4th International PDF/A Conference

PDF/A *Forever*

Long-Term Archiving with PDF

- What you need to know about PDF/A and PDF/A-2
- Legal Certainty
- Accessibility and Metadata
- Archives & Libraries
- Public Administration
- Business to Consumer

PDF/A Competence Center

aiim
Find, Control, and Optimize Your Information

4th International PDF/A Conference

PDF/A *Forever*

Long-Term Archiving with PDF

ISBN: 978-3-9813077-3-3

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie.
Detailed bibliographic data is available at <http://dnb.ddb.de>.

Published by Association for Digital Document Standards ADDS – PDF/A Competence Center, Berlin – www.pdfa.org

Printed in Germany

Layout, design, composition and cover design: Alexandra Oettler; cover photo: Andrea Cabibbo

Printing: Galrev Druck- und Verlagsgesellschaft Hesse & Partner OHG

Preface

Rome – Can you think of a better place for a conference on long-term archiving than the Eternal City? We are pleased to conduct our 4th International PDF/A Conference in Rome and hope we have organised an interesting program and event in cooperation with our Italian partner, 4IT Group.

"All roads lead to Rome", and we have come a long way getting here. The PDF/A standard was published by the ISO five years ago in the fall of 2005. Since then, users in different countries and various industries have recognised the benefits of PDF/A as it definitely helps them meet requirements in the digital world for safe long-term archiving and guaranteed access to documents.

The PDF/A Competence Center was founded in the fall of 2006 and, due to the large interest, has over 100 members today. The information about PDF/A that we make available, with technical explanations and application scenarios, has helped to turn a "formal ISO paper standard" into a well understood and accepted document format in the real world. The first phase was to provide information about PDF/A in German-speaking countries, and then the scope was extended internationally and to sector-specific topics. Today there are numerous active country chapters which provide additional information on a local level, e.g. in Italy, Germany, France, BeNeLux, the US and the UK.

Experience gained in deploying PDF/A solutions has produced benefits in many different sectors, including banking, insurance and engineering. Governmental organizations, archives and libraries now recommend the use of PDF/A, or even require it through legislation. "Mission Accomplished", you might think?

In Europe, PDF/A has achieved widespread usage and the PDF/A Competence Center has contributed extensively in bringing it forward. While "invented" in the US, the adoption of PDF/A there is much slower than in Europe, and Asia is still in the preliminary phases. There are still many things to learn about applying PDF/A in various industries.

The formal name of the PDF/A Competence Center is the "Association for Digital Document Standards" (ADDS). This name was chosen intentionally; we always planned to extend our scope when appropriate. Rome will likely be the starting point.

For this conference, our focus remains on PDF/A, and there have been many developments! The next part of PDF/A, PDF/A-2, was technically completed in the summer of 2010. The formal ISO process leading to publication will still take several months. Approximately five years after the first standard was released, the new part incorporating technological enhancements and reflecting practical experience will be published. Therefore, PDF/A-2 is a central topic of this PDF/A Conference!

We would especially like to thank our partner, 4IT Group, who have done a tremendous job in planning, promoting and organising the 4th International PDF/A Conference. Furthermore, we would like to say thank-you to the authors of these interesting articles, and to the organizers and helpers who have played a part in arranging this compendium.

When the 4th International PDF/A Conference is history, this Conference Proceedings will also remain forever – as a PDF/A file!

Rome, September 2010
Harald Grumser
PDF/A Competence Center
Chairman of the Board

Table of Contents

Forewords

Betsy A. Fanning, AIIM 8
Massimo Maronati, ANORC 8
Hanns Köhler-Krüner, cdm Europe 9
Michael Fray, DIMO 9

Keynotes

PDF/A Making Inroads in the USA 10
Stephen Levenson,
US District Courts and Convener of the ISO PDF/A Committee

PDF/A-2 – Technical Overview 12
Olaf Drümmer,
callas software, Managing Director; Board, PDF/A Competence Center

Receiving Deeds in PDF/A Streamlines Public Administration 17
Michele Piva,
Infocamere

Italian IT network CGN enables 20,000 Professionals to Submit PDF/A-Deeds 19
Andrea Cappellato, CGN
Donato Stellabotte, Technosolutions

Track 1: What You Need to Know About PDF/A and PDF/A-2

Session Intro – Track 1: What You Need to Know About PDF/A and PDF/A-2 21
Johannes Hesel,
Board, PDF/A Competence Center

PDF/A 101 – Introduction to PDF/A 22
Raffaele Bernardinello,
ICT Director, C.M. Trading S.r.l.

PDF/A from Document Management to Graphic Arts 24
Alessandro Beltrami,
Consultant, TechnoSolutions srl

Scan to PDF/A with OCR – Paper Becomes Digital 26
Carsten Heiermann,
President, LuraTech Europe GmbH

Archiving Digital Documents – Conversion of Digital-born Documents to PDF/A 31
Dr. Hans Bärfuss,
CEO of PDF Tools AG and Vice-Chairman of the PDF/A Competence Center

Track 2: Legal Certainty

Session Intro – Track 2: Legal Certainty 39
Dr. Bernd Wild,
Board, PDF/A Competence Center

Electronic Invoices as PDF/A With Respect to Italian Legal Requirements 40
Domenico Barile,
E-Mission S.r.l

Reproducibility of Archived Documents 45
François Fernandes,
levigo solutions GmbH

Archiving E-mails with PDF/A 48
Dr. Bernd Wild,
intarsys consulting GmbH, Board, PDF/A Competence Center

Track 3: Metadata and PDF/A

Session Intro – Track 3: Accessibility and Metadata 52
Olaf Drümmer,
Board, PDF/A Competence Center

Accessibility in PDF and Elsewhere 53
David Hook,
Director Product Management, Crawford Technologies

Accessibility – What PDF/A-1a Really Means 57
Duff Johnson,
CEO Appligent Document Solutions

XMP Metadata Primer 61
Olaf Drümmer,
callas software, Managing Director; Board, PDF/A Competence Center

The Knowledge Map: How to Take Advantage of XMP Metadata 63
Manuel Brunner,
Head of Projects and Services, Intrafind AG

Track A: Archives and Libraries

Session Intro – Track A: Archives and Libraries 66
Thomas Zellmann,
PDF/A Competence Center, Managing Director

Relevance of PDF/A in Archives & Libraries/Digital Preservation 67
Hans-Joachim Hübner,
Satz-Rechen-Zentrum Hartmann+ Heenemann GmbH & Co. KG

Website Archiving to PDF/A – Customer Story: UBS 70
Rolf Günter,
Head of Business Development, Sales and Marketing, PDF Tools AG

nestor – the German Network of Expertise for Digital Preservation 72
Natascha Schumann,
nestor – Kompetenznetzwerk Langzeitarchivierung

Track B: Public Administration

Session Intro – Track B: Public Administration 74
Bernd Wild,
Board, PDF/A Competence Center

PDF/A at the Road and Transport Authority of Dubai 75
Sanat Kulkarni,
eDocuMAN Fz LLC

PDF/A and Records Management in the Netherlands 76
Dominique Hermans,
Owner of DO Consultancy

Track C: Business to Consumer

Session Intro – Track C: Business to Consumer 78
Harald Grumser,
PDF/A Competence Center, Chairman

SAP and PDF/A – PDF/A in Product Life Cycle 79
Dr. Uwe Wächter,
Product Manager PDF Technologies, SEAL Systems AG

Optimising PDF/A Documents for Large Archives 83
Harald Grumser,
CEO, Compart AG and Chairman of the PDF/A Competence Center

Promotional Contributions of our Sponsors and Exhibitors

Exhibitor: 4IT Group, Italy 87

callas software: A Core Provider of PDF/A Technology and Solutions 90

Content and Document Management Europe Limited (cdm Europe) 94

Crawford Technologies – High-Volume Transactional Output Software 95

Exhibitor: Digital Planet, Turkey 98

intarsys – Electronic Signature and PDF/A 99

IntraFind AG – Specialist for Information Retrieval 102

LuraTech – Document Conversion Software 104

SEAL Systems – The Digital Paper Factory 107

About Us

About the Authors 110

About the PDF/A Competence Center 116

It is encouraging to see the continued adoption of the PDF/A standard, ISO 19005-1: 2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1). In the United States, we have held a couple of very successful workshops to promote the standard. These workshops were attended by representatives from all levels of government, healthcare industry, manufacturing, utilities, universities and archives. Organizations worldwide are struggling with how to preserve their electronic documents. ISO 19005-1 and soon to be released, ISO 19005-2 are the answer for these organizations.

As you may remember, we are developing this standard in parts so that organizations who adopt the use of PDF/A for long-term preservation of their electronic documents will be assured that their files will remain valid conforming PDF/A files even as new parts or revisions to the standard are introduced. The first part of ISO 19005 was developed specifically for static documents or e-paper. With part two of the standard, we are continuing to address static documents while updating the standard to comply with ISO 32000-1, the PDF standard, and enhancing a few of the features of a PDF/A file. In response to the many requests for the ability to embed files in a PDF/A file and still maintain the PDF/A conformance, the ISO working group began work earlier this year on a version of PDF/A which will allow for embedded files. We are at the early stages of this embedded files project with a new work item ballot issued to the ISO committee member countries.

As we look to the future, AIIM in cooperation with the PDF/A Competence Center will continue efforts to promote PDF/A bringing a heightened awareness of the standard to organizations worldwide. We look forward to continuing to develop and evolve the PDF/Archive standard to meet the growing needs of the organizations that have adopted it and look forward to seeing more products announced that comply with the standard.

Betsy A. Fanning, AIIM

Italian legislation introduced digital archiving of legal documents in the middle of the nineties, ahead of national and international experiences and best practice definition. This created a long season during which software houses and final users spent a lot of time and energy in experiments and research. As well as this, at that time stable and widespread formats and standards were not yet available.

Since then Italian law makers, with several new measures that slightly confused people in the market, have tried to define more clearly what the requirements are for proper digital archiving and the related process needed to "stabilize" documents' content (Digital Signature and Time Stamp) to make them compliant with long term archiving. ANORC was founded in 2007 with the aim to facilitate the development of the digital archiving market through ongoing dialogue between final users, providers and legislators.

Ever since PDF/A appeared ANORC has promoted and supported its adoption for short and long term digital archiving. At the moment PDF/A is being examined by legislators and it will be introduced as a standard reference, in the new up coming legislation.

Over the last year some important long term digital archiving projects have been activated with ANORC members' cooperation in health service organizations as well as in Public Administration requiring both an "time unlimited" availability of archived documents. In all those projects PDF/A has been strongly recommended and often adopted. A dedicated ANORC team, involving major national filing clerk experts from several Italian universities, is finalizing a "technical rule proposal" (DPCM – Albo on-line) which aims to define the best practice for archiving any measure approved by Italian Public Administration. It will indicate PDF/A as a recommended standard to use in such archiving processes.

Other important organization (i.e. DigitPA) are investigating the way to introduce international standard like PDF/A as mandatory requirement digital archiving assuring a transnational document availability.

Massimo Maronati, ANORC

More and more standards are spanning the world as vendors, consultants and end-users understand and cherish their value. Since its inception in 2005, PDF/A has established itself on the forefront of the drive towards a standard data preservation file format, and is a casebook on how a standard should be supported.

cdm Europe (Content and Document Management Europe) is a relatively new organization which has set a goal of bringing together information management associations, universities and other interested parties to allow cross-border cooperation, exchange of knowledge and communal activities in the field of information management. We were very pleased to welcome the PDF/A Competence Center as our first official member after our founding. The PDF/A Competence Center has already shown the way forwards, demonstrating how international cooperation, exchange of information and working towards a common goal are necessary to achieve one's own goals effectively.

Going forwards, and with the aids of our members, cdm Europe wants to focus our activities in a pan-European way. In addition to PDF/A-2, we have great hopes for activities like CMIS and MoReq2010, as well as the new challenges of the continued information explosion as represented by Social Media and Web 2.0/Enterprise 2.0. An open exchange of experiences, information and best practise as well as the active support of cross-border cooperation is necessary. We are very proud to be present at the 4th International PDF/A Conference in Rome and look forward to the open discussions about PDF/A-1 and -2, as well as other standards dealing with best practise.

Hanns Köhler-Krüner, cdm Europe

Just recently I heard that the sun will cease to exist as we know it today in about 4.5 billion year (give and take a few years). Therefore, some scientists are working on a plan to move us all to Mars. But first they need to create an atmosphere on Mars, so we don't freeze to death. That should be accomplished by smashing an enormous asteroid into Mars, causing the temperature to rise 3 degrees Celsius. The sudden rise in temperature would melt about a trillion tons of water, which is enough water to form a lake covering an area three times the size of Rome.

The downside of this project is that it will still take many thousands of years afterwards before the idea of terraforming Mars can be fully realized. When I heard all this, I found it to be very interesting. But the first thing that came to my mind was this question: "How do they store all their documentation for all these years?" So I forgot about the fact that life will end on Earth, and started thinking about what kind of plan the scientists might have in place to ensure that their documents can be read in thousands of years. "None" would be my guess – but it is only a guess.

Many of us face the same challenges as these "mad scientists". Most of us don't have to worry about saving our private and corporate documents for thousand or millions of years, but one day we will have to have a plan.

Think about how much information and knowledge from the past we still benefit from today. Documents, scrolls, cave paintings, stone carvings etc. have been valuable for us to understand and learn from our ancestors. These are all media that easily can beat the duration of today's media. What will happen in the future, if the knowledge of today disappears? Nobody these days is sitting there, chopping 0's and 1's into media that are expected to last for thousands of years. And that is exactly why we all, corporate and private, need to have a plan to ensure our legacy and the passing on of knowledge.

Nowadays where so many documents, pictures, letters (e-mails) etc. are digital, how do we ensure that we also leave something for our successors? Think of the pictures from your last birthday. How many of them do you have on paper and how many on your hard drive?

Public and private companies also face these challenges. Not so much for sentimental reasons, but more for compliance reasons. Nobody wants paper any more. In DIMO we talk a lot about these issues, and we agree that one of the most serious plans for keeping documents for a long time is to use a format that is standardised, possible to validate and intended for long-term archiving. PDF/A seems like a very good fit.

Michael Fray, DIMO

PDF/A Making Inroads in the USA

Stephen Levenson,

US District Courts and Convener of the ISO PDF/A Committee

Can a four year old standard have a past that anyone would care about? The answer is yes, if the past begins with the adoption of the personal computer and the proliferation of desktop document creation. The first personal computer (PC) was invented by Steve Jobs and Steve Wozniak in 1976 and revolutionized the way we create and store business documentation. During the 1980's, Dr. John Warnock was developing Postscript as a printer language and Dr. An Wang would eventually convert his standalone word processor to a product called MultiMate. MultiMate joined other word processing products for the PC from Microsoft, Corel and others. Word processing would become one of the dominating applications for the PC. This is where future began and the trouble started. The concept of feature rich became problematic for preservation. A tension exists for features that help workflow, but need to be turned off for preservation.

No one ever was really worried about long term preservation because there was always microfilm or paper. The general consensus was that as long as there was the ability to magnify a photographic micrographic image, you had preservation. This seemed to work for homogeneous collections that were easy to organize like land records, library card catalogs, and others. You could gather these up under a camera and film them into a complete set.

Paper became the leading choice of indecision. Many collections were processed in paper and then stayed in that form. As long as business documents were not needed in multiple locations these systems worked, but still paper has many problems. In the United States many paper collections have had mold contamination. At a minimum this requires an expensive cleaning processes if the records are still needed for business purposes. Storms and floods have taken a large toll on primary record collections. Storms like Hurricane Katrina were particularly devastating because they covered a large geographic area and mitigation was unavailable or too late to be effective.

Business processes began to require more functionality. Filing from anywhere (Internet), remote access, and remote processing began to be routine business demands. Thus began scanning and direct acceptance of digital files into the business process.

Digitization was becoming the norm and not the exception. Not many years ago the United States National Archives only recognized ASCII and EBCDIC for electronic formats. After all, what else was there? But the Personal Computer and the proliferation of desktop applications for document creation changed everything.

The cacophony of multiple document creation tools was noise and not instrumentation and a maestro was needed to tune and tame the document creation tools. They did not lend themselves into "easy to organize" sets and some other technology was needed to meet the new challenges. The ability to have consistent reliable formats that you could reliably render far into the future (like microfilm) did not exist. That role was filled by PDF.

The good thing about PDF was that it maintained high product quality through market domination by Adobe Systems, Inc. This was true until the proliferation of many PDF writers started to enter the market at the beginning of the 21st century. There then began the need for more than a ".pdf" at the end of a file format to assure quality and reliability.

A business need arose for a testable independent version of PDF that honoured the needs of the archival community, but could still be a practicable business document format. Could we emulate some of microfilm's properties (e.g. self contained, verifiable, designed for long-term)?

What else is important for long term preservation?

We cannot assume what hardware platforms will be used to open files in the future. In the desktop arena, UNIX, Windows and Mac still fight for market share. In fact, desktops will not be the only choice for opening and

viewing documents. One only has to look at the proliferation of handheld devices from RIM and Apple today to see that it is not very easy to predict where that market may go. It is clear that file formats must be device independent. Also, dependence on external files makes preservation more difficult. It will be an additional burden to assume that the digital object and some external object that is needed to interpret the object will be available twenty years from now. It must stand alone as a self contained object. As an example, if you use a font set that is no longer available on the platform that you are rendering the object on, then a substitute font would be required. This violates the principal of accurate rendering. You are going to get unpredictable results. This will especially be true with pagination and special characters. The future of digital preservation is more self containment and not less. PDF/A is a good start in this direction.

Digital objects must be able to provide documentation for provenance and repurposing. Documents may be collected at the end of a process and verification of that process is essential evidence adding value to the document. PDF/A takes full advantage of XMP to be able to store extensible metadata. In this way the document is able to inherit as much metadata as is necessary to understand fully how it fits into a collection and any other documentation an archivist or record manager might deem important. A minimal portion of the file should be able to be inspected by basic tools. Basic tools should render these parts not only to human readable content, but to machine sortable organization for inspection and audit.

