![intellect logo] REPRESENTING THE UK TECHNOLOGY INDUSTRY

▶ **Document management**

Why the format of office documents matters to your business

Contents

About DeltaXML

DeltaXML is a specialist software company providing tools to manage change in all XML data.
Equally applicable to large corporates and OEMs, DeltaXML software provides the solution for
complete control of XML.

Aimed at professional enterprise system developers, DeltaXML tools allow anyone working on
an XML change issue to quickly deploy a scalable solution within their application. The key
products are designed for use across the enterprise to identify and process change and can be
integrated with both server and client applications.

More information is available at www.deltaxml.com

**This paper is part of a series of white papers issued by Intellect's Document Management Group
following the successful launch of its Guide (Intellect Document Management Group, 2007).**

Audience

The importance of 'standard' document formats is growing. This paper seeks to raise awareness of the development and increased use of new open document formats, and the implications these can have for businesses.

It is aimed at managers who need to access information held in office documents, perhaps as part of a document management system, as part of a business process, or to exchange data between different office applications. With these new open standards for document formats, there are exciting opportunities for better integration in this area. Users need to be aware about what is available and what will be in the future, and the implications of these changes.

This paper is the latest in a series published by Intellect's Document Management Group, which seeks to demystify document management, promote its benefits to the wider business community, identify common problems and propose solutions to some of the difficulties faced by organisations who are newly adopting document management solutions, as well as organisations who are looking to upgrade their document management facilities.

Introduction

In recent years, there has been considerable activity and progress in open standards to enable access to the content of office documents. More and more office documents are now being stored in accessible XML formats. You will notice one of these new formats by the change from .doc to .docx on text-based documents. This has important implications for users of office products, because the content of these documents is becoming accessible to document management systems: applications can share and exchange data, and third-party tools and add-ons are becoming available.

Awareness of these new formats is important: as more and more large companies and government bodies begin adopting them, the likelihood of encountering these new formats will significantly increase over the coming months and years.

As we cannot control what formats others use, we need to be prepared to handle them.

This paper addresses the following issues:

▶ What has been happening to standardise office document formats?
▶ Who is using open standards for office documents?
▶ What implications do these standards have on access to information in office documents?
▶ What are the benefits of open standards?
▶ What are these new formats and how can I access the information in office documents?
▶ What differentiates the standards?
▶ Which office products use which standard document formats?
▶ Conclusion: what does all this mean to users of office products?

What has been happening to standardise office document formats?

Pressure has been growing, particularly from many of the large user organisations, for office software applications to allow access to the content of documents through standard, open formats. Microsoft Office is very widely used, but has only recently begun to allow access through an open standard format. At the same time, open-source and very low-priced office applications are becoming serious alternatives. Many of these are built on the OASIS and ISO standard OpenDocument Format (ODF).

The result has been considerable standardisation activity in this area over the past few years. However, the similar names of these standards and software applications – for instance, OpenOffice.org (the open-source application) and Office Open XML (Microsoft's XML format for Microsoft Office applications) – have led to some confusion.

There are two main emerging standards in office document formats:

1. The **OpenDocument Format (ODF)** is an XML-based document format, developed by the Open Document Format for Office Applications (OpenDocument) Technical Committee at OASIS[1]. In May 2005, the format was approved as a standard through the standardisation organisation, OASIS, and in May 2006 as an official international standard through ISO and IEC[2]. ODF is positioned as an open document format that is not controlled by any single software vendor. Version 1.2 of ODF, due to be released in late 2008, will add a W3C standards based metadata model, which is intended to make processing ODF documents much easier and more powerful.

2. **Office Open XML (OOXML):** At the same time, Microsoft has been developing another format based on XML for all its Office products (Word, Excel and PowerPoint). Office Open XML is increasingly referred to more simply as "OpenXML". The format was approved as an Ecma standard in December 2006 and more recently, in April 2008[3], as an ISO/IEC standard, though its publication has been postponed by appeals against its approval.

1 http://www.oasis-open.org/committees/membership.php?wg_abbrev=office
2 http://www.openoffice.org/press/iso_pr.html
3 http://www.ecma-international.org/news/TC45_current_work/ISO_and_IEC_approve_Office_Open_XML.htm

Who is using open standards for office documents?

