Why Is PDF/A Important for SharePoint Implementations?

SharePoint, Document-Management-System (DMS), File Formats, Portable Document Format (PDF), PDF/A, OpenDocument Format (ODF), Open Office XML (OOXML), Open Source

Rarely has any other platform become as widespread across organizations as SharePoint. Yet SharePoint is no final product, but rather an application platform to host DMS systems and customer specific solutions.

SharePoint and the Lifecycle of the Document

SharePoint applications manage the lifecycle of corporate documents from creation, to multiple revisions and publication, and eventual storage or disposal. SharePoint helps ensuring what kind of documents can be created, which templates are to be used, what metadata shall be included, in what structure documents shall be saved depending on their current phase in the lifecycle, how to control access, how documents are passed on to subsequent processes, what policies are to be applied to document related tasks, what events need to be recorded, which documents are to be preserved, protected or disposed.

www.pdf-tools.com

Dr. Hans Bärfuss is the founder and CEO of PDF Tools AG, an internationally successful software development and sales company. He represents the Swiss Standards Association at the ISO and actively helps with the standardization of file formats and digital signatures. He is also a founding member of the PDF/A Competence Center, an organization with the objective of promoting the ISO standards and PDF/A. By today, the organization counts more than 100 members, including major companies such as Adobe, Kofax, Nuance and T-Systems. Dr. Hans Bärfuss is the Chairman of the Swiss Chapter, speaker at numerous conferences and seminars and publishes articles on the subject of digital documents.

SharePoint and SharePoint solutions implement all these aspects of document management. Microsoft Office tools such as Word or Outlook support the document lifecycle with specific SharePoint functions in order for co-workers to benefit from the SharePoint advantages while using the tools they are accustomed to.

The importance of the document format itself can easily be overlooked in view of the commendable improvements in productivity, collaboration and seamless application integration SharePoint provides. Yet, does the intuitive reflex to obtain everything from one manufacturer solve all problems?

And the Document Format Does Matter!

In an effort to ensure interoperability and long-term access to file contents, the public sector and private industry have requested for some time that any used document format shall be standardized, openly documented and independent of the manufacturer to ensure interoperability and long-term legibility. Better interoperability and the elimination of proprietary file formats reduce the need for open operating systems and open application software. Cost analyses confirm that in the long term open file formats are better in preventing lock-in-effects than open software. So much for the theory – but which file format is most suitable when put into practice?

The PDF, PDF/A, ODF and OOXML formats claim that they are well documented and independent of the manufacturer. Are they all standardized? No doubt. But is each of those formats equally suitable for all phases in a document lifecycle? And are there good reasons to change the format during its lifecycle?

Put those questions to manufacturer of document management and archiving systems, one gets the answer: “We make sure that each document can be retrieved exactly the way it was stored.” But does that suffice? It certainly does not suffice if we need to store the document over a long period of time. The crucial question that should be asked is if we can properly present and read the document in the future – regardless of the hardware, operating system and software at that point.

To further explore this question, we split the lifecycle of a document into two phases, "Working Document" and "Final Document", and examine the requirements with respect to document exchange and long-term archiving. In the end we will come to the conclusion that ODF and OOXML are best suitable for the "Working Document" phase, while PDF or even better PDF/A is best suitable for the "Final Document". In the case of long-term archiving PDF/A in fact is a must.

To support this theory, let us look into the file formats in more detail:

PDF

PDF is known around the world and present in almost any market segment. Most of us use the term "PDF" in conjunction with email attachments or a document that can be downloaded from a Web portal. But as a matter of fact, we only have a rough idea of what exactly it is. The abbreviation PDF stands for "Portable Document Format" and specifies a file format. PDF was developed by Adobe Systems Inc. in the early nineties as a platform independent file format. Based on the experience with its predecessor PostScript Adobe set the following goals: Enable the exchange and presentation of electronic documents, graphically display text and images regardless of their resolution, optimize documents for web-viewing and offer interactive functions.

PDF as an electronic document format is attractive for many reasons. PDF is platform independent. A PDF document that is created on Windows can be further processed on a UNIX server and viewed on a Macintosh computer. The PDF format is based on the established PostScript page description language, and offers many additional functions such as direct access to pages, compression, encryption, interactive navigation, comments or forms. In addition, today PDF is the most frequently used format in the production of print setting in the digital pre-press. Private organizations, public authorities and educational institutions are re-engineering their business processes by replacing their paper-based workflows with electronic information exchange processes.

One of the main reasons for the proliferation of PDF is the PDF-Reader by Adobe. The PDF-Reader has been for free for a long time – costs apply only when creating or manipulating PDF documents. Since the first release of the PDF format, Adobe has encouraged other manufacturers to implement PDF with their solutions. The market responded to this signal with a significant number of independent suppliers offering PDF software and components.