No discussion of this type can occur without discussing XML. Many advocates contend that you convert everything to XML and there you have it you are done. This ignores some critical issues, the biggest of which is that forms and documents are more easily understood by end users. XML looks like computer code. Some end users are ready to work with this, but not many. PDF provides the type of human readable appearance that is expected in business processes. Though, when moving structured database data between processes, XML shines and there is no better format. There is a little known story: PDF and XML are not mutually exclusive. For example, the Brazilian government will store a XML source code in the PDF document private data area. This way, both the PDF format is used and the XML is available for reuse. It is the best of both worlds. A new version of PDF/A will consider incorporating XML. So when it comes to XML and PDF it is not a question of either or but use based on the business case.

Where are we now with PDF/A?

Only good news has happened to date. The ISO committee has published ISO 19005-2. You may ask why does ISO 19005-1 need to be updated if it was to be forever? Well, it will be forever, all documents written to this standard will be able to be read in all future updates. But technology moves on and PDF itself has had some changes. These should probably more correctly be called additions or new features. PDF/A picked PDF up at version 1.4 and ISO 19005-1 is based on that. Since that publication, PDF moved to version 1.7. Adobe then rewrote it and gave to the world and it was renamed ISO 32000-1 after worldwide ISO adoption.

ISO 19005-1 (PDF/A) is based upon the Adobe specification for PDF 1.4. When we update to ISO 19005-2 it will be based on ISO 32000-1 or in plain language it will be a standard based on a standard.

Tremendous excitement has been generated by the PDF/A Competence Center and their activities over the last year. The PDF/A Competence Center is advancing the standard. In addition many educational sessions have been held with many more anticipated. Education has started in the United States and sessions have already occurred in Chicago and Washington D.C. See www.pdfa.org for more information.

Where is the future?

PDF/A 19005-2 will have been voted upon by the time of the 2010 conference. Work has begun on 19005-3. Both standards will be better because a much larger community has made it a better standard. This trend seems to be accelerating with groups like the PDF/A Competence Center growing in membership and mission. PDF/A 19005-3 will be limited to additional file types that may be included in the format. For example XML is under consideration to be an acceptable embedded file type.

At the first international conference on PDF/A I listed in my keynote an acknowledgment to a list of "Hall of Fame" participants that made a critical difference in the emergence and development of this standard. The PDF/A Competence Center has earned its honorable listing.

19005-2 and the coming 19005-3 would not be the quality addition without the wisdom of the PDF/A Competence Center. The standard has benefited from the participation on the ISO committee, AIIM, NPES and the PDF/A Competence Center. We look forward to adding you and your wisdom soon.

PDF/A-2 – Technical Overview

Olaf Drümmer,

callas software, Managing Director; Board, PDF/A Competence Center

The Portable Document Format – developed by Adobe Systems and first published in 1993 – offers a series of beneficial and hitherto unavailable characteristics. In addition to platform independence, this includes reliable rendering of page oriented contents. All components of a file such as fonts, texts, images and graphics are included in the PDF file itself. These advantages made PDF a candidate for standardization for selected usage areas and industries. What all standards have in common is that, as subsets, they restrict the extensive features of PDF and, in this way, optimise it for special application options.

PDF/X and PDF/VT: Standards for the Printing Industry

In the printing industry, it is extremely important that digital print material is reliably reproduced. The technical ISO committee TC 130 developed the PDF/X standard, whose first part was published in 2001 as the first international PDF standard. Since then additional parts have been developed as part of the ISO 15930 series of PDF/X standards. The most recent additions – PDF/X-4 and PDF/X-5 – have just been revised as of July 2010. The PDF/VT standard ("V" for variable data printing and "T" for transactional printing) adds a purely PDF based option to the widely used high volume printing formats like AFP and PPML. It enriches PDF for the purpose of variable data printing and has been published in August 2010.

PDF/A for Long-Term Archiving

PDF/A-1 was published by the ISO committee ISO TC 171 in autumn 2005 as the international standard ISO 19005-1 for long-term archiving. PDF/A is beneficial above all in the administration, archiving, library and publishing, as well as banking and insurance sectors, but also in manufacturing industries since any digital documents – from invoices, books and manuals to technical drawings – can be permanently stored using this standard. Work on the second standard part, PDF/A-2, was completed from a technical point of view in Paris in summer 2010. It is expected to be published early 2011. Just having PDF/A-2 completed, the PDF/A committee within ISO TC 171 started work on an additional PDF/A standard part, PDF/A-3.

PDF/UA: Standard for Accessible PDF

The accessibility of document content is becoming increasingly important, especially in the field of American and European authorities and administrative districts. The PDF/UA standard (UA stands for "Universal Accessibility") is being developed as ISO 14289-1 within ISO TC 171. It aims to ensure that contents of PDF files are properly structured and thus are sufficiently accessible such that assistive technologies like screen readers can extract and present their content accordingly.

PDF/E: The PDF for Technical Documents

ISO standard 24517-1 was published in 2008. This standard supports the reliable and secure exchange of engineering related content.

Last but not Least: PDF 1.7 is an ISO Standard

Since summer 2008, the PDF file format has also been standardized as ISO standard 32000-1. It is based on PDF 1.7. New developments are being incorporated into the second standard part – ISO 32000-2 – also known as "PDF 2.0", which is currently being developed and scheduled to be finalized in the second half of 2011.

Advantages of the PDF/A Standard

As a standard for long-term archiving, PDF/A guarantees the reliable reproduction of documents over many years, regardless of technological developments in hardware and software. PDF/A ensures a homogenous archive in which both digitally created and scanned documents can

be stored. As an international ISO standard, PDF/A is valid throughout the world. The environmental and climate aspect is also in PDF/A's favour. Paper documents can increasingly be replaced by PDF/A documents since PDF/A is able to satisfy all requirements regarding longevity and binding nature (signatures).

PDF/A-2: Central Innovations

Whereas PDF/A-1 is based on PDF 1.4, PDF/A-2 takes advantage of features that only became available in later versions of PDF, up to and including PDF 1.7. PDF/A-2 is no longer based on a specification published by Adobe, but on the internationally approved ISO standard 32000-1. The following overview shows the core innovations of PDF/A-2 which all users should be aware of.

JPEG2000 Image Compression

The powerful compression method JPEG2000 (ISO/IEC 15444) was not supported in PDF/A-1 as it only was introduced into the PDF specification when PDF 1.5 was released. The new JPEG2000 options are interesting for scanned documents, among other things, because higher compression rates can be achieved at a better quality than with the older JPEG format. Furthermore, JPEG2000 offers a highly efficient lossless compression level.

With JPEG2000, libraries and archives can digitize historic maps, books or documents, for example, at the best quality possible, and create size-optimised PDF/A-2 files with JPEG2000 from them. The relevant metadata for the object can be stored directly in the PDF/A-2 document.

Embedded PDF/A Files via Collections

Making use of collections – also called "portfolios" in Acrobat – users can combine several files into one "container PDF". PDF/A-2 can now be used to compile PDF/A collections from several PDF/A files. File formats other than PDF/A are intentionally not allowed inside PDF/A collections. One possible use of PDF/A collections is the archival of e-mails – e-mail attachments (such as Word files) can be converted to PDF/A and archived alongside, but as a separate entity, the e-mail body.

PDF/A collections can also be advantageous in the social security sector, as signatures can be applied to single scanned pages. The PDF/A collection then combines the signed single pages of the whole document. One or more pages can be subsequently removed without affecting the validity of the signatures for the remaining pages.

Transparency

Although transparency is part of PDF 1.4, at the time of the PDF/A-1 standardization it was not yet well enough supported and thus excluded for standard conforming PDF/A-1 files. In the meantime technology has substantially matured, and transparency has become a common characteristic of numerous PDF files.

Transparency is often used for a design element in the form of drop shadows, crossfades or vignettes. It may appear unintentionally in a PDF file, for example if the original file is a PowerPoint presentation, or where text is marked up with highlighting. The use of transparency is now permitted in PDF/A-2.

PDF "Optional Content" or Layers

PDF/A-2 supports optional content, also often referred to as PDF layers. Optional content is helpful for technical construction drawings and plans, among other things, as the contents can be shown or hidden according to topic, such as the electrics or water supply for a building. Layers can also be used to display multilingual contents – such as an international catalogue – in a single PDF file. With the layers function, users can switch between English, Japanese and German text, for example, while the graphics remain the same.

OpenType Fonts

The cross-platform OpenType fonts themselves are standardized as ISO/IEC 14496-22. These fonts provide extensive support for Unicode. OpenType fonts exist as TrueType (suffix ".ttf") and PostScript variants (suffix ".otf").

In PDF/A-2, these fonts can now be directly embedded without first having to convert them – as was necessary with PDF/A-1 – into the older formats PostScript Type 1 or TrueType.

New Conformance Level PDF/A-2u – "u" for Unicode

Conformance level "b" stands for "basic". PDF/A-1b and PDF/A-2b focus on visual integrity. PDF/A-1a and PDF/A-2a ("a" for "accessible") contain additional features. These PDF/A documents also include structural information (such as information about paragraphs, headings or columns) as well as semantic information through the use of Unicode and alternate text. The latter is important to ensure that Copy&Paste from PDF/A files works without problems, and to ensure correct indexing of text. New to PDF/A-2 is the conformance level "PDF/A-2u" ("u" for "Unicode"). As a slimmed-down version of level "a", it of-

fers the advantages of Unicode (text searching and copying text) without having to adhere to any complex structural requirements that may be required by the "a" conformance level. PDF/A-2u is feasible both for digitally-created PDF files and for scanned documents with subsequent text recognition.

Object Level XMP Metadata

In the metadata domain, PDF/A-2 specifies the requirements that are imposed on custom XMP metadata fields for content objects – this goes beyond PDF/A-1 insofar as there only the document level metadata were subject to these provisions. User-de-

PDF/A-2 Innovations for Developers

The following new or improved technical functions of PDF/A-2 should be interesting for developers and programmers in particular.

Feature	New in PDF/A-2
Conformance level "A"	Extended requirements for conformance level "A"
File header	Only file headers from *%PDF-1.0 … %PDF-1.7* are allowed
Structure and tags	Mapping of user-defined tags and standard tags in a role map
Compressed object streams	PDF/A-2 supports compressed object streams, which were introduced with PDF version 1.5
Revision of the restrictions for the implementation	Among other things, the limit to 8191 array objects was lifted
Linearized PDF	No longer regulated by PDF/A-2
Appearance of comments	Annotation appearance is no longer required if the area is empty, or if the annotation is a link or popup annotation
History entry in XMP	If the history entry exists, certain rules apply
ICC profiles	Latest version ICC v4 is supported
Default CMYK	Improved provision
Prepress: overprinting, CMYK	Provisions for the use of overprinting mode and ICC-based CMYK
Spot colours	Spot colours must be consistent with regard to the alternative colour space
Name objects in valid UTF-8	Certain name objects such as for spot colours or structure types have to be encoded as UTF-8
Subset fonts	Provisions for subset of fonts were revised with regard to CharSet and CIDSet
TrueType	Encoding requirements for TrueType were revised (differences array; Adobe Glyph List)
.notdef glyphs	The use of .notdef glyphs (placeholder for glyphs in a font that are needed for rendering but are not present in the font) is no longer allowed
Namespace prefixes	Fewer stipulations regarding prefixes of namespaces
Document requirements key	Not allowed
XFA	The XML-based standard XFA (XML Forms Architecture) is now partially allowed

fined fields also on the level of content objects must now be defined using an extension schema if they are to be PDF/A-compliant.

New Comment Types and Annotations

PDF/A-2 establishes revised provisions around comments in PDF files. Some new annotation types – like 3D annotations – were added to the list of prohibited annotation types, while other new annotations introduced after PDF 1.4 (like text editing comments) are now allowed.

Digital Signatures

From the beginning, PDF/A intentionally allowed the use of electronic signatures. In PDF/A-2 more specific guidance has been added with regard to how digital signatures should be applied to guarantee interoperability. PDF/A-2 carries over provisions from the ETSI/PadES standard. PAdES (PDF Advanced Electronic Signatures) is a set of restrictions and enhancements to the PDF standard in accordance with ISO 32000-1 in order to improve the integration and use of advanced electronic signatures. ETSI has standardized the PAdES standard under TS 102 778.

Switching to PDF/A-2: Considerations and Strategies

The most important fact first of all: PDF/A-2 will not replace or supersede PDF/A-1 in any way. PDF/A-1 compliant documents that were already created will remain valid PDF files for long-term archiving. Archived PDF/A-1 files can remain unchanged in the data archive; an "update" to PDF/A-2 is not necessary here and does not usually make sense, since a PDF/A-1 document is always also a valid PDF/A-2 document.

Creating an Individual Requirements Profile

When reviewing the new PDF/A-2 features, users who discover functions already on their personal wish-list are more likely to benefit from upgrading than those who are completely satisfied with the features of PDF/A-1. A functional, successful archive system relying on PDF/A-1 can remain to be based on PDF/A-1. Once the PDF/A-2 standard is published early 2011, in the initial phase, only a few tools will be available that have already implemented the full extent of the new ISO standard. This may impact the individual schedule for organisations intending to move to PDF/A-2.

Anyone dealing with the topic of PDF/A for the first time has to choose whether or not to use PDF/A-2 im-

Also New in PDF/A-2

Below is an overview in table form of further changes and improvements to PDF/A-2, mostly related to specific technical details.

Feature	New in PDF/A-2
Links	Now also possible in the form of multi-rectangle link annotations (link in the form of several related rectangles at a line break)
Links with PDF collections	Links can be set to, from, or between embedded PDF/A files
Freeform comments	Freeform annotations, such as polygons, are allowed
User unit	Page sizes on a 1:1 scale of up to 381 km feed size (previously 5.08 m) are possible
Units of measure	Support of measurement properties; important for technical documents
Structured PDF	Extended options for tagged PDF
Encryption	Extended options for encryption
Electronic signatures	Extended options
Colours: DeviceN	Maximum number of colourations/colourants in DeviceN
Colours: NChannel	NChannel supported

mediately. Here, too, the availability of the relevant software tools must be verified. In principle, there is nothing objectionable regarding the use of PDF/A-1 for long-term archiving now or in the future, since software is and will remain available, and the know-how that beginners acquire with PDF/A-1 can also be used to a large extent with PDF/A-2.

Even with ongoing projects, if PDF/A-1 satisfies all requirements, then the workflow should remain unchanged. If, however, PDF/A-2 offers crucial features that cannot be implemented based on PDF/A-1, the upgrade should be started at an appropriate point in time.

Literature

PDF/A Standard:
ISO 19005-1:2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1), www.iso.org. (2005).

ISO/DIS 19005-2, Document management – Electronic document file format for long-term preservation – Part 2: Use of ISO 32000-1 (PDF/A-2), www.iso.org. (2010).

PDF/A Competence Center, www.pdfa.org.

PDF Standard:
ISO 32000-1, Document management – Portable document format – PDF 1.7, www.iso.org. (2008).

PDF Reference:
PDF Reference, Third Edition, Adobe Portable Document Format Version 1.4, www.adobe.com. (2001).

PDF Reference, Sixth Edition, Adobe Portable Document Format Version 1.7, www.adobe.com. (2007).

JPEG 2000:
ISO/IEC 15444-1:2004, Information technology – JPEG 2000 image coding system – Part 1: Core coding system, www.iso.org. (2004).

ISO/IEC 15444-2:2004, Information technology – JPEG 2000 image coding system: Extensions, www.iso.org. (2004)

Signatures
ETSI TS 102 778-3, Electronic Signatures and Infrastructures (ESI); PDF Advanced Electronic Signature Profiles; Part 2:"PAdES Basic – Profile based on ISO 32000-1"; ETSI, www.etsi.org. (2009)

PDF Advanced Electronic Signature (PAdES) FAQ: www.padesfaq.net

Receiving Deeds in PDF/A Streamlines Public Administration

Michele Piva,

Infocamere

When legislation and practice converge, complying with the regulations is simpler. That's been the case with the deeds that firms deposit with Infocamere, the Italian Chambers of Commerce Companies' Register. Computer documents from 2009 on have to be produced in PDF/A format. Infocamere carries out systematic compliance checks on them using procedures from callas software, which has been entrusted with the prompt validation of the files sent to the register.

Michele Piva, head of marketing for the document area of the Italian Chambers of Commerce IT department: "We're dealing with a standard that was already widely in use before the Ministerial Decree (of December 10th 2008 sanctioning the use of PDF/A) came into effect, exactly because it is derived from the PDF which firms had been widely using for some time".

Leading edge standard

Of the possible candidates for the long-term archiving of electronic documents, PDF/A won out over XPS (XML Paper Specification), TIFF G4 and JPEG. PDF/A was set as the standard ahead of the legislation. Infocamere has invested in it both for the necessity of making available documents that must keep their legal integrity and value over time – considering that the elimination of paper is a must in public administration – and because it considers that, as it is a reinforcement of PDF, the shift over to it could be taken for granted.

And the statistics are indeed showing this: whereas from the beginning of March to the middle of April already 47% of the deeds presented to the Companies' Register were compliant with the new measures, the documents deposited in the following month and a half achieved a format validation of over 58%.

Looking in particular at company constitution deeds (excluding documents classified as 'other deeds'), the most frequent file formats received by the Chambers of Commerce from amongst the over 580,000 documents analysed in the period under consideration were in a PDF/A compatible format. And for a good 89 of the 105 different Chambers, the percentage of valid files is over 80%. This shows that the adoption of the PDF/A is being generally and rapidly confirmed over the whole of Italy.

According to the ISO 19005-1 standard, PDF/A offers a mechanism "for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for the creation, memorization and rendering of the files". In other words, the PDF/A files are self-contained and, since they incorporate the information (content, colour, image, text and so on) necessary to display the document, they do not require any additional data for its correct visualisation. However, exactly on account of this property, the format must exclude certain functions allowed in normal PDF files such as transparency and multimedia content, and cannot include macro-instructions or rely on links to external resources.

The validation check

In order to facilitate the file production operations, Infocamere has chosen the easiest level of compliance to the standard – which corresponds to the sub-category

PDF/A-1b, for which no explicit logical structure is needed. Despite this, Infocamere has noted that there are some recurring errors: on the one hand, the poor indexing and aggregation of the deeds attached to the form; on the other hand, the widespread habit of creating a copy of the document which then doesn't have the original's representation (and its metadata) and so loses its validity.