Adoption of ODF is probably most advanced in government organisations. Agencies of the State of Massachusetts (in the United States), Belgium, Finland, Germany, the Netherlands, Norway, the United Kingdom (in particular Bristol City Council), and Slovakia are all adopting ODF in some way, though not necessarily across the board. In other parts of the world, governments have adopted open formats to some extent in Australia, Argentina, Brazil, India, Japan, Malaysia, Peru, South Africa, and Vietnam[4].

Adoption of ODF has occurred for a variety or reasons, including short and long-term access; features, benefits and reliability of the office applications; cost; and standards support.

Current versions of Microsoft Office use OOXML, as defined by the ECMA standard, when documents are saved.

OpenOffice.org was downloaded around 1.2 million times per week during June 2008, and the total number of downloads (rather than users) to mid 2008 is roughly similar to the sales of Microsoft's Office 2007 between January and September 2007. With Microsoft announcing its support of ODF, all the major office products will support ODF by mid 2009. An increasing number of hardware vendors are also shipping low-cost computers with Open Office.org or other ODF software pre-installed.

What implications do these standards have on access to information in office documents?

Both ODF and OOXML are based on XML. This means that you can easily access the content of these documents using standard XML software. Indeed, you can even look at the XML in a standard text editor. This makes the content of these documents much more accessible than binary formats (as used by previous versions of Microsoft Office for example), or even RTF.

This access through standard software has enormous implications for content management systems and any automated or semi-automated process that needs to access the content of documents.

For example, you could write software to check that a document conforms to a standard structure, or you could automatically insert standard boilerplate wording in particular sections, and update it in multiple documents as it changes. You could extract the content to convert it into another format for web display or to be transferred to another system. All this can be done without the need for the original application.

4 http://www.odfalliance.org/resources/AnnualReport2007.pdf

What are the benefits of open standards?

One of the major demands driving these standards is long-term archiving. People want to be assured that they can access old data after many years or even decades. Software to access documents held in an open standard format is more likely to be available, whereas a proprietary format may go out of production and become unsupported and the information inaccessible. A proprietary format may also require the original application in order to access the data. However, even data formatted according to an open standard that is very old may become difficult to process and access: open formats are therefore not yet able to solve this problem fully.

Open standards also encourage add-on components to be developed, for example those for OpenDocument Format[5]. Some of these provide an alternative way to access the data without needing the application that created the data.

New and innovate applications are being developed, for example support for export into other formats, applications dedicated to slide show presentation as opposed to editing, support for translation, tools to merge or compare documents.

When several office applications support the same open standard, a user has a choice and can select based on whatever criteria are most important, e.g. ability to handle large documents, reliability, ease of use, cost. Lock-in to a particular application vendor is therefore reduced. This choice can also lead to cost savings for the user.

In some government bodies, providing data in an open standard format may be mandatory. This is an improvement over the situation where data can only be provided in a proprietary format, which may require citizens, or corporations, to purchase proprietary software.

5 http://wiki.services.openoffice.org/w/index.php?title=OpenOffice.org_Solutions&oldid=75508

What are these new formats and how can I access the information in office documents?

ODF and OOXML are both formats that make use of the ZIP file compression mechanism. Each ODF or OOXML document consists of a ZIP file which contains multiple files within it. The files they contain may be graphics, text or style information.

It is only necessary to unzip the documents to access the underlying files. For example, the ODF file for this document contains a file named content.xml file which includes, for the title of this section:

```
<text:h text:style-name="Heading_20_2" text:outline-level="2">
        What are these new formats and how can I access the information in office documents?
</text:h>
```

While the title is surrounded by markup tags, <text:h> and some styling details, the data still remains accessible. You could edit this, zip the document up again and open it in OpenOffice.org to see the change to the text you made.

Similarly, when this document is saved as an OOXML file, the document.xml file contains this:

```
<w:p w:rsidR="00AA3706" w:rsidRDefault="00AA3706">
        <w:pPr>
                <w:pStyle w:val="Heading2" />
                <w:tabs><w:tab w:val="left" w:pos="0" /></w:tabs>
        </w:pPr>
        <w:r>
                <w:t>
                What are these new formats and how can I access the information in office documents?
                </w:t>
        </w:r>
</w:p>
```

Again, the text content is clearly accessible.

