information. In addition, some spreadsheets contain macros that transform them into applications, rather than static containers of data and functions.

Forward Migration

Pros:

•Possibility of preserving all functionality

Cons:

• Risk level may continue to be high

There is no reason to think that future releases of Excel will be any less at risk than current ones, if Microsoft continues to keep its specs unpublished.

• Limited to Windows (and maybe Mac) platforms

For digital repositories that are making a conscious decision to use open operating systems for the sake of preservation, these tools are not practicable.

• Large staff investment in learning/developing API for automatic migration

Although there may be some commercial forward migration tools, the most reliable way to conduct a forward migration is to open a spreadsheet in the current version of Excel, and then save it in the current format. It is probably possible to automate this process using .Net technology, but it necessitates a) the purchase of Excel for the server (providing the server runs the necessary operating system—see above); b) the staff investment in learning .Net to be able to write conversion functionality c) repeated purchases and investments in staff training over time to update to each new version.

Migration to a Non-Excel Format

<u>Excel to text/csv</u>

Pros:

• Conversion process could potentially be very simple and quick
• Not limited to a particular platform
• Many tools are available for migrating Excel files to text/csv.
• Human-readable

Cons:

• Preserves only limited aspects of a spreadsheet

Although this might be an acceptable preservation strategy in certain situations, migration to text/csv preserves only a very limited amount of the potential information in an Excel spreadsheet. For instance, formula results could be converted, but the underlying formulas would disappear. Files would have to be carefully analyzed to determine if this conversion would provide an acceptable outcome.

Excel to HTML

Pros

- Potentially simple conversion process
- Probably not limited to a particular platform
- Human-readable

Cons

- Preserves look, but not functionality, of a spreadsheet

Like the text/csv migration, a migration to HTML would be simple to perform with existing tools, but it would preserve only a very narrow amount of information and functionality.

Excel to PDF

Pros:

- Would preserve more elements of a spreadsheet (e.g., charts) than a text/csv migration
- Probably not limited to a particular platform
- Potentially simple conversion process

Cons:

- Underlying functionality (formulas, macros) would be lost
- Conversion to a published, but proprietary, format still entails some preservation risk
- Not a human-readable format

Excel to XML

Pros:

- Possibility of preserving majority of elements in original spreadsheets (e.g. formulas, charts, outlines, etc)
- Human-readable, non-proprietary format

Cons:

- Assumes either a further migration or special rendering tool

Although XML is human readable, the XML produced to capture most spreadsheets would be nevertheless be too complex for a human to parse, so it would have to be transformed somehow before rendering.

Note:

I haven't actually seen Microsoft's XML for Office documents. Microsoft claims to use XML as a native format in 2003, but it's not clear to me whether they've actually abandoned their binary format. However, it is clear that XML support is integrated into Excel and that it is trivial to save a spreadsheet as XML.

There is still the issue of how to actually perform the migration, using Excel, in a Linux or Unix server environment. One way around this might be to open an Excel spreadsheet in OpenOffice software--which uses XML as its native format--and then save it in OpenOffice's XML. Currently there are some incompatibilities between the OpenOffice and Excel, and this almost guarantees some loss of original elements in the spreadsheet. However, a major focus of OpenOffice 2.0—originally scheduled to be released in the spring of 2005, but still in beta--is to reduce incompatibilities with Microsoft Office. In addition, the OpenOffice 2.0 XML format is being considered as a candidate for an OASIS standard for office documents.

Other notes

Given the wide range of elements that can be contained in an Excel spreadsheet, it seems clear that harvesting in-depth technical metadata about a file is an essential first step in its preservation. A good tool could identify spreadsheets that contain pure data, without formulas; or spreadsheets that contain not only formulas but other elements not easily migrated to a plain text/csv format, such as outlines, charts, and macros. Although the National Library of New Zealand's Metadata Extractor tool does harvest Excel metadata, it doesn't appear to harvest very detailed information. JHOVE does not have an Excel module.

Appendix: Possible Migration Tools

Open Source:

OpenOffice http://www.openoffice.org/
Multi-platform, capable of doing conversions from one format to another.
Questions: API? How accurate are conversions? How up-to-date?

Xlhtml http://chicago.sourceforge.net/xlhtml/
Unix/Linux, Mac. Converts Excel into HTML. Latest version is dated 3/19/02.

Commercial:

Too many to enumerate here.