It is a common misunderstanding that if a PDF is visualised "in PDF/A form", this means that the PDF fully complies with the PDF/A standard. This is a source of a widespread error and can happen even with Adobe Reader 9. Thus the only way to ensure adherence to the standard is the validation control.

To check this, it is necessary to use products such as Adobe Acrobat 9 Pro or pdfaPilot from callas software. The alternative is the free service, accessible from the tools section of Infocamere's WebTelemaco site, which employs the same tool used by the Companies' Register office and which enables the validation of any document whatsoever supplying the result online, complete with details of any non-conformities encountered.

Negative outcome – rejected form

For the validation procedures, Infocamere has used the server version of pdfaPilot and has developed an application which makes appropriate calls to it, determining the correspondence to the PDF/A-1b specifics. A good number of the people Infocamere works with, from service professionals to notaries and accountants, get a great deal of benefit from the tool on WebTelemaco. And, once Infocamere has received the files, the checks are regularly repeated. If the outcome is negative, the form is rejected.

The management application was created over the space of three months following approval of the Ministerial Decree and required an investment of around 30,000 Euro. This expense has already been repaid by making available a standard that guarantees long-term document validity and integrity: a guarantee of uniformity to a defined standard and, therefore, also of quality. Among the further benefits, apart from the return on investment, is the improvement of cash flow. Compliance to the standard ensures homogeneousness, which is a key element for its subsequent conservation and concession – on payment – to distributors and information agencies who use it for commercial purposes.

Italian IT network CGN enables 20,000 Professionals to Submit PDF/A-Deeds

Andrea Cappellato, CGN

Donato Stellabotte, Technosolutions

Final balance sheets, constitutional deeds, changes and operations on company shares, business transfers, modifications to company agreements, acquisitions, mergers, dissolutions and cancellations: the common denominator to all the annexes delivered to the Companies' Register is PDF/A format. "An obligation" says Giada Marangone, Head of Communication at CGN services: "which we adapted to both early and well". The first Italian IT network, CGN, brings together more than 20,000 professionals from the fields of accounting, law, tax and labour. They benefit from services for their day-to-day work such as Chamber of Commerce forms, tax obligations, legislative updates, data banks as well as consultancy and training.

The reasons behind the choice

Software is also one of the tools CGN supplies to the professionals in their network. So, for the conversion and validation of the files to be deposited in PDF/A format at the Chambers of Commerce, CGN set out to find an adequate tool. And the choice – after a complete analysis – was pdfaPilot from callas software.

It's a product that can be used both as a stand-alone application and as a plug-in, which CGN's team of developers has integrated into its own data transmission platform through which the professional members access its services. The search for a suitable product started by looking at the ones listed on the www.pdfa.org website. Various tests were carried out, since there are tens of thousands of forms presented to the Chambers by the members of the network and it was necessary to ensure a totally suitable conversion to PDF/A.

The objective was to find the right product to meet the indispensable requirements of lightness, ease-of-use and non-invasiveness with regards to the operating system. CGN periodically supplies its members with a CD containing the programs which enable interaction with the network's data transmission platform. Since the PCs which the various professionals' offices work with are very heterogeneous, sometimes with rather outdated operating environments too, it was fundamental that the application could operate without changing the registry file and without requiring particular configurations. All these were the key characteristics that CGN found in the generation and conversion tool already adopted by Infocamere, which CGN then decided to use themselves.

The CGN team has thus enriched its new services data transmission platform by giving the client's associates the same application. The software gets copied from a CD to a directory on the hard disk and can be used right away. The package's interface is typical for the Windows environment, doesn't require particular setups to function, and the execution of the operations takes place transparently.

A professional who has to produce a deed in PDF/A format, whether from a text processing application or from a PDF file, launches the virtual printer and gets the PDF/A document which is later validated by callas software. This may then be sent by Internet to the Chambers of Commerce Register. pdfaPilot does a good job not only in converting PDF files generated by Microsoft Office applications but also from all other sources. However, there are situations where the tool is not able to repair a PDF file accordingly. This is a difficult situation for

somebody who is not a computer expert. A certain expertise with the various stages between generation and transformation of the format is needed. To simplify the matter, and to supply valid answers in recurring cases, CGN has created a document that summarizes the solutions to the most frequent questions received by their call centre.

A necessary move

Based on the conversion capabilities of their software solution provided to the network members, and on the basis of their precise instructions on how to generate documents for compliance down the line, the users have been able to independently solve most of the problems.

The final step for the professionals is the transmission of the deeds in PDF/A format to Infocamere's web "front office" – Telemaco. First though, the form's sensitive data is transmitted to CGN for a final check, ensuring the "cleanliness" of what is going to be sent. In general, the task of CGN is to lighten the workload of the professionals who belong to their network by speeding up their operations. This clearly stated role, alongside free membership, may be the decisive factor for the professionals when deciding between the different providers.

Session Intro – Track 1: What You Need to Know About PDF/A and PDF/A-2

Session Chair:

Johannes Hesel,
Board, PDF/A Competence Center

The articles in this chapter contain the conference presentations from Track 1: What you need to know about PDF/A, for novice to intermediate level users.

Track 1 covers the basics of PDF/A and, after reading, you should have a fairly good overview of the format. More detailed information about PDF/A can be found in the subsequent chapters, or also in the PDF/A Competence Center's "bible": "PDF/A in a Nutshell". The book is a very good follow-up to the lectures and is available in English, French, German, Italian and Spanish.

Four interesting articles are included in this chapter, beginning with "PDF/A 101 – Introduction to PDF/A" by Raffaele Bernardinello of CMT Group, Italy. Raffaele describes in his article that PDF/A is the answer if a document format with nearly all the advantages of PDF is required (e.g. cross-platform availability, full colour support, full text retrieval, free available viewers for all platforms …), which also guarantees a certainty of reproduction, independent of the hardware or software used over the lifetime of the document.

This is a real requirement in many industries. For example, our customers – who are typically doing any kind of engineering – have to archive their drawings, specifications and documentation quite often for 30 years and longer. For them, TIFF/G4 was the only accepted archiving format in the past; some of them are in fact still storing their information on microfiche to guarantee persistent and certain reproducibility. Today however more and more of them, especially in critical industries like aerospace and defence, pharmaceutical industry or in the energy sector, are migrating to PDF/A as their unique archiving format.

When I describe PDF/A in a three minute talk, I like to say that, quite simply, PDF/A is just a subset of PDF that ensures long-term readability and, through it, every PDF becomes a good PDF.

After the introduction and description of PDF/A in general by Raffaele Bernardinello, Alessandro Beltrami from TechnoSolutions srl in Italy explains the role of PDF/A in the Graphic Arts (GA) world. File formats for the GA industry must be flexible on the one side and reliable in reproduction on the other side. PDF/X and PDF/A both offer flexibility and reliability, and will therefore coexist in the GA world in the future. This will especially be true after PDF/A-2 is released, offering more flexibility than PDF/A-1 does today.

Two articles follow which cover the two major sources of original documents.

"Scan to PDF/A with OCR" by Carsten Heiermann of LuraTech discusses documents that began as a piece of paper. The scanning of documents has been carried out for a long time already, and Carsten compares the older electronic formats with PDF/A. He describes the new and modern features which PDF/A allows for in scanned documents like high compression, metadata, colour scanning and full-text OCR.

The second group of documents includes the so-called "Digital-Born Documents". Dr. Hans Bärfuss from PDF Tools AG, Switzerland, describes possible sources of digital documents and the options for converting these into PDF/A. He also discusses different use-cases like e-mail archiving and how PDF/A can help in such an application.

I would like to thank the authors for their excellent contribution of articles, and hope the chapter makes a worthwhile reading for you!

PDF/A 101 – Introduction to PDF/A

Raffaele Bernardinello,

ICT Director,
C.M. Trading S.r.l.

A new PDF-based standard capable of guaranteeing document management, whether they need to be archived or reproduced in hard copy in large volumes. How do you make your documents compatible with this new standard?

Using PDF (Portable Document Format)

Right from its launch by Adobe in 1993, PDF has proved to be the most user-friendly format for managing electronic documents of any kind, ranging from accounting documents, maps and designs to books. Features such as easy interchange, the search functions provided, the provision of free viewers (not only Adobe), easy migration and legibility among the various operating systems (Microsoft Windows, Apple, Unix, Linux, Sun etc.) are just some of the many reasons which have resulted in PDF becoming "de facto" the most commonly used standard in the world for exchanging electronic documents.

The majority of document management, ERP and image management systems now have their own internal layout systems with native output in PDF. PDF files are not big, they can be easily archived and read on virtually any system and by everyone.

However, there are some areas of technology which still come under document management where PDF has not become the dominant format yet, for example, in the area of high-volume printing where AFP is still the benchmark format, or else scanning where TIFF format still plays an important role.

On the other hand, as a result of technological progress, most companies which need to manage their own documents are becoming independent when it comes to creating PDF documents. This also makes them independent in terms of managing these documents, whether this involves long-term archiving or sending the documents electronically to the recipients.

This is also the trend for the future. It is not for nothing that HW vendors have been adapting to this "de facto" standard for a few years now. This has led to the appearance of high-volume printers with management software capable of handling PDF as native input files as well, in addition to AFP format, along with high-volume scanners which generate PDF files directly, plus other developments.

Why use PDF/A?

Can we say absolutely that PDF is the right document management standard to meet every need? The reality is not as simple as this. PDF offers a host of benefits, but is not, unfortunately, a standard. Anyone who has had the chance to work with PDF is aware that PDF itself does not guarantee that a document in this format can be displayed, archived long term or printed in the correct form.

Proper character management, proper colour profile management or the document's actual dimensions are variables which are difficult to control exclusively via PDF. The actual development of PDF's capabilities (features for managing layers, encapsulating images, film clips etc.) means that a file's compatibility and ability to be displayed properly over time are not guaranteed.

In the particular case of high-volume printing (which is still a fundamental aspect of the document management process nowadays), problems caused by not managing some of the characters or images properly or relating to the management of the actual colour mean that the way in which the file is displayed sometimes does not match the original in the way the person who created the document intended. Original fonts are also substituted by others. These are just some of the problems which people managing PDF have ultimately had to resolve.

Documents sent to print where some characters disappear, letterheads which are superimposed because the viewer has substituted the character when displayed on screen, or else documents with abnormal dimensions or even actual porting from Windows to Mac or else Linux, AS400 etc.,

where character management is still an issue which needs to be dealt with today. These are a few of the problems with using PDF which are still unresolved at the moment.

But what is PDF/A?

PDF/A attempts to provide a solution to all these problems which can be summarised in a single statement: the certainty of reproducing the document correctly, irrespective of the HW and SW used. Consequently, the international community created an ISO specification, which is nothing more than a PDF subset. Created in 2002, PDF/A stipulates the basic rules for having a standard which can guarantee long-term archiving for documents.

The PDF/A initiative was kicked off in 2002 by AIIM (Association for Information and Image Management), NPES (National Printing Equipment Association) and the Administrative Office of the U.S. Courts. By 2005, PDF/A had been published as ISO 19005-1, where it is the cornerstone standard for electronic document file format for long-term archive and preservation. Today, AIIM provides the lead on the PDF/A ISO Standard and the PDF/A Competence Center is the major industry association supporting PDF/A, especially in Europe where adoption rates are higher than in North America. With all this in mind, it is easy to understand why the PDF/A standard is rapidly being required by governments and implemented by industries around the world.

The PDF/A standard is "a file format based on PDF which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rending the files".

The current PDF/A specification, PDF/A-1, is based on the PDF 1.4 specifications and has two levels. Adoption of the first level (PDF/A-1a) ensures the preservation of a document's logical structure and content text stream in natural reading order. This is critical when the document is displayed on a mobile device (for example a PDA) or other devices. This feature is commonly known as "Tagged PDFs". Some PDFs are created with sufficient information to meet this requirement; many PDFs created by production business processes do not contain this information and so fit into the second level.

The second level of compliance is referred to as PDF/A-1b. This level is the minimal standard that ensures the rendered visual appearance of the file is reproducible over the long-term. Specifically, PDF/A-1b ensures that the text (and additional content) can be correctly displayed (e.g. on a computer monitor or in hardcopy), but does not guarantee that extracted text will maintain the same structure as presented in the original document.

But using this standard, created for long-term archiving, may provide the standard to be used as the basis for all delivery or reproduction requirements, including high-volume printing. All this obviously leads to the request from the companies using this technology for a guarantee that their documents can be reproduced correctly, thereby providing a standard which eliminates all the issues mentioned above. In other words, a standard allowing documents to be printed, archived and displayed in the same way every time, whether now or in 100 years' time, without having to deal with the problem of medium type (paper, DVD etc.), viewer or type of computer which will be used to display them (Windows, Mac, Linux etc.).

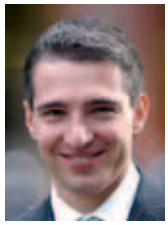
How to become PDF/A compliant

Most companies which use PDF nowadays do not generate PDF/A files and many are unaware of the aims and obligations associated with using an ISO standard format like PDF/A. However, processes and software are becoming available now on the market which can verify and convert a PDF file into a PDF/A file. This provides the guarantee for companies generating these files that their own documents will not be subject to variation over time and will be able to be printed or displayed in the same way every time, regardless of the computer or printer required to reproduce it.

There are a variety of software packages on the market for converting PDF to PDF/A, but none of them can guarantee 100% at the moment that they can resolve all the PDF document management issues, particularly the management of fonts and colour profiles, especially when the document is generated on different platforms to those to be used for the conversion process. Nevertheless, this is the way to go. The aim is to come up with a universally recognised standard which makes vendors, software houses, service providers etc. follow the same rules, thereby providing a guarantee both for anyone creating only electronic documents and for anyone who has to print these documents.

There is still a long way to go, but the guidelines have now been set out. There is definitely still a great deal to be done in providing information about the benefits of such standardisation. The new standard needs to be promoted at every level, among those involved in this area, as well as to both public and private companies.

PDF/A from Document Management to Graphic Arts



Alessandro Beltrami,

Consultant,
TechnoSolutions srl

Talking about PDF/A and PDF/X, or other specific PDF standards, means talking about two different worlds: Information Technology (IT) and Graphic Arts (GA). These two worlds are tightly connected in most document management environments but were created using different approaches.

For 30 years, IT has supplied the GA world with digital technology in graphic reproduction: the first colour electronic prepress system (CEPS) was introduced in 1979 in Milan, Italy, and started a revolution in GA technology [1]. The digital IT world has therefore existed for a few decades, but GA document management has centuries of history: the Gutenberg Bible is dated 1455. This is the reason why it is very hard for IT people to understand the GA methods for designing documents: the centuries of graphic design have left their mark on the language, working patterns and printing technology.

Basically, GA people search for reliability in a digital file. Reliability is the opposite of flexibility. For example, a PostScript file gives a lot of flexibility but only a medium level of reliability, whereas a CopyDot file presents the maximum of reliability but is difficult to handle. In the GA world, PDF is the balance between reliability and flexibility.

Today, the GA approach is influenced by people who work with analogue technology (films, analogue plates, torque…): this is the reason why the current PDF technology is not used by 100% of GA companies. Although standards like PDF are widely used, specific standards like PDF/X are not applied as often as managers and IT personnel would like. Fortunately, the scenario is changing very quickly with new generations of graphic artists who start their profession as digital natives. The use of proprietary file formats by specific composing software is declining compared to the use of specific PDF files: only complex graphic productions (ex. some packaging) are not yet covered by PDF ISO standards.

GA documents are created with a specific printing device or printing process in mind and contain elements that must be reproduced in the same way as designed on the original software. Operations like reflow, font substitution and change of resolution are not desired or not possible in most GA documents.

The first PDF/X standard (PDF/X-1a) was referred to as “digital film”: meaning that the document contained exactly the same content that user expected to see after the printing process (after impressing the analogue film). The other PDF/X versions that ISO TC130 defined (-3, -4…) contain certain information that need to be interpreted by the printing process workflow, like colour management, levels or transparencies, external references etc. This approach could be called “shared responsibility”

It is very important to consider this double approach (“digital film” vs. “shared responsibility”) when we talk about archiving or handling GA PDF documents. In most cases, GA digital data can be converted to PDF/X-1a. This is the scenario that can enable us to implement both the PDF/X and PDF/A standards. But new composing software produces content that can only be correctly exported in PDF/X-4; in this case it is very difficult to implement both the PDF/X and PDF/A standards today: we must wait for the PDF/A-2 file format. The other specific PDF formats like PDF/VT or PDF/E could also have an important role in some companies.

PDF/VT represents variable data printings, i.e. when a layout is fixed and the text changes every page (e.g. invoices). PDF/VT is strictly derived from PDF/X-4, and it therefore has the same limitations for PDF/A as mentioned above, until PDF/A-2 is released.

PDF/E is totally ignored by GA people, even though we hope to find in the future, and especially in the architectural world, mixed competences in companies (repro houses and copy services) that print and archive this form of document. Architectural projects are examples of an environment where the documents used require features of PDF/E for sizes, PDF/X for content presentation features and PDF/A for archiving purposes.

A typical GA workflow generates PDF directly from the applications or through a conversion of PostScript files. In the past, not all applications were able to correctly generate PDF files, and this is one of the reasons why some old GA people are suspicious of PDF. If we analyse the archive of a prepress company, we'll find a lot of documents where the text is converted into vectors (high reliability, low flexibility). In this case it's very hard to produce PDF/A-1b files with a minimum of text content. Automatic indexing and other features cannot be created.

Today we must accept having PDF/A-1b compliance files in the GA industry because they can be integrated with specific metadata to summarize the content: PDF/A-1a is not useful for the conversion of GA documents.

What are we expecting for the future?

PDF/A is supporting more and more features, and PDF/X potentially could be used for modern, cross-media production. A vision could be a company that produces content in both ways, to have the strongest graphic appearance and reliability and the strongest long-term archiving features.

