Why Microsoft Word may be bad for your health

davee@sungate.co.uk [a]

[a] *http://www.sungate.co.uk/*

Microsoft Word has become a de-facto standard for word-processing and document exchange, which is unfortunate because it has many limitations and undesirable features. A common refrain is that people use Microsoft Word simply because "everybody else uses it". The effective monopoly of Microsoft Word in the mass-market word-processing arena is not generally considered a Good Thing.

Abstract

This report raises a number of issues that the average user of Microsoft Word may not be aware of, or may not consider important. However, some of these issues are potentially extremely serious in terms of confidentiality and privacy and therefore need to be highlighted. The points raised apply equally to all versions of Microsoft Word on all platforms, unless otherwise indicated. As far as possible, the report is fair and balanced despite the author generally holding less than favourable views of Microsoft.

Key words: Word-processing; privacy; confidentiality

1 Reasons for concern

There are a number of reasons why one might want to think twice before using Microsoft Word to write new documents. In addition, there can be issues in the reading of Microsoft Word documents that one has been sent.

1.1 Hidden content

Microsoft Word document file sizes are huge compared to the amount of text they contain. What is in the rest of the file? Without being aware of what constitutes the remainder, one should be extremely wary of distributing [1] Microsoft Word document files electronically, since they may contain information that for reasons of confidentiality or privacy should not be made available to recipients or to anyone else. Some examples of hidden content appear in section 2 on page 2.

[1] Distributing means passing from one individual to one or more others either on floppy disk, on CD, via email or making available for download across a network, including the Internet.

1.2 Extra work required

On many occasions, Microsoft Word documents are distributed when plain text will do just as well or would in fact be far more suitable. A plain text copy will be smaller to send, will get to the recipient quicker (especially for large documents) and will be easier for *everyone* to read quickly and easily, particularly of course those recipients who do not have Microsoft Word installed. Furthermore, it will even be quicker to read for those who *do* have Microsoft Word installed, because there will be no need to launch an external application in order to read it.

1.3 Inter-version compatibility

The technical specification for the *DOC* file format is a Microsoft corporate secret and changes with each new version of Microsoft Word. This is anti-competitive behaviour which makes it almost impossible for software other than Microsoft Word to manipulate *DOC* files; other word processing software cannot compete on a level playing-field. This policy led a judge in the United States to describe Microsoft's actions as those of a "predatory monopoly". Moreover, these different formats are not fully compatible with each other and exhibit what is known as "backward incompatibility". A truly reputable software company would ensure *not*

only that new versions of the software can read documents created in older versions of the software *but also* that older versions of the software can still read documents created by a newer version. Microsoft's solution to version incompatibilities such as this is often simply to reply "You need a newer version, please send us more money." This does *not* sound like the voice of a reputable company.

The on-screen and printed appearance of *DOC* files depends on many factors and therefore a document may appear differently across multiple systems. How does one know whether a recipient will see the document as intended, or even whether it will be readable at all? For recipients with which one is unfamiliar, it could be considered discourteous or possibly even arrogant to expect that they ought to have Microsoft Word installed. One should attempt to ascertain their preferred format in advance. Microsoft Word is not ubiquitous and some may choose not to use it for document exchange, for reasons related to those outlined in this document.

1.4 Hidden formatting

Hidden formatting in a *DOC* file is a functional, rather than a privacy or security issue, but will have been experienced by almost anyone who has ever worked with a moderately complex Microsoft Word document. Copying formatted text from one *DOC* to another, for example, often gives unwanted results that cannot be easily reformatted as desired without rewriting the text in full. These problems arise because special formatting is stored in the file that cannot be accessed directly. Preparing document content in plain text form avoids this problem.

1.5 Long-term document archival

People keep documents for many years. This happens because they want to be able to read their documents in the future. Why should it be necessary to continue paying for new software just to be able to read one's own documents? If one changes software in the future, will these documents still be readable? Only ten years ago, Microsoft Word did not enjoy the monopoly it does today. How many pre-Microsoft Word documents does one still have the software available to read? An open file format not controlled exclusively by any single corporation is required.

1.6 Viruses

The increased use of Microsoft Word documents has fuelled the rise in the transmission of macro viruses. This issue has been discussed before and will not be repeated here. Briefly, even though virus protection may be installed, it can be defective or out of date. Why take the chance?

2 Hidden content in Word documents

The *DOC* file format is a "closed" secret, so there is really no way to determine for certain what "metadata" has been stored in it. Remarkably, even Microsoft themselves admit that *"[Some metadata] is only accessible through extraordinary means, such as by opening a document in a low-level binary file editor."* In response to criticism from users, Microsoft have issued some guidelines describing how to remove different types of metadata. The author has tested some of these procedures [2] and found that they are far from straightforward for the typical user and not 100% effective.

Analysis of *DOC* file contents has shown that the following are typical examples of what is stored in most files.

2.1 Author information

When a new document is created, authorship information is saved with the document. In addition, similar authorship information for the template on which the document is based is saved with the document. "Authorship information" has a very broad scope, including: author's name, author's initials, author's company or organization name, author's computer name, name of the server on which the document is stored, other file properties and summary information, names of previous document authors and template information. It would appear to be possible to remove some authorship information from the document after creating it, although this process is not perfect; it is almost impossible to remove the template-based details, without modifying the template itself and recreating the document from this modified template. Otherwise, the name of the template author is saved to the *DOC* file and there is no simple way to remove it! The newest version of Microsoft Word [3] has a single facility which purports to remove all personal information from a document. Assuming this works as advertised, it should have been in place long ago and one still needs to be aware of the existence of document metadata and know that it can perhaps be removed.