There are many XML processing tools now available, allowing the processing of documents in these formats to be easily be integrated into a content management or other enterprise information systems.

What differentiates these two standards?

Or, as many have asked, why do we need more than one format for office documents? Do we need dual standards, or is this the start of a standards duel?

These two standards do address different needs. ODF is a general-purpose office document format designed to allow a rich interchange between different applications, and sharing of formats across text editing, spreadsheets and presentations.

OOXML was designed by Microsoft to support not only the latest version of its office products, but also previous versions: one of its strengths is that it preserves some of peculiarities of older versions of Microsoft Office, and so allows consistent processing of legacy content. It is not designed primarily as an exchange format with other systems but more as an accessible and long-term archive format for Microsoft Office applications. This provides an alternative to the PDF archiving format, PDF/A, with the benefit of more structural information.

Underlying each standard is an abstract model of office documents. ODF and OOXML have different underlying models. For example, the way that tables are handled is different, and the way that auto-numbering is handled is also different. So conversion between these two, although possible, is not perfect – it is important to note that some data may be changed or lost.

In terms of size, the ODF specification is just over 700 pages long, while the OOXML specification is nearer 6,000 pages. This makes the threshold to access OOXML much higher than that for ODF.

Which office products use which standard document formats?

The table below shows how some applications support the different formats. It is meant as a general guide only and does not go into the detail of the different versions, as these are constantly updated and altered. Any prospective user should check the versions of both formats and applications carefully, as there are significant differences between versions.

In May 2008, Microsoft announced that full ODF support would be available in Microsoft Office with Service Pack 2 (SP2) in the first half of 2009[6]. When using SP2, it is intended that customers will be able to open, edit and save ODF documents.

In the table below, the 'default' format is the format that is normally used by the application, and therefore the one that will preserve all the internal structure of the content. Some information may be lost when a non-default format is used for content edited in that application.

Application	ODF	OOXML
OpenOffice.org	Yes - default	Yes
NeoOffice	Yes - default	Yes
StarOffice	Yes - default	Yes
IBM Lotus Symphony	Yes - default	No
Novell OpenOffice	Yes - default	Yes
Google Docs	Yes - default	No
WordPerfect Office X4	Yes	Yes
Microsoft Office	With add-on[7]	Yes - default

6 http://www.microsoft.com/presspass/press/2008/may08/05-21ExpandedFormatsPR.mspx
7 Sun offers the ODF Plug-in for Microsoft Office which allows Microsoft Office users to open and save ODF documents from within Microsoft Office

Conclusion: what does this all mean to users of office products?

Although simple documents can be exchanged between applications using these standard formats quite well, users will experience more difficulty in exchanging complex documents. The situation is however improving. Some organisations are performing automatic conversion of all incoming documents, for example, as part of an email system. Many of the open-source and low-cost office products support ODF, and with Microsoft's expected support for ODF in the first half of 2009, the ability to interchange documents should be considerably improved.

These developments in office document standards mean that where previously access to the content of office documents was difficult and required processing of proprietary formats, content is now increasingly available in formats defined by open standards and based on XML. The content is therefore accessible using a wide variety of commonly available programming tools. This is making the exchange of office documents easier, providing more choice in office applications and encouraging the development of new applications for processing office documents in innovative ways. These benefits can have tangible positive repercussions on organisations' ability to effectively manage and share their documents, resulting in greater overall business efficiency.