References:

[1] The development of international graphic arts standardization forum – Beijing September 2009 – proceedings – D. McDowell

Scan to PDF/A with OCR – Paper Becomes Digital

Carsten Heiermann,

President,
LuraTech Europe GmbH

PDF/A documents are not only generated from digital sources – a large percentage of documents are created from scanned hard copies received by mail or from files that are being converted to digital form. In such cases, the company has no access to the original files, and the documents that need to be converted into electronic documents are merely paper copies. PDF/A is preferable to other electronic formats because it is an ISO standard and a target format that provides a range of benefits with regard to archiving and reusing content.

From analog to digital

The digitalization of paper documents (letters, files, invoices, photographs, and many more) is part of everyday life in many companies and institutions. There are many common processes depending on the various intended uses of the documents concerned.

Previous solutions for scanned documents

In the case of documents that only exist in black and white, such as invoices, TIFF G4 has often been used, a format that is still in use today. This format was developed for fax transmissions. If original colour documents exist, the JPEG image format is a popular choice. Other less common formats include PNG and BMP. In certain cases, special formats such as 'JPEG in TIFF' are preferred, in order to reduce the file size or create multi-page files, for example.

Disadvantages:

These older methods are subject to a series of disadvantages in comparison with the digitalization of documents using the PDF/A format. Users who still work with these older formats today will be confronted with problems such as the following:

■ **Format variety:** Because different file formats are required for different tasks, the older procedures do not result in a uniform format for scanned documents. In certain circumstances, users have to use a different viewer for each format. As a rule, each display program is operated differently. Only one viewer is required to display PDF and PDF/A – one of them, the Adobe Reader, is even available free-of-charge.

■ **Loss of information:** PDF/A is capable of adopting content in a one-to-one fashion. Other, older file formats cause the loss of detailed information. One example is TIFF G4, which can only display content in black and white.

■ **Image quality versus file size:** When using image file formats, users are often faced with a choice between bad quality or large files. For example, if using JPEG, the size of a file can only be reduced if the user accepts a consequential reduction in its quality. This disadvantage is particularly irksome when displaying text, and it can impede readability.

File size versus display quality: The images are derived from a page in DIN A4 format and the file sizes also refer to DIN A4 with a resolution of 300 dpi.

■ **The myth of revision-safe TIFF:** The commonly held opinion that storing documents and data in TIFF is sufficient to make them revision-proof is a falsity. Every format can be manipulated and it is especially easy to make slight changes to the TIFF format. Archive formats can only form

a revision-proof solution in the overall context of the system in which they are being used, whereby the systems themselves – for example, a DMS or financial accounting system – provide the required security, rather than the format used.

- **Non-uniform metadata:** If a file archive comprises a large number of documents in different formats, it is not possible to achieve standardized metadata for all the formats used. Each file format tends to build on its own proprietary solution, making standardization impossible. PDF/A provides a uniform metadata system. The standard XMP (Extensible Metadata Platform) integrates any additional information directly into the PDF file itself, making it permanently accessible. This means that users can call up specifications such as the author, access permissions, keywords, and copyright directly and without resorting to the use of a database.

- **Full-text search:** Most image formats do not support text recognition (OCR) for files. As a result, they – unlike PDF – do not permit a full-text search.

- **Laborious data recall:** Image formats only allow data to be recalled via databases, not at file level. Example: You want to find "Simon Sample's" personnel files. The database can localize all documents that name this person, but cannot highlight the exact location of the hit on the correct page. In the case of large documents, this can result in extremely time-consuming searches and – in consequence – high costs.

The PDF alternative

PDF is a modern, standardized alternative. Digitalization via conversion to PDF is already a popular choice for users who wish to standardize document formats (Image2PDF) or enable full-text searchability. PDF also permits the use of newer, more powerful compression formats, such as JPEG2000. Many users have switched to PDF in order to achieve metadata uniformity. Using PDF eliminates all the disadvantages of the older formats, but – even so – the traditional PDF format is not the best solution for every single usage area. If generating PDF, it makes sense to create PDF/A straight away! If you decide to use PDF as your archive format, it makes sense to use the PDF/A variant, since this is the only format that was developed as an ISO standard for long-term archiving.

Full-text search options in PDF/A

PDF enables text searches at file level. This improves the usability of the documents concerned in many areas, such as the following:

- Electronic libraries – after download
- Manuals, design documents, and construction files in archives for product liability purposes
- Documents that are sent to customers, tax consultants, or attorneys

Improved compression in PDF/A

For documents in black and white

An increasing number of customers who process black-and-white documents recognize the advantages provided by PDF/A. In the case of black-and-white documents, the JBIG2 compression format (standardized in ISO/IEC 14492) is particularly effective.

This compression format is positioned as an alternative to TIFF G4. JBIG2 allows users to choose between lossy and lossless compression. This technology, which is – as yet – not well-known, has been implemented in PDF/A and is available in Adobe Reader since Version 5.

The JBIG2 compression format significantly reduces file sizes for best-quality text (these values refer to a scanned page of DIN A4 in 300 dpi).

For colour documents

Colour is an important bearer of information. It can have both content-related and semantic significance. The processing of colour documents increases the productivity of employees and thereby helps companies to reduce costs.

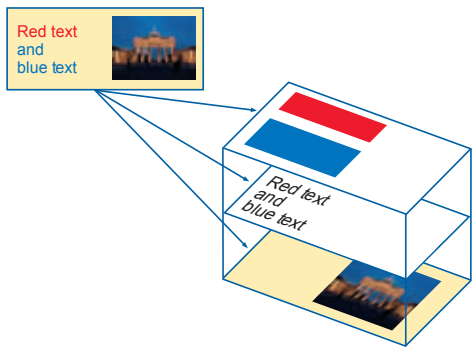
A study that was instigated by Kodak found that employees work better with colour documents, which bring the following advantages:

- Around 14% better comprehension of documents
- Around 70% improvement in decision-making ability
- 80% improvement in reading accuracy

Colour helps employees to understand content. Many colour documents lose important details when they are processed. For example, highlighted text can become illegible when scanned in black and white.

If all documents are scanned in colour rather than being separated into colour documents and black-and-white documents, the pre-sorting effort (which accounts for around 75% of the costs) is drastically reduced. This method also means that there is no need for changed scanner settings or rescans of a single document.

In the case of colour documents, powerful compression of the image data can reduce file sizes significantly. MRC compression – which is also known as JPEG2000 (JPM) – can drastically reduce file sizes without causing a visible decrease in the display quality.



The LuraTech layer procedure combines the benefits of the sharp display of colour images and text with a particularly small file size.

LuraTech uses a procedure that efficiently solves the problem of file size reduction in its Scan-to-PDF/A solutions. The division of each document into three layers that are converted entirely separately from each other enables the separate compression of text, colours, and images. Three-layer technology produces optimum quality by digitalizing a compressed original that splits the content into text, image, and colour layers using modern MRC procedures.

And, with the soon available PDF/A-2, there are even switches in Acrobat Reader, that allow to switch of background and foreground layers, so that even very low contrast scanned documents can be read very good. Switching off a background image of an old worn out scanned document before printing, helps to reproduce a version of that document, that has e.g. a crisp and clear coloured view on the data, without reprinting the "yellowish" background of old documents.

PDF/A – usage examples

The three case studies below show the advantages of the digitalization of documents via conversion to PDF/A for personnel records, knowledge bases, and credit files.

PDF/A for personnel records in a service company

This company is a services company that has a global turnover of 7.1 billion euros and a turnover of 420 million euros in Germany alone. A total of 220,000 employees work for the company worldwide, 14,500 of them being based in Germany.

The project

The task definition was as follows: 14,000 personnel records of around 150 pages each needed to be digitalized. This corresponds to a total processing volume of 2 million pages. These documents must be available to 200 authorized employees with access from 70 locations. The paper documents existed in black and white, greyscale, and colour. The solution was the conversion of the original documents into the ISO future-proof PDF/A variant of the PDF format along with effective compression to reduce file sizes as much as possible. The OCR (Optical Character Recognition) process prepared the scanned text for full-text searching.

The results

The uniform conversion of the document set into PDF/A enabled all personnel files to be safely retained in digital form. The ISO PDF/A standard guarantees the suitability

of data for long-term archiving. It makes it significantly easier to use the data, since employees now have access to documents that support full-text searching. The electronic search function replaces visual searching, resulting in a high accuracy of hits at the same time as saving time. Choosing the PDF/A format also results in files that are up to 60% smaller than if using TIFF or JPEG. Lastly, the smaller file sizes cause a significantly lower network load and permit direct access to data.

The advantages at a glance:

- Safe data storage for decades
- PDF files that support full-text searching
- Small file sizes (up to 60% reduction in size)
- Lower system load and quick access

The DAK: Migration of knowledge base to PDF/A

The DAK's INFO services needed to be digitalized to provide uniformity. The DAK (Deutsche Angestellten-Krankenkasse) is the second largest health insurance company in Germany, with 6.2 million members and 12,000 employees working in 750 branches.

The project

The internal information archive, which contains around 300,000 pages of text, took the form of image files before the migration. Most of the text was stored in TIFF format, with more recent additions in PDF. The stored information – originally stored on microfilm – was already partly digitalized, but using a mix of formats. TIFF, for example, neither saves space nor provides full-text search options. This archive is growing constantly, with around 3,000 new documents each year. Each file can have 50 or more pages. The aim of the project was to create a uniform archive with as low a file volume as possible while enabling digital data recall.

In order to optimise the possibilities provided by the info service and to make them future-proof, the DAK decided to archive the knowledge base in PDF/A format. The DAK used LuraTech's PDF/A solutions to carry out the migration. During this initial project, the DAK was able to gain early experience of the new PDF/A format that will be of use in later projects.

The results

The employees of the DAK's INFO service can now enjoy the advantages of easy and quick full-text search functions. The smaller file sizes allow information to be accessed more quickly. Naturally, a program for displaying the data must be installed on employee's PCs. The DAK uses Adobe Reader, which can be downloaded from the Internet free-of-charge. Thanks to PDF/A, the DAK's data is now suitable for long-term archiving in accordance with the ISO standard. Lastly, the DAK has gained practical experience from this reference project with regard to further data archiving using PDF/A.

The advantages at a glance:

- Files that support quick full-text search options
- Reduction in required disk space
- Quicker, easier access to documents for users
- Long-term readability
- Only one viewer required (Adobe Reader)
- Acquisition of practical PDF/A experience
- PDF/A for the decentralized scanning of credit files

An American finance company

In Tennessee, the headquarters of an American finance company, 'check into cash' procedure documents were digitalized and stored in a data archive in PDF/A format. The financial service provider concerned has 1,200 payday advance centres in 30 US states.

The project

The service provider required the decentralized scanning of credit files. Documents were to be processed in colour throughout. Lastly, the switch to the new system was to improve the transmission of data to headquarters.

The results achieved for the centres:

The centres now benefit from quick data transmission thanks to the implementation of the LuraDocument PDF Compressor, which creates PDF/A documents via scanning and data conversion procedures. All documents can

be processed in colour. This means that the centres do not need to sort documents into colour documents and black-and-white documents before digitalizing them. This has resulted in a considerable decrease in processing time.

The results achieved for the headquarters

The headquarters, where the PDF/A documents are stored, have benefited from a reduction in the required disk space since the modern data compression procedure used yields significantly smaller file sizes. Smaller file volumes also cause a noticeable reduction in administration costs. Last but not least, the company's headquarters benefit from the long-term readability of data and safe archiving in accordance with the ISO standard.

The advantages at a glance:

- No need to pre-sort documents into colour documents and black-and-white documents
- All credit files can be read in a single process
- Reduction in file sizes
- Quicker transmission of data
- Safe long-term archiving of credit files

Conclusion: PDF/A is the optimum format for scanned documents

PDF/A is the optimum format for scanned documents. It can be implemented in every single company and institution without any major technological problems. Anyone considering digitalizing paper documents today should choose the modern, standardized PDF/A solution straight away. In an environment where other formats have been used to archive scanned paper documents up until now, clearly defined, well arranged projects provide an opportunity to experience the advantages of PDF/A and gain practical experience of this new format.

With the upcoming new part two of the standard (PDF/A-2), there are even better results for scanned documents In PDF/A, using JPEG2000 the compression ratio and quality can again be improved and using the layer switches in PDF viewer software now allows even better access to the documents content.

Archiving Digital Documents – Conversion of Digital-born Documents to PDF/A

Dr. Hans Bärfuss,

CEO of PDF Tools AG
and Vice-Chairman of the
PDF/A Competence Center

1 Introduction

When compared with the preservation of data in its original format, there are many advantages to archiving documents and data from digital sources into PDF/A. The source applications are rapidly being developed further. As a result of this, after only a few years, the readability and the authentic display of data can no longer be guaranteed. Furthermore, a company must maintain all of the applications that are used and all of the platforms on which they operate. This incurs considerable costs. Even for documents and files that are created digitally, PDF/A is an excellent choice for long-term archiving and comes with great advantages with regard to uniformity, searchability and cost-effectiveness.

2 Development of digital documents as archive materials

The ECM model from AIIM distinguishes between five major processes in the management of business information: Capture, manage, deliver, preserve and store the documents. These processes can be easily assigned to the following PDF/A functions:

The ECM model from AIIM and the associated PDF/A functions

Digital documents are created in all of the mentioned processes and PDF/A is also important in all of these processes, although in different ways, as explained in the following.

What are the typical sources of digital documents that are later archived, and in which processes do these emerge?

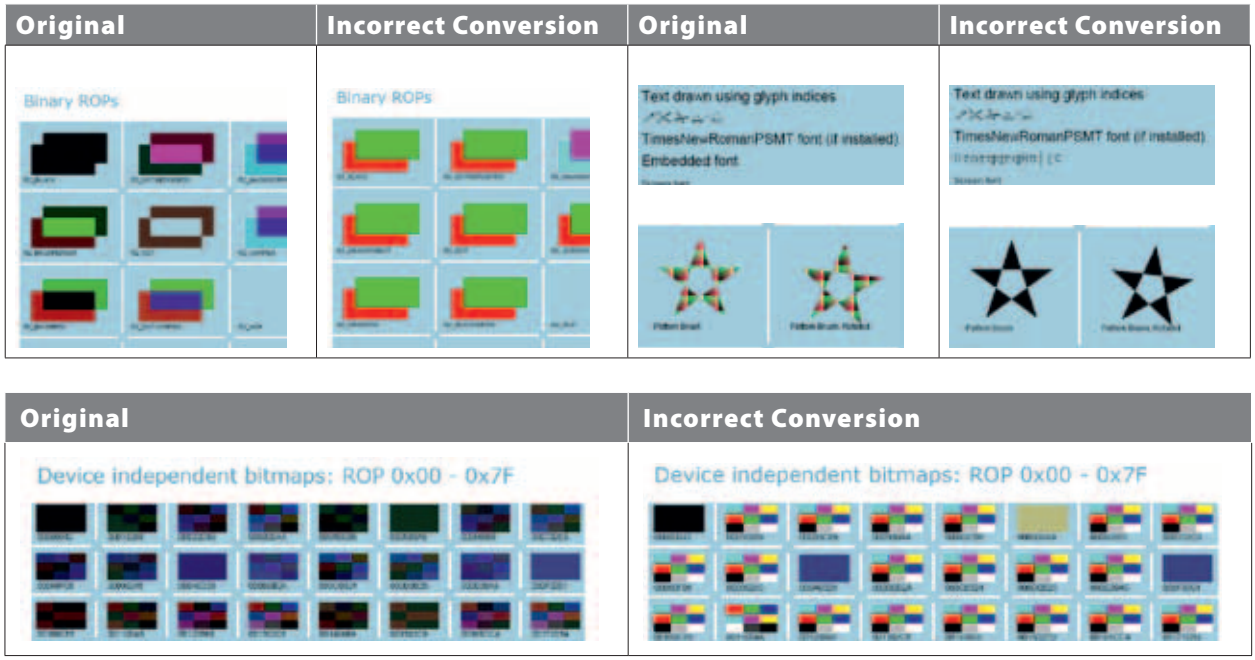
Process AIIM ECM Model	Use case	Applications/Examples
Capture	Inbox	- Scans with or without OCR - E-mails with or without attachments
Manage	Office, graphics and construction	- MS Word, Excel, PowerPoint, Visio, etc. - Illustrator, InDesign, Photoshop, etc. - CAD: Autocad, 3D Studio Max, etc.
Deliver	Outbox	- Print data streams: PostScript, PCL, AFP, etc.
Preserve	Archive migrations	- Masses of TIFF and other files, including source data (metadata, object relationships, etc.)
Deliver/Capture	Electronic data exchange	- SWIFT, EDIFACT, etc.

3 Attributes of analog and digital sources

Digital documents can emerge from analog and digital sources. Some parameters are relevant for their subsequent long-term archiving:

Attribute	Analog	Digital
Sources	Scanner, raster images	Standard and proprietary formats from applications and data streams, in file storage, mailboxes and attachments
Quality of the source	Good	Large differences
Complexity of the source	Low	Can be very high
Product differentiation	Compression rate, performance	Quality
Biggest challenge	OCR recognition rate	Loss of information during the conversion

From these differences, it is clear that we require different strategies for handling different sources, both in the general outline and in detail. These strategies are required both for the employees of IT departments, the records manager and for manufacturers of conversion products. The challenge here lies not only in creating a document that conforms to the PDF/A standard but in interpreting the source in such a way that the visual appearance corresponds to the original document. The following diagram shows the results of conversions to PDF/A whose form conforms to the standard, but whose visual appearance does not sufficiently correspond to that of the source:



Correct and incorrect conversions: In both cases, the result was a document that conforms to PDF/A, but, in the case of an incorrect conversion, does not correspond to the original document in any way.