2.2 Text from previous revisions of a document

Many users are unaware that Microsoft Word has a concept called "file versioning" (not to be confused with different versions of the software package itself). When activated, this allows a series of distinct versions to be

[2] using Microsoft Word 97

[3] Microsoft Word 2002 at the time of writing

stored within the document, reflecting the revision history of the document. Every previous version of the document can be retrieved. Despite the fact that this increases file sizes substantially, it is a useful facility, but it must be used with caution. Imagine a previous revision of a manuscript which indicated an intent to give formal acknowledgement to certain individuals but that this decision was subsequently changed. Comments made while the document is still a draft can come back to cause embarrassment and confidentiality issues at a later date.

Note also that file-versioning is switched on by default for some versions of Microsoft Word. Comments and previous revisions can be removed from a document, *but only if one is aware of their presence in the first place.* For example [4], a small icon appears in the status bar to indicate that file-versioning is active for the current document; this may be spotted if looked-for, but otherwise would probably not be noticed or any significance attached to it.

2.3 Other hidden content

The "closed" nature of the *DOC* file format means that, even when trying hard to "clean up" one's document, there is no way to tell for certain what metadata remains in the file. The difference between a *DOC* file and a document stored in plain text is that from a privacy standpoint the former might reflect *"what you see is what you get"* whereas the latter is *"what you see is* ALL *you get"*.

3 Workarounds

It is all very well to suggest that Microsoft Word has failings and shortcomings but users need to deal with these issues. Various workarounds are available.

3.1 Save documents in other formats

The first option is to continue to use Microsoft Word but, instead of distributing documents in Microsoft Word *DOC* format, to use an alternative file format such as plain text or perhaps *RTF*, "Rich text format", which retains much of the formatting of *DOC* files without the associated metadata and virus risks. Editing plain text versions of draft manuscripts is highly recommended from a scientific point of view, because it allows one to focus on the content, rather than the appearance, of a document. It is well known that when a journal receives a document submission in Microsoft Word format, or indeed any other format, the document is first converted to plain text before being prepared for publication. Time spent tinkering with superficial formatting issues such as fonts and layout is almost certainly wasted.

[4] based on inspection of Microsoft Word 97 for Windows

3.2 Printed or published copies

In some circumstances, it is possible to distribute copies of one's document in printed or "published" form, such as in PDF format. This carries none of the risks associated with the *DOC* file itself.

4 Alternative software

There are, potentially, many alternatives to Microsoft Word. Some software packages are similar in nature to Microsoft Word but may be specially tailored to particular areas of expertise; other packages take a completely different approach.

4.1 Other word-processing software

There are other word-processing packages available for Windows, with interfaces that may be familiar to those used to using Microsoft Word. Examples are WordPerfect and Lotus Word Pro. The difficulty with switching to one of these packages is, of course, that they too have their own, propriertary document file format, in the same way as Microsoft Word. In fact, given that they are less popular, their security and privacy issues will have undergone less scrutiny. One could apply similar suspicion to the contents of these "closed" document formats as to *DOC* files. The same interoperability problems will also arise as with sharing *DOC* files.

Of course, there is no particular need to restrict ourselves to using Windows for document creation. Many other operating systems have suitable software available, for example Unix and MacOS.

4.2 LaTeX

This document was typeset using LaTeX. Those with a mathematical or statistical background may be familiar with it because it has a sound reputation for being highly suitable for scientific writing, especially for work containing equations. It approaches document-writing from the perspective of "logical structure" based on a defined document style and makes creation of professional manuscripts remarkably straightforward, once one is familiar with this concept. All LaTeX source files are plain text files and as such are perfectly portable, shareable and safe. The final, printed document is normally *Postscript* or *Portable Document Format* (PDF), which is designed to appear identically across platforms and all postscript-compatible printers. These file formats are also "open" and documented, so that the absence of metadata can be easily verified by inspecting the files before distribution. Traditionally, LaTeX has been used under Unix, but versions for Windows are also available.

5 Discussion

There are of course many circumstances when using Microsoft Word is a perfectly reasonable thing to do, and may in fact be the preferred tool. As discussed, however, there are always many issues to bear in mind when doing so, especially when sharing documents with others. Some of these relate to functionality and interoperabilty, others to confidentiality and privacy.

There is a huge amount of additional information, the so-called "metadata" that is sent along with the document text when sent as a *DOC* file, that any competent professional would never dream of including when sending a paper copy, for example. Similarly, one would not normally want older, superceded drafts of a document to be sent in addition to the final version; this is effectively what happens when distributing a *DOC* file consisting of multiple document revisions.

The problem of the *DOC* file format is very real. Consider the huge number of documents that have been written and saved in this format. Decades from now, will these documents still be readable by contemporary software? Only Microsoft knows the format of all those files. One thing that can be relied upon, however, is that plain text files will always be readable, even if the software used to create them has long disappeared.

In an ideal world, there would be a standard document file format, so that everyone could exchange documents freely and easily but would not be tied to using a single piece of software. Putting to one side the option of simple plain text, such a thing almost exists already. In fact, there are several "open" file formats which could be used for documents, all of which are specially-formatted text files. One of these is *HTML* (hyper-text markup language), which is the language used to write web pages. The simple idea of *HTML* is that documents are text files which can be edited by any piece of software one happens to be using (even Microsoft Word) and can be displayed using a web browser; the web browser interprets the content of the *HTML* text file. Another technology, known as *XML* (eXtensible markup language), is in its infancy and works in a similar way, but is much more flexible and versatile.

Hopefully, this document will have conveyed the message that the distribution of *DOC* files generated by Microsoft Word can be potentially hazardous. Distribution can breach confidentiality, result in an invasion of privacy, cause a variety of interoperability issues and transmit viruses. The author strongly urges all users of Microsoft Word to bear these warnings in mind at all times and to consider not distributing *DOC* files!
