



## OpenDocument for Libraries and Archives

### *Preserving history*

It is an irony that the computer revolution has made it harder, not easier, to preserve our works. Today we can read a Shakespeare play written 400 years ago, but we can't read a Word document that we saved 5 years ago.

The digital medium might preserve a perfect copy of a document, but what good is that if 50 years from now nobody has a copy of Word 97 to read it with? The computer revolution is not 50 years old. The first PC appeared 25 years ago. Windows 95 appeared 10 years ago. Documents written 5 years ago are usually unreadable today. How can we hope to preserve information that people can read 400 years from now?

### **Understanding the problem**

When NASA scientists decided to go back to the Viking data, they found that the software used to read it no longer existed and the engineers who designed those computers were dead. Fortunately, they found paper printouts of the Viking data. But what if they hadn't? How can we ensure that the digital documents we make today will be read tomorrow?

### **OpenDocument is XML**

XML files are fundamentally text files with structure. Take this example:

```
<text:h text:style-name="Heading">
  European Union
</text:h>
<text:p text:style-name="Standard">
  The European Union is a supranational union of 25 member states from the European
  continent. It was established under that name in 1992 by the Treaty on European Union
  (the Maastricht Treaty).
</text:p>
```

Even if all knowledge of the format is lost, one can always read the plain text to obtain the information stored. XML is the most future-proof way of storing a digital document.

### **OpenDocument is human readable**

Not all XML files are created equal. Some, like Microsoft's OXML, are designed purely to be used by a computer. Others, like OpenDocument, are designed to be as understandable to a human as possible. This would allow a future historian to discern more information about the document, even if all knowledge of the document format had been lost.

### **OpenDocument is an open standard**

Of course, the best situation is to never lose knowledge of how the format works. The best way to ensure this is to use an open standard. An open standard is one that is maintained by an independent standards group. It is in the public record and can be implemented by anyone.

If you store a document in a format owned by a single vendor, when that vendor is gone, knowledge of the format goes with it. If you use an open standard, a public record remains. OpenDocument is an open standard, maintained by both the ISO and the OASIS standard groups.

## OpenDocument is platform independent

- The files used by NASA to store the Viking data were tied to one type of computer. When NASA wanted those files, these computers no longer existed.
- Microsoft Word documents are tied to one platform. To one company. Windows 95 came out 10 years ago. Can you depend on this company being here 200 years from now?

The OpenDocument format is not tied to any type of computer, or any application, or any type of software. So it will still be accessible to whatever systems our descendants are running 300 years from now.

## Did any archivists help design ODF?

Yes. The OASIS committee that designed the OpenDocument format included representatives from two important archivist groups:

- **The National Archives of Australia**  
This is a group tasked with preserving a country's history. How long do you want to preserve your country's history for?
- **The Society for Biblical Literature**  
Here is a group that needs to deal with complex multilingual documents (including several dead languages) and needs them to be fully accessible a thousand years from now without any loss.

## Sustainability factors

The Library of Congress, through its National Digital Information Infrastructure and Preservation Program, has produced a list of Sustainability Factors for the suitability of a digital format for preserving digital information.

## Disclosure

*"Disclosure refers to the degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content."*

The full OpenDocument specification is public, available on the OASIS committee web page and is free to redistribute. As an XML format, any RelaxNG validator can validate OpenDocument files. The OpenDocument spec is maintained at two open standard organizations (OASIS and ISO).

## Adoption

*"If a format is widely adopted, it is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions."*

OpenDocument is adopted by many applications including the major vendors save for Microsoft.

*"In some cases, the existence and exploitation of underlying patents may inhibit adoption, particularly if license terms include royalties"*

OpenDocument has no patent restrictions, no license restrictions, and no royalties. It can be freely adopted by any software maker including open source software.

## Transparency

*"Transparency refers to the degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text-only editor."*

Transparency is a primary design goal in OpenDocument. The format is plain text XML, and the tags are designed to be as human readable as possible.

*“Transparency is enhanced if textual content is encoded in standard character encodings (e.g., UNICODE in the UTF-8 encoding)”*

OpenDocument text is encoded in Unicode in the UTF-8 encoding.

*“compression inhibits transparency... Archival repositories must certainly accept content compressed using publicly disclosed and widely adopted algorithms that are either lossless or have a degree of lossy compression that is acceptable”*

In OpenDocument, compression is optional. When compressed, the algorithm used is ‘ZIP’. This is a publicly disclosed, widely adopted, lossless compression algorithm.

## **Self-documentation**

*“Digital objects that are self-documenting are likely to be easier to sustain over the long term and less vulnerable to catastrophe than data objects that are stored separately from all the metadata needed to render the data as usable information or understand its context.”*

OpenDocument is self-documented. It is divided into separate portions including content, metadata and "styles". All the rendering information is stored in styles, which are defined in the styles section of the document.

## **External dependencies**

*“External dependencies refers to the degree to which a particular format depends on particular hardware, operating system, or software for rendering or use”*

OpenDocument is designed to be platform independent. It does not depend on any particular hardware, operating system or software. Indeed, it is already supported by many applications for at least 8 operating systems (Linux, Windows, Mac, Solaris, FreeBSD, Symbian, OpenBSD, AIX, React OS) and 4 hardware architectures (Sparc, PowerPC, Intel, ARM).

## **Impact of patents**

*“Patents related to a digital format may inhibit the ability of archival institutions to sustain content in that format.”*

OpenDocument comes with no patent restrictions of any kind.

*“the existence of patents may slow the development of open source encoders and decoders and prices for commercial software for transcoding content in obsolescent formats may incorporate high license fees”*

There are already two open source applications with full support for OpenDocument (OpenOffice.org and KOffice).

## **Technical protection mechanisms**

*“Content for which a trusted repository takes long-term responsibility must not be protected by technical mechanisms such as encryption, implemented in ways that prevent custodians from taking appropriate steps to preserve the digital content and make it accessible to future generations.”*

While ODF files may optionally be encrypted, by default they are not. If encrypted, the encryption follows an industry standard (RFC2898).