4 Converting digital sources to PDF/A

4.1 Why convert?

Long-term archiving of digital data to PDF/A offers great advantages:

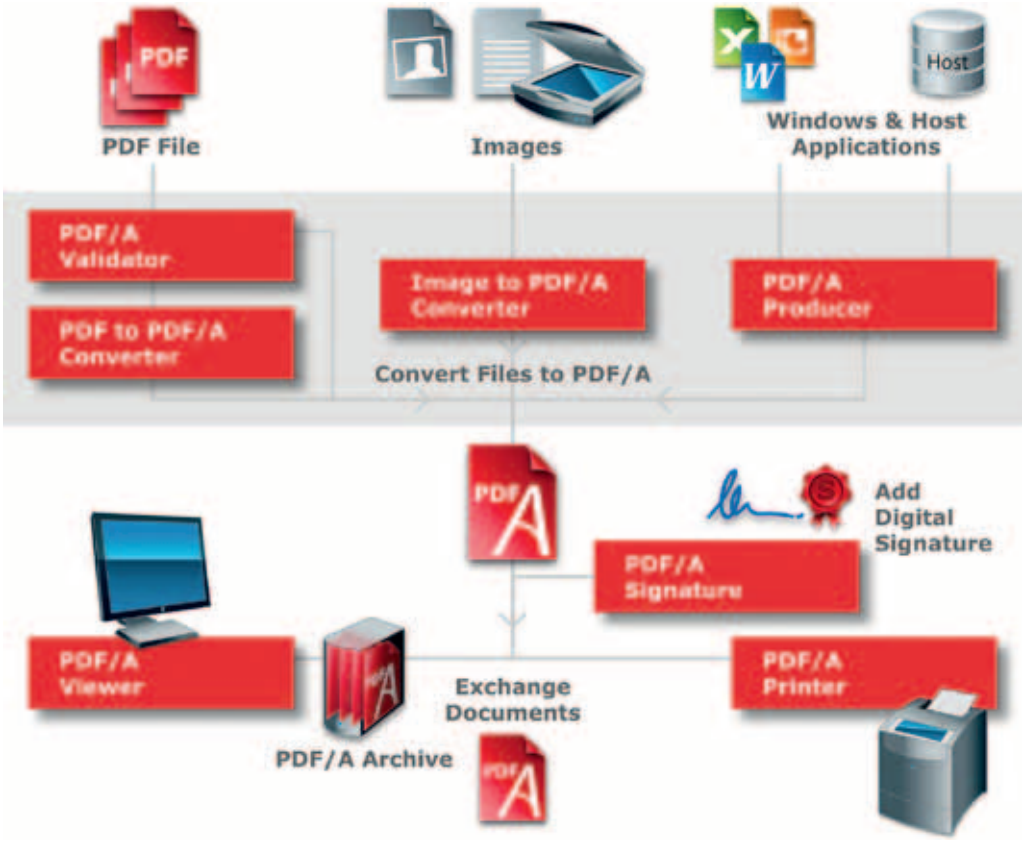
- The user does not have to maintain the original "native" applications and the platforms on which the applications operate.
- Users depend less on software manufacturers because all of the relevant information is saved in one ISO-standardized format and this format is manufacturer-independent.
- Simplified processing due to the fact that the archived data is standardized into one format.
- Option to perform a full-text search in all of the stored data.

These advantages also involve an economic benefit that must not be underestimated.

Of course, when compared to the native formats, archiving in PDF/A also has a few disadvantages, for example, due to the loss of interactivity or the built-in "functionality" of the native format. MS Excel should be used as an example here. MS Excel offers calculation formulas for content and these are lost during the conversion. Therefore, for these formats, it always makes sense to also archive the original document and to use the archiving in PDF/A as a fallback variant. With "interactive" files, the time for archiving can be chosen so that there is hardly any need for further changes (Document Lifecycle Management). In certain formats (for example, e-mails), the original document may have to be saved due to compliance reasons.

4.2 Overview Development and conversion processes

In the following complete overview, the development of digital documents (above) is particularly relevant:



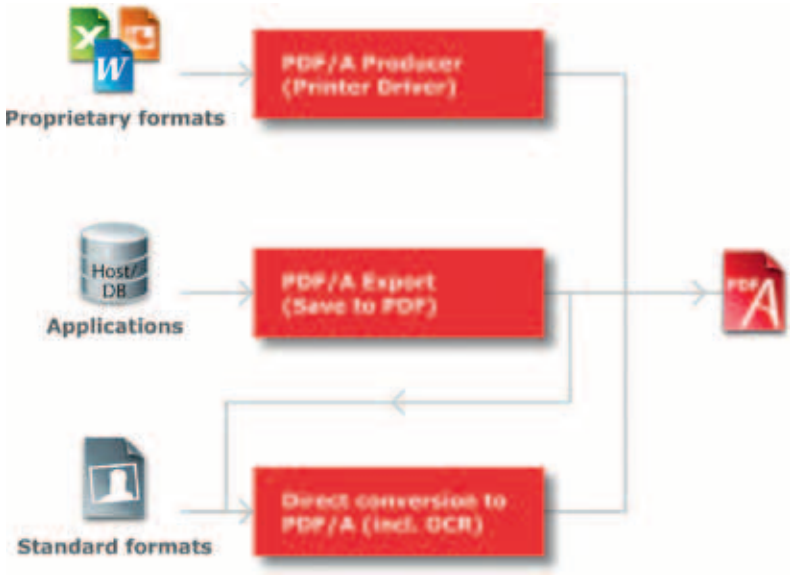
Complete overview of the PDF/A processes, with particular focus on the development of digital documents.

The easiest way to create PDF/A from proprietary formats such as Office documents, CAD drawings, etc. is to use an effective printer driver, also known as PDF Producer, PDF Creator or PDF Converter (for example, Adobe Distiller etc.). This "detour" via a printer driver is required because, so far, most native applications do not have a Save to PDF function. This function is now available for MS Office 2007 but it must be downloaded as a separate add-in.

The process of archiving e-mails, including attachments, to PDF/A (for example, from MS Outlook) is more complex. There are currently only a few providers of this type of functionality, for example, PDF Tools AG with their 3-Heights Document Converter Service, which converts an e-mail and its attachments into a single PDF/A document.

From databases, ERP systems, etc., PDF/A is usually controlled using an export function (Save to PDF). Often, these files must be post-processed because they do not completely conform to the standard. Another option here is the direct, programmatic creation of PDF and PDF/A files. In this process, the contents from any sources can be merged, for example, for processing personalized printed materials. PDFlib GmbH is one of the leading providers of these tools.

Specific tools are usually used to convert images and, in this process, an OCR function is important for the creation of metadata and for the searchability of the texts. In spite of this, even in scanned documents, we cannot underestimate the complexity of such applications, particularly in the areas of multiple formats (for example, dozens of variants of TIFF), colours, fonts and compression and segmenting procedures (for example, Mixed Raster Content). LuraTech offers the leading products in this area.



Converting digital sources to PDF/A using various conversion procedures

All conversion software in all of the areas must take into account the specific obligations and prohibitions from PDF/A, for example, the embedding of fonts, colour profiles and metadata (as XMP).

4.3 General Challenges

From a general perspective, when creating PDF/A from digital sources, we are confronted with the following challenges:

Area	Challenge
Colours	If the colour profiles from the sources are missing, assumptions are made about the colour space.
Fonts	If fonts (or glyphs) are missing, replacement fonts must be selected. To do this, the text must be a Unicode text.
Transparency	The flattening of transparency is complex and may lead to the loss of information (fonts, vectors, etc.)
Levels, interactive and multimedia elements	Only the "Print Preview" is retained
Actions	Functionality (JavaScripts etc.) is lost
Digital signatures	Check, document, sign again

4.4 Converting E-mails

An e-mail can contain all types of documents, interlaced archives and much more (executable files etc.). In addition, the e-mail can contain internal or external references (e.g., HTML mails) and different systems, interfaces, file systems and data streams are involved. The process of archiving e-mails, including attachments, is therefore effectively the "supreme discipline" of archiving in PDF/A, since all of the challenges in connection with converting sources that were originally analog or digital must be solved using one single product.

To solve this, a different conversion strategy must be selected for each individual element of an e-mail: The e-mail body and attachments are converted individually and, only then, are merged into a single document. In this PDF/A document, each attachment can then be identified using a so-called bookmark entry. By doing this, the structure of the e-mails can also still be traced at a later point. In addition, information, such as tables of contents from Word documents, is not lost, because these are mapped as a second level of hierarchy in the bookmarks and are linked accordingly in the PDF/A. Even the handling of digital signatures poses a challenge when archiving e-mails.

4.5 Converting websites

The topic of archiving websites is relatively new. This basically involves retaining the contents and state of one's own website in a way that is legally trustworthy so that the required evidence can be provided in legal or other procedures.

The difficulty when archiving websites is that the output using a print driver does not normally represent the authentic appearance of the website, because websites are usually specially prepared for printing. To be able to bring forward trustworthy evidence, this "true to the original" is crucially important.

Therefore, from the website, a "Capture" function is used to create an image that is merged with the relevant text and other information (fonts, colour spaces, etc.) to effectively produce a "vectorised, searchable screenshot". Another complex issue is the handling of external links and the internal link structure of a website. In addition, it is necessary to decide on one browser and one browser version because different browsers and browser versions display websites differently.

4.6 Converting on the client or on the server

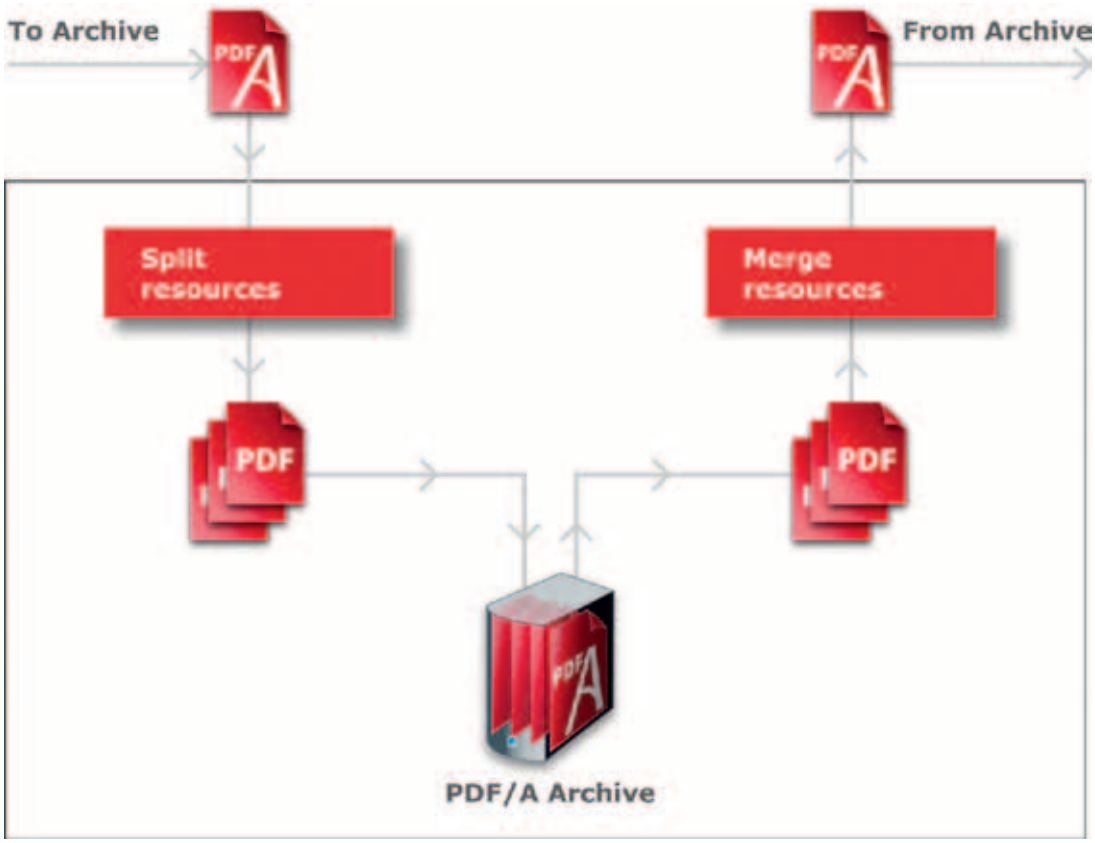
We must consider the following aspects with regard to the question of whether conversion software should be installed on individual clients or on a central server:

Attribute	Client	Server
Scaling workstations	Small amount	Large amount
Distribution	Complex	Simple
Robustness for the users	Depends on the creator-application	Independent
Performance for the users	Restricted by the client	Scalable
Supported source formats	Restricted by the installation	Scalable
Application support	Local	Central

4.7 Font handling in mass archiving

Single, individual PDF/A documents can be directly archived. When archiving large quantities of similar PDF/A documents (for example, telecom invoices etc.), the situation often arises in which the documents contain the same fonts, logos or other corporate identity elements that must also be archived for each individual document. The repeated saving of collective resources (fonts, images) is undesirable and reduces the acceptance of PDF/A.

To solve this, the archive system can be upgraded using an add-in that separates the shared resources and saves them in only one instance for all documents when performing mass archiving of PDF/A documents. When a document is accessed, the shared resources are again merged with the document to produce a complete PDF/A document. This procedure can also be used for digitally-signed documents, but, during the signing process, the document must already be prepared for the separation of resources.



Concept for preventing superfluous saving of resources (e.g., fonts) in mass archiving.

4.8 Legal security with digital signature

The process of digitally signing PDF/A files that derive from digitally-created documents brings greater legal security. Depending on the application, the user must be clear about what the signature really provides. In any case, in a qualified electronic signature, it is absolutely clear at what time the conversion occurred and whether the document has been changed since the conversion. It is also clear who performed the conversion process in a company.

However, the uncertainty that arises from the "dynamic" source (e.g., a database) of such a PDF/A document cannot be dispelled. Nor is it possible to verify whether the created PDF/A document actually corresponds to the appearance of the original document (e.g., a Word document) or whether all of the information that is contained in the document (e.g., contents and e-mail attachments) actually exists in the PDF/A file. To increase the credibility of such documents, the entire process must be certified. This is therefore a topic that transcends the simple use of digital signatures. However, such certifications require a certain volume of data so that this is worthwhile for service providers, manufacturers of software and systems and large companies.

4.9 Quality assurance by validators

"Trust is good, control is better": This, of course, also applies for PDF/A documents and products that create PDF/A. Or that claim to create PDF/A. Not all the products that are labelled as PDF/A are actually PDF/A products. In extreme cases, the archiving of company data can be crucial for the existence of a company. For example, in a court case, if the exonerative records have not been prepared or have not been prepared correctly.

It is therefore important to use tools that ensure the highest standards of quality. Validators exist to determine if a tool fulfils this prerequisite. These validators also need to be checked. For this task, the PDF/A Competence Center created a freely-available test suite that systematically breaches the standard and then checks that a validator can identify all of the breaches.

The use of a validator is not only important when evaluating a tool, but it is also important in the operational processes. A validator should therefore be used regularly to check the conformity of the created PDF/A documents – as a permanent quality check. This is because different sources, application versions, etc. may lead to different conversion results.

5 Summary

PDF/A is beneficial as a format for archiving digital documents and can lead to considerable cost savings in comparison to archiving in the native format. However, the devil is in the details with this and we must not underestimate the complexity that arises depending on the source of the digital documents. It is therefore essential to collaborate with specialists in this area and this collaboration can protect users from unnecessary costs accrued through incorrect processes etc.

For both day-to-day business and from a strategic point of view (e.g., in legal cases), it is very important that information can be accessed quickly and securely. Discrepancies in this area can result in damage to a company's image or in substantial financial consequences. Processes for archiving directly from digital data are therefore given top priority.

Session Intro – Track 2: Legal Certainty

Session Chair:

Dr. Bernd Wild,
Board, PDF/A Competence Center

The archiving of information and documents takes up more and more space in the modern information society. On the one hand, the number of new documents, whether paper-based or already electronic, increases every year. On the other hand, legal and organisational regulations demand that these documents are maintained for many years or even many decades. It is not unusual for business documents relating to tax to be archived for 10 years, or over 30 years for documents in the health sector and up to over 90 years in the plant engineering or aircraft industries.

In addition to the length of time for the archiving, the scope of the information that is to be archived also changes. Where this was restricted several years ago to a relatively few documents that were seen as business-relevant in organisations, all communication between business partners and within an organisation is coming to the fore as a result of rules, e.g. SOX or the more generally formatted "Compliance". This particularly affects the archiving of e-mails, which has come to represent a fundamental medium of business communication.

One of the main advantages of PDF/A is that it supports embedded electronic signatures. This means that for scanned documents, electronic invoices, forms and contracts you can have a legally binding signature that freezes their content reliably. The basic requirements for electronic signatures, and the supported standards, is presented by Andrea Valle in "PDF/A and Digital Signatures".

One of the success stories in using PDF/A and electronic signatures is clearly eInvoicing. While public authorities make requirements on authenticity, integrity and readability of invoice documents, they also request a stable and consistent electronic archiving of tax-relevant documents. This brings us to the presentation by Domenico Barile: "Electronic Invoices as PDF/A With Respect to Italian Legal Requirements", which describes the possibilities and potential of electronically signed PDF/A invoices in the light of Italian tax legislation.

Since PDF/A was the first standardised document format for long-term archiving suitable for archiving different document types, and is also suitable for PDF, we must pay careful attention to several aspects regarding the conversion of PDF documents to PDF/A. This particularly concerns the handling of character sets that must be embedded in order to achieve a reproduction of a document according to the PDF/A standard that is true to the original. The presentation by François Fernandes: "Reproducibility of Archived Documents", deals with the hazards that you must consider when converting PDF documents to PDF/A.

Although the conversion of PDF documents to PDF/A has been slowly developing into an established practice, the use of the ISO format during the archiving of e-mails is still in the early stages. Among other things, this is due to the fact that an e-mail often does not represent a homogenous document, but is more of a container for various file formats. The requirements that exist when archiving e-mails in PDF/A, and the relevant approaches, are the theme of the last presentation by Dr. Bernd Wild: "The Challenges in Archiving e-mails with PDF/A".

Electronic Invoices as PDF/A With Respect to Italian Legal Requirements

Domenico Barile,

E-Mission S.r.l

Italian laws prescribe requirements for electronic documents, their reproduction and storage. Let's take a rapid look, focusing only on some specific regulations that a service creating electronic invoices needs to conform to:

- A bill dated January 23rd, 2004 from the Ministry of the Economy and Finance contains the fundamental definition of an electronic document: static and non-modifiable documents are drawn up in such a way that the content cannot be modified during access and storage, and the document remains unchanged over time.

- Article 3 of this regulation defines the requirements for electronic documents relevant for tax purposes: they need to be electronic documents issued by affixing a time stamp and a qualified electronic signature, in order to ensure certification of date and their authenticity and integrity.