The successful evolution of PDF has enormously strengthened the trust in this format. The only criticism applied to the format was that it was Adobe proprietary, which resulted in a demand for an internationally accepted standard. No surprise as such that PDF has evolved to the ISO-Standard (ISO 32000) for the electronic document exchange. The first part of the standard (PDF 1.7) was published in 2008; the second part (PDF 2.0) is scheduled for fall 2011. The standard also provides the basis for additional application specific ISO-standards, with the most prevalent ones being PDF/X for document exchange – particularly in the graphics industry, PDF/A for long-term archiving and PDF/VT for the high-volume printing of transactional documents with variable data.

PDF/A

The main initiators for the implementation of an archiving standard for electronic documents were AIIM (Association for Information and Image Management), NPES (National Printing Equipment Association) and the Administrative Office of the United States Courts (AO). Their goal was the definition for a standardized format for electronically archived documents. The result of this initiative was the ISO-Norm 19005, which defines a file format based on PDF called PDF/A.

This format offers a mechanism to present electronic documents in such way that the visual appearance will remain intact over a long period of time – irrespective of tools and systems used for their creation, storage or retrieval. The standard does not define the method, meaning or purpose of archiving. Instead it defines a norm for electronic documents, which shall guarantee that a document can be reproduced authentically in the future.

The PDF format itself does neither guarantee the long-term reproducibility, nor does the independence from software and output devices. To support these two principals, the existing PDF standard had to be restricted and extended at the same time. Hence a document may not directly or indirectly point to an external source. An example for that is an external picture. Certain functions of PDF such as support for sound and video for instance are also not permitted. On the other hand, other options such as the embedding of fonts into the document are mandatory.

In essence, the PDF/A standard defines and clarifies selected properties of the PDF reference 1.4 – stipulating if they are absolutely necessary, recommended, restricted or prohibited. The PDF/A standard (ISO 19005) is serialized, where the first part (PDF 1.4) was introduced in 2005 and the second part (PDF 1.7), based on ISO 32000, will be published in spring 2011.

ODF

The OpenDocument Format (ODF) is an open XML-based document format for the office applications text editing, spreadsheet processing and presentations. The format relies on other open standards wherever possible, such as formats of multimedia content or fonts. Initially, the ODF format was developed by Sun as a file format for the OpenOffice applications. A technical committee at OASIS further developed the format and published it in 2005 as the OpenDocument format (ODF). Version 1.0 and 1.1 of OpenDocument have since been certified as ISO 26300 and version 1.2 has been in draft since 2009.

