Preservation Options for Word 10.0 /XP/2002

Date: July 26, 2005
Author: Grace Carpenter

Introduction

In its simplest form, a Word document is just text. In its most complex, it contains not only text but embedded spreadsheets or graphics and also macros. In between these two extremes are a range of aspects that a user may or may not find essential to a document, and any preservation strategy must take into account not only what the document contains, but what the user feels is its essential core.

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think that future releases of Word will be any less at risk than current ones, if Microsoft continues to keep its specs unpublished.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 document in the current version of Wordl, 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 Word 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-Word Format

<u>Word 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 Wordl files to text/csv.
• Human-readable

Cons:

• Preserves only limited aspects of a document

For any document in which the formatting is not essential to interpreting the document, this is a viable (although perhaps unappealing) option. Files would have to be carefully analyzed to determine if this conversion would provide an acceptable outcome.

<u>Word to HTML</u>

Pros

- Potentially simple conversion process
- Probably not limited to a particular platform
- May be able to preserve look-and-feel
- Human-readable

Cons

- Preserves look, but not functionality, of a Word document

Like the text/csv migration, a migration to HTML would be simple to perform with existing tools, but it would not be able to preserve more interactive parts of Word documents (such as macros)

Word to PDF

Pros:

- Would preserve most aspects of a document, other than macros
- Not limited to a particular platform
- Potentially simple conversion process

Cons:

- macros
- Conversion to a published, but proprietary, format still entails some preservation risk
- Not a human-readable format

Word to XML

Pros:

- Possibility of preserving majority of elements in original document
- 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 documents 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 Word and that it is trivial to save a Word document as XML.

There is still the issue of how to actually perform the migration, using Word, in a Linux or Unix server environment. One way around this might be to open a Word document in OpenOffice software--which uses XML as its native format--and then save it in OpenOffice’s XML. I gather from what I've read that there are some inconsistencies in formatting between OpenOffice and Word. 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.

Appendix: Possible Migration Tools

Open Source:

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

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?

WvWare http://wvware.sourceforge.net/
Linux/Unix-based set of utilities/libraries (?) for doing conversions from Word to a variety of formats, including text, HTML, and PDF. Dates on web page suggests it is not an active project; latest “news” is from October 2001.

Commercial:

Too many to enumerate here.