- Article 6 specifies that electronic documents must be made legible and, on request, available on hardcopy or electronic medium at the place of storage, in case of verifications, audits or inspections.

This law is an update of many older ones and was passed to guarantee that electronic documents will have the same value as paper documents, especially for tax purposes and in case of verifications and audits. Electronic documents can be created in many ways, but only the ones that are completely comparable to paper documents conform to the law.

The main purpose of the law was to integrate electronic invoices (that only make up a small portion of all documents received by companies) into the inspection system: making a hardcopy of them would leave the method of inspection as before. The electronic invoices should replace the paper pages. Since the inspection results could have an important influence on our customer's business, these legal requirements also have some implications on the data-stream format of electronic documents. The electronic processes used for representing electronic documents will be compared in the following section.

XML vs. PDF

In the past, TIFF files were used for generating, reproducing and storing documents. But we had to improve services to give our customer more flexibility, usability and other new special services. TIFF files were not viable for the radical upgrade we wanted to do, so we focused on XML and PDF. As opposed to TIFF, these file formats not only represent a document, but can in addition be a carrier of other information and used in different systems.

Therefore we examined XML flows and PDF files to understand which of them could be the right format to respect all legal requirements for electronic document representation, particularly for electronic invoices.

A digitally signed XML flow is always non-modifiable: it can't contain anything that modifies its content, so its representation will never change without invalidating its digital signature.

A digitally signed PDF which includes JavaScript, or other advanced features that could be used to modify its content, could theoretically change its representation itself without invalidating the digital signature. However, a PDF can only be used for electronic invoices, according to Italian law, if it doesn't have features that could modify its content. This is a disadvantage to using standard PDF files.

Both formats support time stamps as well as a qualified electronic signatures, and they are recognised world-wide.

Although PDF files are usually human readable (in the same way as paper documents, if they are correctly interpreted by a browser's rendering engine), printable and can be sent by e-mail, an XML flow is not human readable

without specific knowledge and its own XML schema. It can be printed and sent by e-mail, but its paper representation doesn't look like the original document (the paper used before switching to electronic document). This is a disadvantage for XML, because human readability is absolutely necessary for verifications, audits or inspections.

For XML dataflows, Italian law restricts the addressees who may receive a document, and allows the use of an XML dataflow environment to only those customers with a signed agreement. This restriction could generate a complex management of contracts with customers which can be avoided with PDF, because PDF files can be used by everyone without any restrictions.

Electronic Invoice for E-Mission

Since we must guarantee that a document cannot be modified, the only way to create electronic invoices that conform to Italian law is to use PDF files. However, there aren't any universally recognised methods for creating an unalterable PDF – or at least, there weren't…

PDF/A

PDF/A-1 is an ISO standard for long-term archiving. It was published on October 1st, 2005 based on the PDF Reference Version 1.4 from Adobe Systems Inc.

PDF/A files are viewable now and in the future on every system, independent of producer, operating system and viewer: it's a "self-contained" format with all resources embedded in the same file.

PDF/A stores objects, allowing for an efficient full-text search in an entire archive, and requires only a fraction of the memory space of original or TIFF files without loss of quality. A typical invoice, consisting of an image with one or two different standard fonts for text, saved as a TIFF file at 300 DPI requires 50Kbytes. Saved as a PDF/A file it requires at most 35Kbytes.

Metadata like title, author, creation date, modification date, subject, keywords, etc. can be stored in a PDF/A file. PDF/A files can be automatically classified based on the metadata, without requiring human intervention. Electronic documents in PDF/A format can contain metadata for automatic archiving or for carrying data to receivers. Creating a proper XMP extension schema makes it theoretically possible to load all metadata that are contained in the files in another representation, or to load all the information that are necessary for receivers into a PDF/A file.

JavaScript and advanced features that add interactivity or modify content are not allowed, so PDF/A could be the right file format for saving electronic invoices according to Italian laws. Digitally signing a PDF/A file of an invoice respects all the restrictions dictated by Italian law. PDF/A files could be checked by third party analysers and validators. Then the inspectors would have the instruments to check that the PDF file is really a PDF/A file and is conforming to legal restrictions.

Electronic Invoices as PDF/A for E-Mission

Respecting the PDF/A standard and using qualified electronic signatures can guarantee compliance with all Italian legal requirements – now and probably in the near future.

Since PDF/A features may be the only ones that perfectly conform to Italian legal requirements, the electronic invoice for E-Mission is a PDF/A conforming file, with time stamp and qualified electronic signature embedded and external UBL (Universal Business Language) data linked to the document.

UBL (Universal Business Language), a little widening

Since its approval, XML has been adopted in a number of industries as a framework for the definition of the messages exchanged in electronic commerce. The widespread use of XML has led to the development of multiple industry-specific XML versions of such basic documents as purchase orders, shipping notices, and invoices.

While industry-specific data formats have the advantage of maximal optimization for their business context, the existence of different formats to accomplish the same purpose in different business domains is marked by a number of significant disadvantages as well.

- Developing and maintaining multiple versions of common business documents like purchase orders and invoices is a major duplication of effort.
- Creating and maintaining multiple adapters to enable interfacing across domain boundaries is an even greater effort.
- The existence of multiple XML formats makes it much harder to integrate XML business messages with back-office systems.

- The need to support an arbitrary number of XML formats makes tools more expensive and trained workers harder to find.

The OASIS Universal Business Language (UBL) is intended to solve these problems by defining a generic XML interchange format for business documents that can be extended to meet the requirements of particular industries.

- A library of XML schemas for reusable data components such as "Address," "Item," and "Payment" – the common data elements of everyday business documents.

- A set of XML schemas for common business documents such as "Order," "Despatch Advice," and "Invoice" that are constructed from the UBL library components and can be used in generic procurement and transportation contexts.

UBL is designed to provide a universally understood and recognised commercial syntax for legally binding business documents. It operates within a standard business framework such as ISO 15000 (ebXML) to provide a complete, standards-based infrastructure that can extend the benefits of existing EDI systems to businesses of all sizes. UBL is freely available to everyone without legal restrictions or licensing fees.

UBL schemas are modular, reusable, and extensible in XML-aware ways. As the first standard implementation of ebXML Core Components Technical Specification 2.01, the UBL Library is based on a conceptual model of information components known as Business Information Entities (BIEs). These components are assembled into specific document models such as "Order" and "Invoice". These document assembly models are then transformed in accordance with UBL naming and design rules into W3C XSD schema syntax. This approach facilitates the creation of UBL-based document types beyond those specified in this release.

The decision to use UBL files to carry data to receivers as external XML simplifies the service:

- when an XMP extension schema is created, all the PDFs that contain the metadata must load it (need to be self-contained)

- UBL is a standard; creating a new XMP extension schema could be a duplicate

- not all customer's users need to receive accounting data using the service

The link between PDF/A and UBL is made by filenames and standard metadata saved in the PDF/A.

E-Mission.Weaver

The purpose of E-Mission.Weaver is to automate the entire supply cycle and to carry electronic invoices from suppliers to customers. E-Mission.Weaver can be used from the order to the payment, and in each phase an electronic document can be generated as PDF/A for visualisation and storing in document management systems.

The figure depicts a common flow between a company and its suppliers and customers when using E-Mission. Weaver. E-Mission is the link between parties: it receives data spools from the suppliers, generates electronic invoices and creates UBL containing accounting data (the Italian law allows for a service provider to emit invoices for another company). Then it sends all the electronic invoices and the accounting data to the service user (E-Mission customer).

The flow can also be directed to the end-customer. The invoice spools are received by E-Mission, which creates electronic invoices for the service user. The data extracted from the spools is saved in UBL, enabling the end-customer to automatically load it into his accounting system.

The cycle can be a closed-loop with the return of invoice recording data and payment data. All of the documents emitted are optically stored using Archiva services.

E-Mission S.r.l. started its business in October 2008 as a spin-off of Archiva S.r.l., which has been on the market since 1979 with substitutive optical document storage. All document distribution services were entrusted to E-Mission with the target of completely automating and improving them.

The main activities of E-mission are:

- examination and updating of all national and international standards for document storage and transmission

- development of new technologies that include data and document transmission and sharing between different and heterogeneous business partners

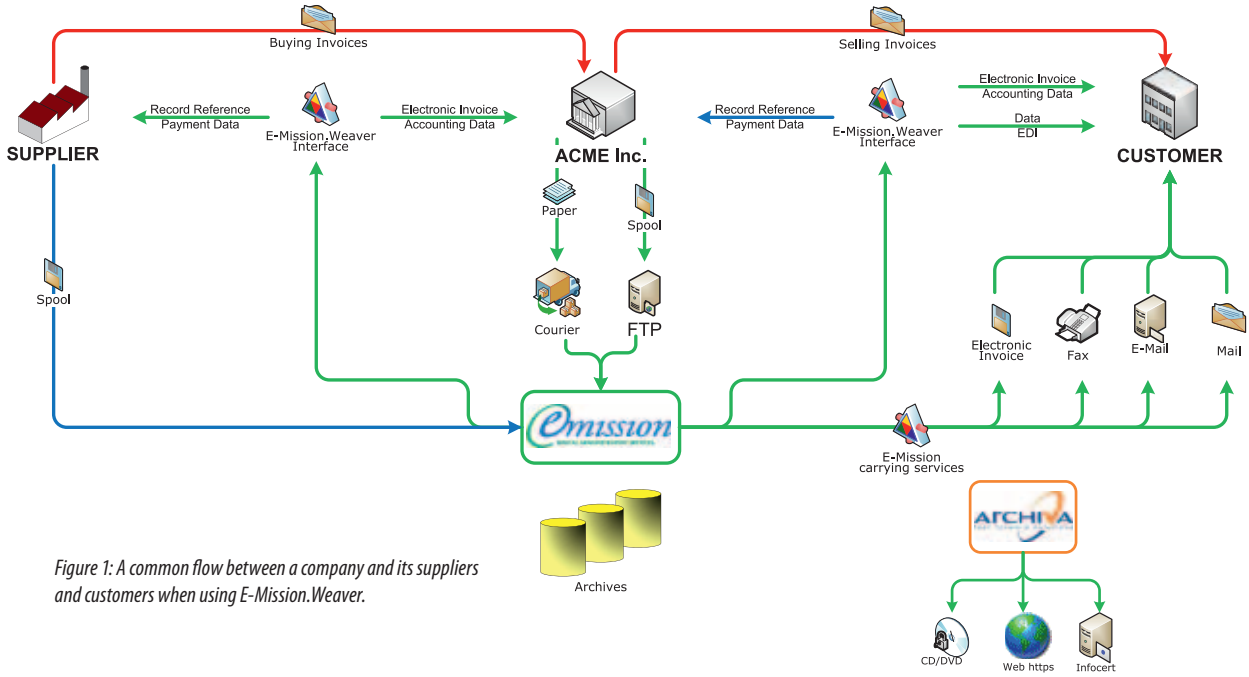


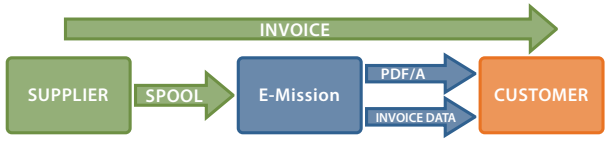
Figure 1: A common flow between a company and its suppliers and customers when using E-Mission.Weaver.

- consultation to develop dedicated customer solutions that result from an analysis of their corporate requirements and processes
- implementation with customers of the data transmission and sharing services, using the solutions that have been selected
- in-house management of the processes that provide the services

Currently the company handles the transmission of documents of any type and nature (commercial, tax, administrative, technical, etc.) in hardcopy and through electronic mailing services. With Weaver it has completed its range of services being offered with the transmission of the data contained in these documents, creating a new transmission engine open to all new technologies and all present and future communications standards (XML, EDI, etc.).

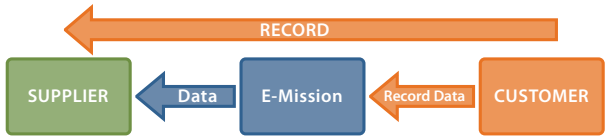
Weaver – Basic Implementation Level

The first implementation phase is the generation of invoices for a supplier (E-Mission emits invoices in his name according to Italian laws). Electronic invoices in PDF/A (with time stamp and qualified electronic signature embedded), and data contained in them saved as UBL files, are transmitted between the parties.



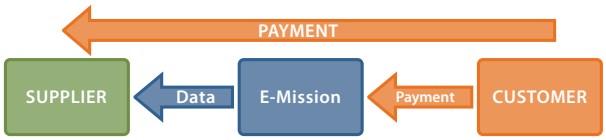
All data exchanges are made by using a custom web service. UBL data is only used for storage information, but it is sent using the interface to avoid customers having to keep up to date with UBL upgrades.

The customer can put references to invoice records in the database to E-Mission.Weaver, delivering them to the supplier when he logs into the service. This helps control the cycle.



After payment, the customer gives the payment reference for downloaded invoices to E.Mission.Weaver. These

are delivered to the supplier to close out the cycle (after bank account verification).



This level of implementation can be adopted by all E-Mission customers, independent of whether they are suppliers or customers (or both).

Service structure

E-Mission.Weaver was created on a server structure that guarantees the best levels of security, redundancy, consistency and traceability. The web service is a WCF service (Windows Communication Foundation) based on SSL layer and accessible by all types of systems. The binding forces to use a specific certificate for all requests to the E-Mission.Weaver's service is another level of security and it is also a way of having traceability for all transactions.

The use of distributed sessions increases consistency levels to the best available, avoiding differences between supplier and customer accounting systems.

All PDF/A files are generated using PDFlib libraries (PDFlib, PDFlib TET and PDFlib PLOP). We worked with the PDFlib development team to support the use of Italian smartcards during the electronic qualified signing of PDF/A.

Weaver – Extended Service Levels

In another implementation, E-Mission.Weaver integrates the management of delivery notes and invoices with the transfer of transaction data and PDF/A documents between suppliers and customers, for both types of documents.

In this instance, the process starts with the delivery of goods (or services). The supplier sends a delivery note spool to E-Mission, which in turn creates PDF/A electronic delivery notes and UBL data for all customers. When a customer logs into the service, E-Mission.Weaver transfers to him all of his electronic delivery notes and all the data contained in the UBL, for direct loading into his warehouse system.

Then the process continues with the management of the invoices linked to already processed delivery notes processed as described above.

In this phase, customers are able to load the transaction data into their warehouse system, even before the goods arrive!

In the next version that's under development, E-Mission.Weaver will handle the complete supply chain with the integrated management of orders, delivery notes and invoices, including the data transfer and the exchange of the PDF/A documents, between suppliers and customers.

The process starts with the order being placed and entered by the customer and introduced into the E-Mission. Weaver service, sending an order spool to E-Mission. When the supplier logs into the service, E-Mission.Weaver sends him all the pending orders. After this step, the process will continue with the management of delivery notes and associated invoices.

The process finishes by releasing the process data into EDI workflows, enabling service users to fulfil all their customer requirements without any limitation.

Italian law and PDF/A

Finally, there is an important news about PDF/A and Italian law: for the first time, the Italian Government has written a law that refers to PDF/A instead of generic PDF. The Prime Minister's decree dated December 10th, 2008 concerning "balance sheet presentation to registry office in XBRL" (eXtensible Business Reporting Language) contains the possibility of using PDF/A file format instead of XBRL files for the balance's presentation, if the accounting system is not able to generate XBRL files.

It might seem like a small detail, but the decree is a part of the "eGov2012" plan and it could be the starting point for officially adopting the PDF/A format as a standard in Italy.

That's what we and the rest of the PDF/A Competence Center are hoping for…

References:

http://www.pdfa.org
http://ubl.xml.org
http://www.pdflib.com
http://www.funzionepubblica.gov.it/ministro/pdf/DPCM_10dicembre2008.pdf (only in Italian)
http://www.governo.it/governoinforma/dossier/piano_e_gov_2012/ (only in Italian)

Reproducibility of Archived Documents

François Fernandes,

levigo solutions GmbH

Documents must be archived. Electronic archiving has become a universally recognized and practical method of digitally maintaining information. The formats that are used vary from simple raster formats (BMP, PNG and so on) to formats that have complex structures (MO:DCA, AFP) and also include PDF and PDF/A. As the complexity of the individual formats increases, the requirements for the structure and completeness of the documents must be adjusted accordingly and realized consistently. The aim is to ensure that you can reproduce these documents even after a long time. It only becomes clear at the time of the reproduction whether the criteria for a successful reproduction were also consistently implemented and realized.

What are the criteria in this case? With a view to reproducing documents at a later stage, we can divide the criteria into three categories:

- The standard on which the document is based must be clearly identifiable.
- All of the required information must be available.
- The document must comply with the standards.

These seem to be very basic criteria and, in theory, they should be met without any problems. Unfortunately, in this topic, there is also a significant gap between the theory and reality. The aim of this track is to identify the details of the requirements for a successful reproduction and to highlight the reason why PDF/A is suitable for archiving. The aim is also to highlight how these requirements can be used in PDF and PDF/A.

The standard on which the document is based must be clearly identifiable.

You want to reproduce a document after 10 years of digital archiving. To do this, you require the relevant tool that can interpret and reproduce the document.

How do you choose a suitable tool (for example, a viewer, a converter etc.)? The data format of the document is extremely important here. Based on the data, the format and the method of interpretation should be identifiable. The format must therefore be identifiable. Depending on the format, a problem may occur if insufficient information is available. An example of this is simple text documents that do not usually contain any information about the type of data. If the document was not set to ASCII when it was created, problems also occur when identifying a suitable encoding. As a result, for example, special characters may not be correctly converted. In contrast to this, it is very simple to identify PDF documents: The header of a document (specifically "%PDF-1.x") specifies not only that this is a PDF document, but also the version of the PDF specification on which this document is based.

You can use two basic elements to identify PDF/A documents. Firstly, the document contains the PDF header that is mentioned above. Secondly, the XMP metadata contains further information about the PDF/A conformity. This information specifies the PDF/A version on which the document is based and the conformity level that was reached.

All of the required information must be available.