OOXML

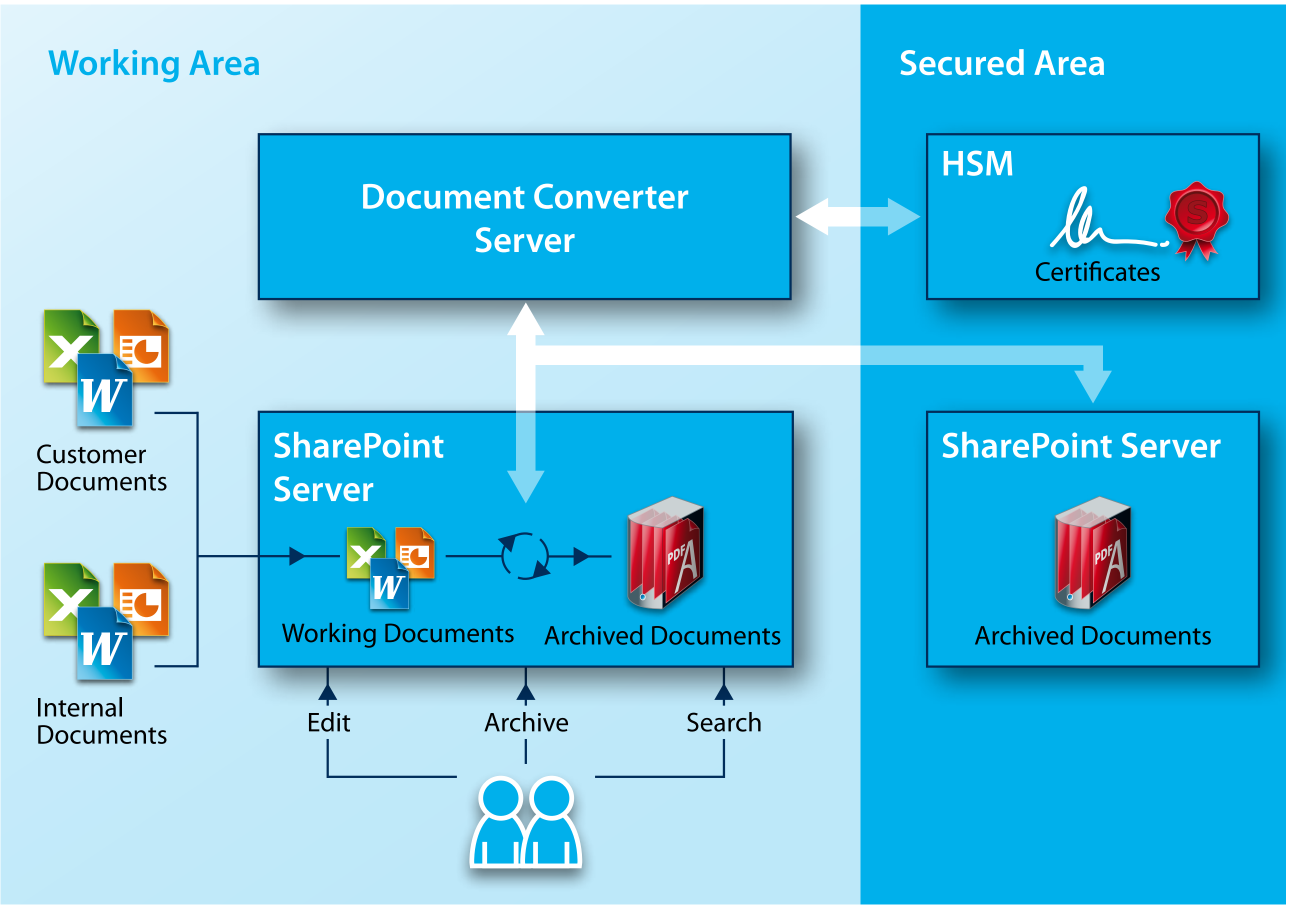
The term OOXML is an abbreviation of Open Office XML and it too is an ISO standard. Open Office XML was developed by Microsoft and submitted for standardization to a working group at Ecma International, where it was published in 2006 as Ecma-376. In 2008 the standard was published as ISO 29500. The entire standardization process was complicated, took setbacks and was covered in reports of irregularities. As the name suggests, the format was intended to make proprietary file formats of MS Office products such as Word, Excel, PowerPoint available in a standardized and publicly accessible format. The widely used XML syntax was used for its basis. The standardization process was to reflect the existing capabilities of the proprietary documents in XML, to expand individual capabilities, to document and to enable interoperability between applications.

At the emergence of OOXML there were already over 400 million users who produced an estimated 40 billion documents annually using the above mentioned tools. As a consequence OOXML not only had to reflect an enormous amount of existing documents, but it also had to support a wide range of new applications. Examples of such applications are the automatic creation of documents from business process data, the extraction of data from documents and the re-insertion of that data into business applications, the targeted and automated processing of documents and many more.

Long-Term Archiving and Legal Requirements

The probably most critical challenge of the digital age lies in the long-term archiving. We are accustomed to producing an exponentially increasing amount of documents. To this day, without much consideration, we still link the digital output to the tools we use to create it. Our experience however tells us that it can be exceptionally difficult to visually reproduce such documents after 10 or 20 years, and even if so – with significant loss of content information. Preserving the financial and intellectual investments in these documents is becoming an urgent priority.

The power of the file formats poses significant problems to archivists: Functions such as encryption, dynamically changing content and dependencies of external resources such as fonts are not manageable in the long term. A solution had to be found fast; and for the lack of alternatives – archive documents were simply converted to TIFF. Nothing is wrong with the TIFF format in general, but as the term "Tagged Image File Format" ever so aptly describes – it is a file format for raster images that can be extended readily using mostly proprietary functions. TIFF does not offer standardized functions such as searchability, embedding of



metadata or digital signatures. The lessons learnt with TIFF however have contributed to the development of the PDF/A format, which was developed as an initiative by AIIM (Association for Information and Image Management), NPES (National Printing Equipment Association) and the Administrative Office of the United States Courts (AO) to specifically address archiving requirements.

PDF/A satisfies all important requirements of archivists such as static content, predictable and true visual reproducibility of documents regardless of the platform and software; as well as no references to external source, free of encryption and patent rights, searchability, embedding of metadata and digital signatures and more. The vision of "digital paper" finally came true with PDF/A. It is the same properties of PDF/A and the hard work of many volunteers, including the initiation of the PDF/A Competence Center, which have made PDF/A to the de-facto-standard for archiving. The proof can be found with the many recommendations, guidelines and legal requirements of public authorities and governments in many countries around the globe.