Glossary

ODF	An OASIS standard for text, spreadsheet and presentation documents. Version 1.0 has been standardised as ISO/IEC 26300.
OOXML (Office Open XML)	An Ecma standard (Ecma 376[8]) originally developed by Microsoft for Microsoft Office products (Word, Excel and PowerPoint). It has recently been approved, subject to appeal, as an ISO/IEC Standard (ISO/IEC 29500).
OpenOffice.org	A free, open-source software product for editing text, spreadsheet and presentation documents which stores data using the ODF standard.
RTF (Rich Text Format)	A textual representation for text documents developed by Microsoft for cross-platform document exchange between different applications. RTF includes styling information. It pre-dates XML.
XML (eXtensible Markup Language)	A text-based markup language in wide use in many areas including XHTML and office applications.
XHTML	The current standard for HTML which also conforms to the XML standard.
ZIP	A compression format, used by ODF and OOXML, which converts one or more files into a compact representation.
W3C	The World Wide Web Consortium.
PDF	Portable Document Format.
PDF/A	An ISO Standard (ISO 19005-1:2005[9]) for using PDF format for the long-term archiving of electronic documents.
ECMA	European Computer Manufacturers Association (ECMA)[10], an international standards organisation for Information Communication Technology (ICT) and Consumer Electronics (CE).
ISO/IEC	ISO and IEC formed a Joint Technical Committee known as the ISO/IEC JTC1[11] to deal with overlap in areas of standardization and work related to information technology. Its official mandate is to develop, maintain, promote and facilitate IT standards.
ISO	International Organization for Standardisation[12].
IEC	International Electrotechnical Commission[13].
OASIS	Organisation for the Advancement of Structured Information Standards[14].

8 http://www.ecma-international.org/publications/standards/Ecma-376.htm
9 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920
10 http://www.ecma-international.org/
11 http://www.jtc1.org/
12 http://www.iso.org/
13 http://www.iec.ch/
14 http://www.oasis-open.org/

Intellect's document management series

The group's first paper, **'Document management concerns the whole board - a guide for all directors'**, was launched at a reception on 21 November 2007.

Aimed at non-technical board-level executives, it has been written to help organisations and come to terms with the information-heavy future that organisations everywhere are facing.

'Implementing document management - recommended practices and lessons learned' provides advice and guidance on how to implement document management solutions, looking at issues that arise specifically in document management projects.

'An approach to maximising your investment' is aimed at organisations that are already familiar with the concept of document management, and have already bought into the business case for it. It assists those who are ready to embark on a new document management programme of work, or indeed already have a document management solution in place but have plans for further work in this area.

'Collaborative Working Environments' is aimed at public and private sector organisations that face the challenges of forming teams across disciplines, on different sites, or working across time zones, when running projects and programmes.

'Safeguarding information, reputation and corporate productivity - a guide for information governance' is aimed at both public and private sector organisations, and provides guidance on how the requirements set out in the Cabinet Office paper 'Data Handling Procedures in Government' can met in large part by document management solutions.

'Addressing Information Risk and Compliance' provides guidance on how to comply with internal and external regulatory pressures, and minimise risk to businesses though document management'

'Document management for SMEs' looks at the business case for document from the SME perspective.

'Why the format of office documents matters to your business' looks at the growth of 'standard' document formats, and the implications of open formats for businesses.

'Migration: a hidden danger lying in wait' discusses the rationale behind migration, what issues to consider when migrating information, and how to avoid some of the common pitfalls associated with this element of document management.

To download any of the above reports visit www.intellectuk.org/docman

To keep up-to-date with all the developments taking place in the document management sector, please email documentmanagement@intellectuk.org

For all the latest news from Intellect, why not subscribe to our industry newsletters?
Visit www.intellectuk.org/newsletters to subscribe online.

Intellect is the trade association for the UK technology industry.

Intellect provides a collective voice for its members and drives connections with government and business to create a commercial environment in which they can thrive. Intellect represents over 800 companies ranging from SMEs to multinationals. As the central hub for this networked community, Intellect is able to draw upon a wealth of experience and expertise to ensure that its members are best placed to tackle challenges now and in the future.

Our members' products and services enable hundreds of millions of phone calls and emails every day, allow the 60 million people in the UK to watch television and listen to the radio, power London's world leading financial services industry, save thousands of lives through accurate blood matching and screening technology, have made possible the Oyster system, which Londoners use to make 28 million journeys every week, and are pushing Formula One drivers closer to their World Championship goal.

In the past 12 months 14,500 people have visited Intellect's offices to participate in over 550 meetings and 3,900 delegates have attended the external conferences and events we organise.

The technology industry contributes over 10% of UK GDP and directly employs over a million people in the UK.

For more information visit www.intellectuk.org

intellect
REPRESENTING THE UK TECHNOLOGY INDUSTRY