If the data format of a document could be identified, the reproduction then occurs (using the suitable tool). It is then important that all of the required data exists and is complete. What does this mean exactly and what information does this concern? For PDF, it is not easy to provide a qualified answer with regards to the required data. It is clear that the body of the document must follow the guidelines that are defined in the standard version. However, we do not immediately recognize that additional standards (or at least the specified formats) are used in a document. As a result, for example, TrueType, Type1 and

(as of PDF 1.6) OpenType fonts are integrated into PDF documents. Image data formats such as JPEG, JBIG2 and JPEG2000 (PDF 1.5) may also be used.

An important topic, and one of the reasons why PDF is increasingly used, is the integration of textual contents without having to destructively integrate these into an alternative format. This would be destructive because a format that is based on raster is often used for archiving (for example, TIFF). The text data can continue to be visually reproduced but cannot be used for any other purpose. Thus, texts and fonts in a PDF document are of particular interest and are therefore the focus of this examination.

According to the PDF specification (not PDF/A), it is optional to embed fonts in a document or not. This obviously contradicts the requirement that all of the information must be available. If the PDF specification also allows documents without embedded fonts, why is it important that these fonts are embedded for the reproduction of archived documents? For this, we must consider in more detail the process when displaying textual contents.

To be able to describe the details as clearly as possible, the concepts that are used below must be clarified. Three basic components are important for texts: The character codes, an encoding and the glyphs. Character codes are binary data that usually consist of one or more bytes. These codes aim to represent a specific character. A glyph is the graphic representation of a character that is provided by a font. Character codes and glyphs are not directly linked to each other. An encoding establishes an assignment. This specifies how to map the character codes to the glyphs. A simple example for this is ASCII, which, for example, defines that a byte that has the value 0x41 must be interpreted as the character "A".

How is this procedure specified for PDF? To reproduce textual contents, the system reads the character codes that are stored in the document and maps these to the characters in a font. This procedure is described below using TrueType as the example. The procedure for other font formats varies somewhat in order to uncover the idiosyncrasies of the formats. However, the basic principle is similar.

Figure 1 displays the basic procedure for mapping character codes to TrueType glyphs. A number of steps are required to reproduce the characters. Firstly, the character codes must be extracted from the PDF document. To map the character codes to glyphs, an encoding is required. For this, the encoding is defined in the document using a basis encoding and any adjustments. You can now map the character codes to the character names, for example, /A, /at or /atilde. Based on these character names, you can use two methods to select a glyph.

- The character names are mapped to the Unicode values using a defined table (the "Adobe Glyph List"). The Unicode values are mapped to the glyphs using an internal mapping table of the TrueType font.

- If the first method fails (for example, for special characters etc.), the system uses the character names to select the glyphs using the Postscript table of the font.

When creating documents, we meet a number of assumptions regarding the font that is to be used. Examples of these assumptions are the expectation that a specific Unicode value leads to a specific character, or that a specific postscript name can be used to select a glyph. If the font is available in the document, these assumptions do

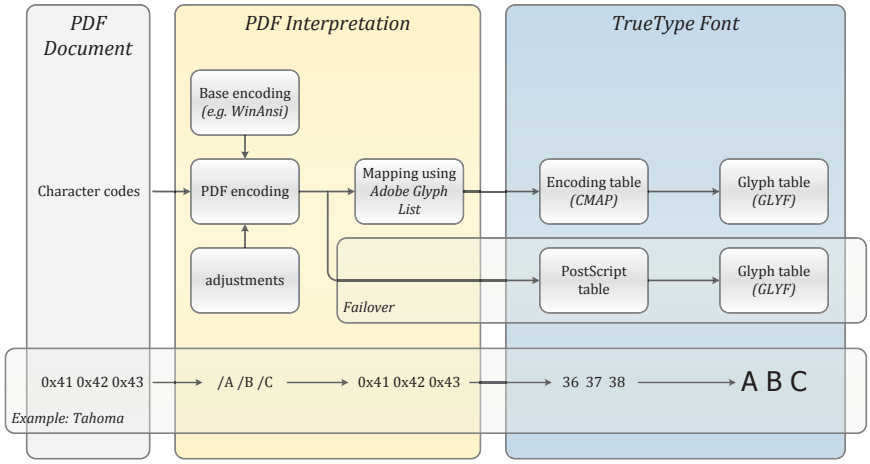


Figure 1 Locating characters in the example TrueType

not cause any problems, because the system already determines which font to use for the reproduction when it generates the document.

But how does this affect non-embedded fonts?

In such a case, the processing application must find this special font and, if it does not exist, must find a suitable replacement. The application uses the data from the system to limit the search for a replacement. Usually, the application searches through the existing fonts in the system and selects a suitable replacement font. If, in the worst-case scenario, the system has only one font, the procedure is not very likely to succeed. The affected assumptions may or may not apply for the font that is used as a substitute. The results depend on the options for the processing system and, in certain circumstances, for the operating system that is used. This may frequently result in texts that have random characters, or a selection of incorrect special characters (for example, checkboxes).

What help can PDF/A offer?

The defined aim of PDF/A is to prevent these problems. This is achieved by, for example, requiring that fonts are always embedded. You can use this restriction to ensure that all of the required information is available. In addition to requiring embedding, PDF also clearly regulates the additional specifications regarding the fonts and texts. If some attributes are marked as "optional" in the PDF specification, these are required for a valid PDF/A document. These clear definitions increase the chances that the document will be successfully reproduced, even after a long period of archiving.

The document must comply with the standards.

In accordance with PDF/A, all of the required data (fonts, colour profiles and so on) is now embedded into the document and is available for the reproduction. In spite of these preparations, in some documents, the images in some documents and the texts in others are displayed incorrectly or not displayed at all. What is the problem here? In addition to completeness, another important requirement arises and this requirement is extremely important for a reproduction. All of the data must conform to the standard. This means that even additional resources (such as fonts, images and so on) must be correct.

Problems with embedded resources are rarely identified at first glance but these may lead to problems in the long term. If, for example, the data for a font is incorrect, the processing application must deal with this. In this case, there is no specific information about how the data must be processed. If the data in a font is incorrect, heuristics are often used to attempt to identify the intention of the data. These heuristics are not documented and this is a detail of the implementation of the processing applications. In this case, most of today's products already behave in different ways. The future development of the applications is in debate.

If the data cannot be used, despite any corrections, it is common to use a font that already exists in the system as a substitute. This is the same as the situation in which the font is not embedded in the document. This situation also has all of the associated disadvantages. The same problems occur not only for fonts but also for all types of additional data. JPEG decompressors contain, for example, a number of heuristics to determine the colour space of the image data if this has not been explicitly defined.

Summary

PDF is a very comprehensive format. It offers a large number of options when processing contained data and when preparing additional information (for example, metadata). To successfully reproduce a document so that it is true to the original, a number of must be observed and adhered to. This process is made easier by using PDF/A because potential problems are clearly distinguished by the standard and, where required, optional information becomes compulsory. As a result, PDF/A documents are, to a large extent, self-explanatory. Questionable or incomplete aspects in the PDF reference were made clearer, excluded from PDF/A or restricted.

PDF/A is a future-proof format and is being used more and more. It offers various advantages when compared to the "classic" formats, such as TIFF and MO:DCA. The completeness of PDF/A and the fact that PDF/A is an international standard are good reasons to use PDF/A for digital long-term archiving.

In this context, we must also mention the work of the PDF/A Competence Center: It is made up of a large community that concerns itself with the topic of PDF/A. In this way, we can keep the PDF/A specification in the spotlight and we can exchange experiences. An additional group is the TWG (Technical Working Group), which deals with the details of the standard PDF/A and, through teamwork, clarifies many questionable aspects of the standard.

Archiving E-mails with PDF/A

Dr. Bernd Wild,

intarsys consulting GmbH,
Board, PDF/A Competence Center

The legal requirements for archiving business documents are increasing the need to also archive e-mail correspondence. This raises the question of which archiving format is most suited for the long-term archiving of e-mails. The PDF/A-1 standard also opens up new possibilities for using a uniform long-term format that contains the document character of an e-mail, can be searched for full text and, at the same time, can contain important metadata. A concept for converting e-mails to PDF/A documents is introduced and particular attention is paid to the process of handling file attachments.

Motivation for E-mail Archiving

Business communication via e-mail has overtaken classic postal delivery by far. If there was no e-mail, many business processes could not be performed in the time that is available and to the level of quality that is required. Originally intended as a transport medium for notifications, e-mail systems have developed into more of a "Document Management System" thanks to their capabilities to exchange documents and information of any type and the flexible storage in e-mail accounts and folders. Along with the file system, e-mail systems are therefore one of the most important document storage and management systems. Often, unless the life-cycle of a document (creation, revisions, completion, release) is managed using a genuine DMS (Document Management System), it can only be retraced using the mail history.

Current studies assume that between 35% and 70% of all business communication and information in a company is now transported and stored using e-mail. Due to the importance for transaction processing, this medium is gaining more focus in legal regulations that are putting the legal regulation to obtain data and the burden of proof for e-mails on an equal footing to paper-based documents. In addition to the legal restrictions (data protection), the process of archiving all e-mail traffic also faces technical restrictions. Therefore, issues surrounding data volumes, the ability to search in the saved e-mails and the handling of spam e-mails must be solved. When it comes to the commercial law obligations to archive data for at least ten years, the main issue is how to choose a suitable storage format and how to handle file attachments.

The "10 Golden Rules" of E-mail Archiving

Based on European regulations and directives, and the archiving recommendations of the German Association for Information Management (VOI), we can state the "10 Golden Rules" for e-mail archiving:

1. E-mail archiving has to ensure the authentic and unaltered storage of all information contained within an e-mail, including formatting and copyright information, attachments and electronic signatures.

2. Archived e-mails must be reproducible without loss of information.

3. Scalability concerning number and data volume of the e-mails is required.

4. It should be possible to delete an already archived e-mail partly or in total from the production mail system.

5. Encrypted e-mail has to be archived, either decrypted or else encrypted together with the appropriate decrypting key.

6. Electronically signed e-mails or attachments have to be archived. Additionally, provisions should be made to ensure long-term stability of the signatures by renewing digital signatures.

7. Attachments and the mail body text should be converted into a document format which is appropriate for long-term archiving.

8. Single instance archiving should be used in order to eliminate duplicates.

9. The legal requirements of the local country concerning archiving have to be fulfilled.

10. E-mail archiving has to be regarded as being only one component in an overall archiving concept, and not a stand-alone system.

While these rules have to be implemented by an appropriate archiving system regardless of the document formats used, there are indications which lean towards PDF/A being recommended as the ideal candidate. Especially the requirements for a long-term archiving document format (7), the lossless storage of information (2) and the support for electronic signatures point to PDF/A as a basic format for e-mail archiving.

Architecture

An e-mail consists of three parts: the header, the body and the optional attachments. Although the e-mail body comprises the contents of an e-mail that can actually be read, the header consists of attribute-value pairs that contain meta information about the e-mail. In addition to the date the e-mail was sent and the sender address, this also includes the destination address and the subject of the message. As well as these attributes that are required in accordance with the standard RFC 5322 [1], the header often contains routing information about the mail gateway (envelope sender) that participates in the transport, specifications for the coding of the mail text and a mail identification number. Since sending e-mails using the SMTP protocol is based purely on ASCII (as was also the case in the early days of the Internet), all of the additional formatting and enhancements must be coded accordingly to an ASCII basis. MIME is the established standard for this and is defined in RFC 2387 [2]. The coding of the e-mail header may be text, HTML or MIME. For compatibility reasons from mail clients, in addition to the message that is coded in HTML or MIME, a textual representation of the message is often inserted. This is particularly the case for formatted e-mails. If the sender also wants to send file attachments (non-text items such as images, PDF files, Office files etc.) as well as the actual message, these items are also coded in MIME. The participating mail client programs often have no knowledge of the attached files and, instead, they simply run MIME coding or decoding. Therefore, specific application software is responsible for displaying e-mail attachments and this software must be available on the target client system.

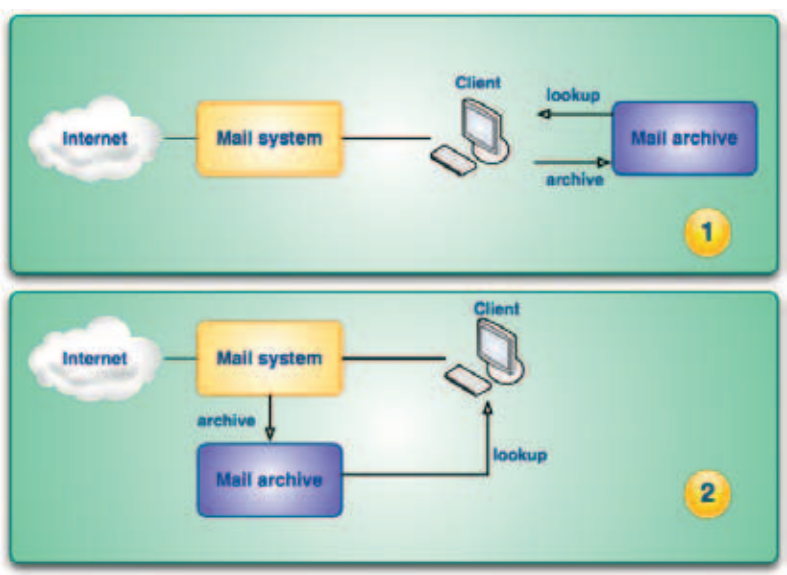


Fig. 1: System architecture in e-mail archiving

As displayed in Fig. 1, when archiving e-mail traffic, we can choose between a client-side approach (1) or a server-side (2) approach. In client-side archiving, within the e-mail client program, the client selects which e-mails must be archived. This can be done using manual selection, rules such as date created or date received or receiving addresses. When archiving is requested, the affected mails are then completely stored in the mail archive. Depending on the implementation, information that refers to the e-mail address in the archive may remain in the e-mail system. A web application is usually used to search for archived e-mails. This application should offer the option to search for full text and the option to search using specific criteria

from the e-mail. The e-mails that are found are then displayed as HTML pages. Any attachments remain unchanged, which is why the original programs are required at the workstation in order to open attachments. This approach cannot allow for universal archiving, because the end user makes the individual decisions regarding the archiving.

In the server-side archiving, the entire incoming and outgoing mail traffic is stored according to rules. In addition to an efficient spam filter, care must also be taken with regards to who can access the central archived e-mails by searching in this way.

PDF/A and E-mails

Most of the e-mail archiving solutions that are currently available save the e-mail that is to be archived either in the original format (text, MIME) or as an HTML page. In contrast to electronic documents, the visual appearance of the body of the e-mail is not usually the most important aspect for e-mails. Yet the majority of exchanged e-mails are based on simple text and can be read, created and processed using command prompt tools. As a result, a preset layout does not have to be kept. The proportion of formatted e-mails is increasing but, due to the predominant information characters in an e-mail, there is no specification for the appearance of a graphic format. The prominent attribute of the PDF/A format does not come into effect here. The metadata of an e-mail plays an important role. This metadata is frequently used for verification purposes. In this case, PDF/A, supported by XMP (eXtensible Metadata Platform), is a powerful tool for the structured storage of metadata and has the ability to reproduce attributes of the envelope sender and attributes that are specific to the mail system. A PDF/A file can therefore reproduce the complete range of information from an e-mail. Due to the XMP structure, targeted searches for metadata from the e-mail can be performed, irrespective of the archive system that is being used. When storing a PDF/A e-mail in the archive, its meta information can be used for the specific index management of the archive system.

File attachments require special attention because these cannot always be properly converted into PDF/A. While the automatic conversion of word-processing documents (for example, Microsoft Word or OpenOffice) is possible and useful, this is very problematic for files from programs such as Microsoft Project or Microsoft Excel. This is either because print areas can only be defined interactively or because the dynamic attributes of the file are essential for the information content. If media formats such as MP3 or MPG are also used, then no conversion can be performed at all. However, PDF/A conversions can be performed for many image formats.

A PDF/A-based Approach

Fig. 2 shows a basic approach to e-mail archiving in PDF/A.

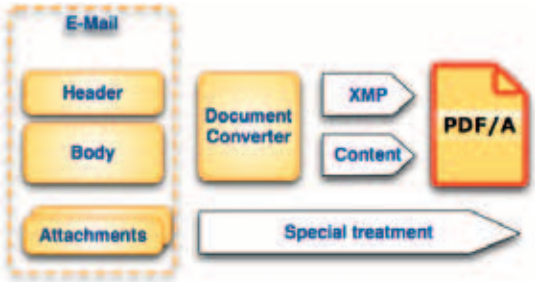


Fig. 2: An approach for converting e-mails to PDF/A

The three main components of an e-mail are each handled in a particular way. The header and body are fed into a converter. For the body, this converter converts the text or MIME-coded contents into PDF/A. The metadata from the e-mail header is converted into an XMP structure and embedded into the PDF/A document. An XMP schema is a prerequisite for the XMP conversion. This XMP schema defines the attributes that can be used and their semantics in the form of tags. An important requirement of the PDF/A-1 standard [3] is that the schema must be embedded. For new attributes, an automatic schema generation can be performed or these attributes can be stored, in condensed form, in one single extension attribute.

Attachments must be dealt with in a special way. A decision about whether or not to perform a conversion is based on the conversion matrix and the file type of the attachment (see fig. 3).

Since, from experience, most file attachments are Word, OpenOffice or PDF files and these formats can be converted to PDF/A, this results only in a small number of file attachments, which must be saved as unchanged source documents in the archive. However, for many convertible source documents, it may be useful to save the original as well as the PDF/A equivalent, particularly if dynamic contents should be retained.