The Swiss government has committed to the PDF/A format in its legislation for the exchange of digital documents. The same trend can be observed in other countries such as France, Austria, Norway, Denmark, etc. where PDF/A is the standard for the public sector. National and state archives also prefer to receive documents in the PDF/A format. A clear sign that the PDF/A format has prevailed are the many projects in the private sector, which are not driven by regulatory requirements, but rather by sustainable economics.

Interoperability and Conversion

Document standards claim that they support "interoperability". What does that mean exactly? Only complicating the matter is the fact that very different format properties are meant. For instance, the term interoperability can mean that various application programs present the document the same way, or it can mean that applications interpret the document structure and content identically. The quality of interoperability becomes evident once a user wants to convert the document from one format to another. Why is that?

Most document formats separate the layout from the document structure and content, such as PDF, ODF and OOXML do. Depending on the application, one or the other aspect is more important. With office documents for example the structure and content are of importance – while for printing the layout is more important. The phase of the document lifecycle also plays a role. During the early phases of changes in the document there is a demand for content and structure – whereas in the final phases the demand shifts to the layout. The document formats PDF, ODF and OOXML are quite different with respect to the mentioned aspects. The strength of ODF and OOXML is found in the structure and content, whereas with PDF the strength is conversely in the layout. Not that the formats would categorically ignore the opposite aspects – but they support them rather reluctantly. This comes to no surprise if we consider that PDF finds its roots in the graphics industry, as opposed to ODF and OOXML, which find their origin in office applications.

These insights confirm the theory that PDF – and in particular PDF/A – is best suited for the final stages of documents. On the other hand it is unlikely that one would want to use PDF as the format for editing documents. In this phase of the lifecycle it is clear that ODF and OOXML are better suited. The consequence of this is obvious: Documents managed in SharePoint must be converted to PDF/A if they are to be archived. The preparation for archiving as well as for document exchange requires a perfectly layout-true conversion of ODF or OOXML to PDF/A.

The SharePoint platform is well prepared for this task. As extensible platform it can be complemented with a Document Converter Service, sometimes also referred to as a rendition service, which can convert ODF or OOXML documents automatically or user-initiated to PDF/A and flag the converted documents with a "ready-to-archive" attribute – typically in the form of an electronic sign-off signature. Additional automated background processes control the storage of documents to the archive and ensure that documents are seamlessly integrated into SharePoint's text search capabilities.

To guarantee an impeccable image of the document layout, the Document Converter Service requires making use of the native office applications. For ODF this is OpenOffice and for OOXML the corresponding Microsoft Office applications. A study by the Fraunhofer FOKUS Institute has shown that the interoperability between ODF and OOXML can in many cases be very difficult if not impossible. Ambiguities in the description of the standards result in many cases in unpredictable representations of the layout.

The Conversion to PDF/A is Necessary

The SharePoint platform has established itself quite successfully in many organizations, where it improves productivity and the collaboration of "information workers". SharePoint applications manage the document lifecycle in organizations from creation to storage. The document format plays an important role in this, because documents can surpass the lifetime of the creation, manipulation and archiving systems used. As such, the format must be open, documented, not proprietary and standardized. An open format is more economical in the long run than open platforms and applications (Open Source). This may be one of the reasons for the high level of acceptance of SharePoint and Microsoft Office.

But not every format is perfect for each phase during a document lifecycle: ODF and OOXML are recommended for the phase of the "Working Document" whereas PDF/A is a "must" for the "Final Document". Consequently, the document must be converted from ODF or OOXML to PDF/A at the time it transitions from one phase to the other. For this task, there are professional Document Converter Service applications, which are integrated on top of SharePoint. With the help of the SharePoint platform, these applications automate the conversion process and they ensure that text searches of PDF/A documents will be transparent to the user.

Many organizations have already implemented Document Converter Service solutions, such as the largest life insurance company in Switzerland with over 700 insurance consultants in over 40 agencies. The scenario is this: Microsoft

Office documents from business processes – company-internal as well as customer documents – are managed and archived in SharePoint Server. In the past, TIFF was used as file format for archiving. The company switched to PDF/A in order to make the documents searchable, to apply digital signatures and to guarantee traceability. At the completion of the business case, the documents are converted to the PDF/A format and metadata is added prior to storing them in the archive. The users control the conversion of the documents directly from within SharePoint where they can access the PDF/A documents from the standard user interface. At the same time, documents are replicated into an autonomous and robust long-term archive – also based on SharePoint – which protects the documents from unauthorized access. A digital signature is applied to ensure the authenticity of the documents and to prevent these documents from subsequent changes.

PDF Tools AG
Kasernenstrasse 1
8184 Bachenbülach
Switzerland
Tel: +41 43 411 44 50
Fax: +41 43 411 44 55
pdfsales@pdf-tools.com
www.pdf-tools.com