By breaking down and converting a complex e-mail with attachments into one or more PDF/A files, the integ-

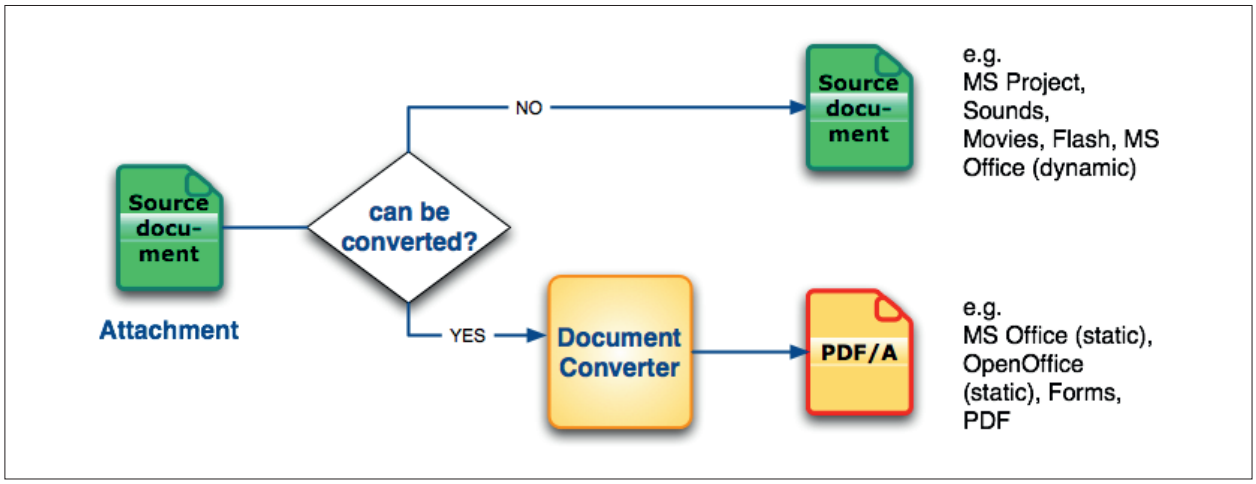


Fig. 3: Case distinction when converting file attachments

rity of the e-mail is lost. This integrity was guaranteed by the MIME container in the original e-mail. Since, in accordance with the PDF/A-1 standard, PDF/A files cannot avail of embedded PDF/A files (file collections) or references that can be resolved externally, the archive system must perform this task. This must secure the entity of the original e-mail in contrast to the PDF/A file of the mail body and the converted or unconvertible file attachments. Current archive systems include the relevant precautions for this. This should become easier when the PDF/A-2 standard is released because then even embedded files can be used in PDF/A-2 files. However, in this case, only PDF/A-1 or PDF/A-2 documents can be embedded.

Conclusion

The PDF/A-1 standard opens up new opportunities for the long-term archiving of e-mails. By rejecting the option of saving e-mails in different original formats and by converting to PDF/A-1, a uniform archive format can be used. At the same time, the mail metadata can also be completely integrated into the archive document. This means that when extracting from the archive, these specifications remain associated with the e-mail header and a formatted visualization is possible. For most Office formats, the file attachments can be converted and this conversion permits a "frozen" display of the attachment. If it does not seem to make sense to convert the data, you must archive the original file attachment.

References

[1] IETF, RFC 5322 (2008).
[2] IETF, RFC 2387 (1998).
[3] PDF/A Competence Center, TechNote 0009: XMP Extension Schemas in PDF/A-1 (2008).

Session Intro – Track 3: Accessibility and Metadata

Session Chair:

Olaf Drümmer,
Board, PDF/A Competence Center

The year 2011 is going to be an interesting one. Not only will PDF/A see the addition of a second part – "PDF/A-2" – to catch up with technological developments; some related standards will also enter the stage.

- A new standard – called PDF/UA – will address accessibility of PDF documents. It has been worked on for about five years already and is going to be finalised and published in 2011.

- The world of metadata for file formats like PDF will see a very important step as well: Adobe is currently releasing their XMP (Extensible Metadata Platform) specification to ISO, and if all goes well, XMP will be an official ISO standard in 2011.

Accessibility

The PDF/A committee in ISO always saw the need to not only cover archival in the sense of visual reproducibility, but to also capture and preserve as much semantics and content structure as possible. In order to achieve this, the conformance level "A" was introduced (although some say the letter "A" stands for "advanced", linking the "A" to "accessible" may be more adequate). Level "A" requires that text can be mapped to Unicode and that the semantic structure of the content – for example its reading order – must be reflected in the tags used to structure the PDF. One of the aspects the PDF/A committee was not able (and actually never attempted) to achieve was to offer strict rules or even guidance on how to ensure or enforce good quality tagging. There are also no provisions in PDF/A on how to best take advantage of structure information inside a tagged PDF.

This is exactly what PDF/UA is taking care of. The preparation of the PDF/UA standard has taken quite some time for a reason – it was very important to the PDF/UA committee to get it right as much as possible the first time around. Not only should the use of the PDF/UA lead to better structured PDF, but it should still remain feasible and cost efficient to get there, whether for those producing tagged PDF content, for software vendors developing tools for creation of well structured PDF, or for makers of assistive technology.

The two accessibility presentations in this track will bring you up to speed regarding accessibility in PDF, as well as PDF/A, and beyond. Both speakers – David Hook, Director Product Management at Crawford Technologies and Duff Johnson, CEO, Appligent Document Solutions – have been involved in accessibility for a very long time, and both have actively contributed to the development of PDF/UA.

Metadata

During the past decade several PDF related standards began to make use of XMP metadata instead of using the simpler and less powerful "Document Information" mechanism, which essentially is a list of key value pairs in PDF syntax. Even PDF itself is moving towards using XMP exclusively for any general purpose metadata. The next version of PDF, to be called PDF 2.0 and to be published as ISO 32000-2 in late 2011 or in 2012, will deprecate the use of the Document Information mechanism.

While some communities like photographers make very active use of XMP for embedding information like copyright notices, keywords or descriptions, metadata in the form of XMP inside PDF files has not yet become as prominent and widely used in the world of document management and archiving. A number of companies who switched from other ways of associating metadata with their documents though found substantial advantages in using XMP. In today's world, where exchange of structured data using XML-based formats is more widely understood than ten years ago, XMP turns out to be a natural fit for most metadata needs. The metadata sessions track will get organisations started looking into the use of metadata inside PDF and PDF/A, and will illustrate the power and cost efficiency of XMP over other approaches.

Accessibility in PDF and Elsewhere

David Hook,

Director Product Management, Crawford Technologies

We often forget that a significant percentage of all printed pages produced every year are not books, manuals, newspapers and other such documents. The bulk of printed output are the insurance policies, bank statements, telephone bills, invoices, statements and other documents that are sent to our homes. The quantity of these documents produced annually number in the hundreds of billions of pages. In the United States alone, over 80 billion transactional mail pieces are sent annually, each including multiple pages.

In the customer communications industry there are many significant trends taking place:

- The migration from printed documents to electronic documents; highly relevant to our focus on PDF/A
- The requirements for organizations to archive these transactional documents for legal and compliance reasons; also highly relevant to our focus on PDF/A
- The introduction of new production technologies, such as low-cost full color and other developments, which are drawing attention for their cost reduction and marketing potential
- The need to further protect customers' private and confidential data in both the physical and electronic production workflows for these documents

In addition to these trends we now must add the ability to provide document accessibility solutions. Companies have to serve the needs of those consumers who cannot access printed or electronically presented document content.

This article draws on our company's 15 years of experience providing solutions to the transactional customer communications industry, our experience providing document accessibility services and our own extensive research.

Read on to find out more regarding the various physical and electronic options for document accessibility, how PDF/A forms the foundation upon which accessibility for electronic documents may be built and how PDF/UA will become the key standard that will serve this important community.

Why does PDF/A matter for Transactional documents?

Most organizations now realize that they need to keep the transactional documents that they send to their customers in an archive format for compliance, legal, workflow, data extraction and other purposes. As organizations contemplate the architecture and file storage formats involved, PDF/A is clearly one of the leading standards due to its fidelity, popularity and reliability.

The fact that PDF/A is the defined international standard as an "electronic document file format for long-term preservation" speaks volumes. In other words, PDF/A is becoming significantly important to the customer communications transactional document industry.

However, many organizations do not understand the two conformance levels for PDF/A, and the implications surrounding which conformance level they choose.

PDF/A-1b is the minimum conformance level for PDF/A and it ensures that the "rendered visual appearance" is reproducible over the long term. Most organizations in this industry use this conformance level today for

PDF/A. However, it does not guarantee that extracted text will be legible or comprehensible – this is an important consideration that we will discuss below.

PDF/A-1a is the full compliance level and preserves the natural reading order and content text stream. This requires significant structural tags, and provides:

- Text reflowing capabilities which are important for mobile devices such as smart phones and other mobile devices due to their small screen size. The Gartner group predicts that mobile devices will outnumber PCs as the most common web access device by 2013, so this is an important consideration.
- Reliable extraction of text from the PDF to facilitate down stream production workflows including integrity checking, and the additional ability to use data from within the PDF for indexing and retrieval purposes.
- Much of the foundation upon which to build accessible PDF/UA documents.

Due to our industry's need to reliably store these documents in 'true fidelity' format, the need to properly support PDF on mobile devices, the need to reliably use content within PDFs for production workflows and the need to build PDF documents that are accessible, we strongly recommend that organizations go well beyond the minimum standard supported by PDF/A-1b.

To meet the minimum requirements a transactional document:

- Must be tagged as per ISO 32000-1, 14.8
- Must be tagged in the logical reading order
- Must properly tag all headings as per ISO 32000-1:2008, 14.8.4.3.2
- Must properly tag tables according to ISO 32000-1:2008, Table 337 and Table 349 – this is very important for transactional documents.
- Should use bookmarks whenever possible to assist in navigation – this is very useful for long business invoices and for insurance documents.
- Must use descriptive alternate text to image tags – this would help a user to understand that a graphic depicts electricity usage history in a utility bill, for example

What are Transactional Customer Communications?

Transactional Customer Communications are the documents sent to you via the mail or electronically that record 'point in time' information about your contractual and financial relationship with an organization that provides services for you. These documents officially communicate your transactions, amounts due, the minimum payment you must make, what happens if you don't pay on time, interest charges, insurance coverage details, the change in your investments and other information.

Organizations that produce these documents include banks, wealth management organizations, credit card companies, insurance companies, telecommunications companies, utility companies, government departments and other organizations that are compelled by regulations to deliver these documents to their clients.

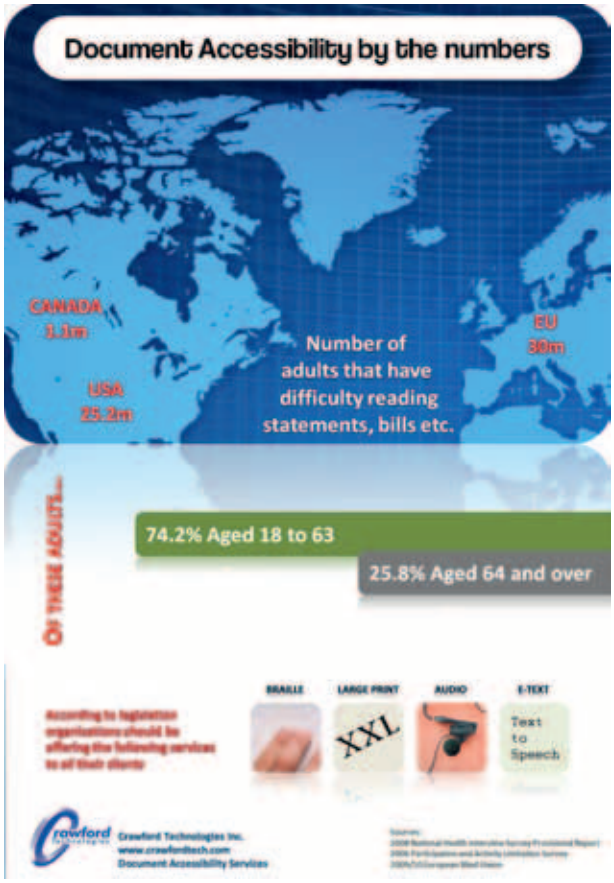
Who needs accessible documents?

There are many reasons why individuals may request an accessible document. Major vision limitations include diabetes, clinical blindness, old age and reading difficulties (e.g. dyslexia). Finding information about how many individuals need accessible documents is a challenge itself.

Our 'Document Accessibility by the numbers' infographic details some of the regional facts, but your company's customer needs will be determined by the types of product or services you supply and the demographic of your customers.

Research in CrawfordTech's division for accessible document creation and mailing services shows the following pattern: 6% of 'alternate format clients' request Braille, 92% request large print, 0.5% request an eText document and 1.5% request an electronic audio file.

CrawfordTech's DAS division currently provides accessible document creation and mailing services, so we are in a position to discuss high level usage patterns and share the trends we are seeing.



Organizations do not really know how PDF documents are currently being used for 'read aloud' or text-reflowing purposes – they simply can't tell how users are interacting with their PDFs – but due to the fact that tagged PDF is not broadly adopted, our opinion is that PDFs are not yet widely used for this purpose.
It is important to note, however, that interest in this capability has increased significantly in 2010.

Why should I take the time and effort to ensure my documents are accessible?
Some organizations have taken the time and effort to make their facilities, systems and documents accessible. In fact, some companies have provided accessible documents for nearly 20 years. For them, it was always 'the right thing to do'.

Other organizations are only now becoming interested in making their documents accessible due to lobbying and awareness efforts, new and pending legislation, customer satisfaction and the significant costs associated with class-action law suits and litigation.

The bottom line? Leading organizations are already making their transactional customer communications accessible and they are following industry trends and standards, such as PDF/UA, for their output. Those who have not yet implemented their document accessibility programs are looking for leadership from standards bodies, such as the PDF/A Competence Center, to help them.

Approaches to Creating Accessible Transactional Documents
There are two approaches to creating accessible documents for your clients:

Approach 1 – Do it yourself
Just as it sounds, this approach means acquiring the software and equipment required and integrating it into your current systems and workflows. It also means developing or hiring accessibility expertise within your organization to maintain and enhance your document accessibility deployment strategy for transactional documents. This strategy will evolve as standards such as PDF/UA are published, as other technology trends impact accessibility and as demographic shifts impact accessible document usage.

We predict that some very large organizations will wish to adopt this approach to transactional document accessibility to maintain control of the diverse needs of their organization.

Approach 2 – Outsource
As with most outsource, SaaS (Software as a Service) or Cloud computing approaches, this means finding a vendor that is responsible for being current on all of the trends, standards and developments in accessible transactional documents. You will then work with that organization as both a service provider and consultant to produce all of the electronic and physical documents in accessible formats based upon your specific industry vertical and customer demographics needs.

We predict that most organizations will utilize this type of approach due to its cost-effective nature.

Why is PDF/UA so important for these types of documents?
It is important to restate that PDF/UA is built upon the strong foundation of PDF/A. A PDF/UA document can also be a conforming PDF/A document. This means that

a properly prepared PDF/UA document will meet all of the existing PDF document needs of this industry for accuracy, retention, longevity and the other key PDF/A benefits.

It is critically important to have a PDF standard for accessible documents:

- Users of an accessible PDF can be assured that all software and devices they use will be compatible and will utilize all of the accessible features to meet their specific needs. Let's face it; these individuals have enough challenges getting access to the wide variety of documents that surround them today. They want their accessibility tools to work reliably and consistently. Having problems with their accessibility technology is an unnecessary frustration that they shouldn't have to endure.

- Providers of accessibility technology and assistive devices will only need to implement their particular features against this PDF/UA standard. This is a relatively small community of vendors and standards are highly important to them. This standard will help them keep their development and support costs low, improving their profitability and ongoing viability. In turn it will also ensure that their products are available at reasonable prices for their customers.

The publication of this single, defined accessibility standard for PDF files will be an important development for transactional document producers, accessibility solution vendors and for individuals that need document accessibility.

In Conclusion

PDF/UA is a welcome standard for organizations that wish to be compliant and serve all of their customers equally and fairly. It is an important cornerstone for organizations that have to serve the needs of their clients and the wide variety of their accessibility needs. It is a welcome and consistent standard for the visually disabled community, the assistive technology they use and the vendors that support them.

Accessibility – What PDF/A-1a Really Means

Duff Johnson,

CEO Appligent Document Solutions

Executive Summary

PDF/A-1a is the higher of the two conformance levels for PDF/A. This article explains that the "a" stands for "accessible", and provides an overview of the end-user, business, regulatory and operational significance of conformance level "a". Finally, we introduce PDF/UA, the forthcoming International Standard for accessible PDF.

The role of accessibility in PDF/A

In HTML, accessibility is simple. The content and the logical structure organizing paragraphs and headings and images into a document are seamless.

PDF is a different world; a world of objects, coordinate references, dictionaries and content streams. Deep within the core technology of PDF, the characters, words and paragraphs and pages so clearly evident to the visually-oriented reader have no logical connection to each other at all.

PDF was originally designed to provide multiplatform fidelity on screen and in print, where the only objective was painting a picture. There's no such thing as a "paragraph" or even a "word". Runs of text are known as "TJ operators". TJs appear in a sequence suiting the software that produced the PDF. Don't confuse words with TJ operators; that's like confusing a sentence with the movements of the print-head used to physically print that sentence.

It may seem too obvious, as it were, for words. It's not. The order in which content-streams occur in the PDF file, commonly referred to as the "reading order", is something of a misnomer. In this context, "reading order" actually refers to the order in which a computer reads the file's contents. Humans, by contrast, read in "logical order". The two often appear similar but should never be confused with each other. Unfortunately, many developers have interpreted typesetting arrangements as equivalent to logical order, with disastrous results.

As of 1999, PDFs could be made accessible through "tags" – the addition of logical ordering structures (headings, lists, tables, footnotes, form fields, etc.) to document content (text, images).

Tagged PDF makes PDF/A-1a possible, because tags are the mechanism for expressing logical document-structuring concepts in PDF files. Since tags organize the non-visual means of accessing content on the page, correct tagging is essential to the intent of PDF/A-1a.

Given the history of PDF and the way most PDFs are built, achieving logical in addition to visual reproducibility is a substantial challenge. Requiring a reproducible visual appearance over the long-term is profoundly different from requiring that the same document's contents be accessible. The two conformance levels of PDF/A exist to allow for both.

Who needs accessibility?

Conventionally, the typical consumer of assistive technology (AT) is a blind person equipped with a computerized braille reader or "screen reader" software. Their chosen AT device provides text-to-speech, keyboard interaction or other features to make computers usable to those without sight. There are many disabilities, however, and a correspondingly wide variety of assistive technology devices, both software and hardware, are available to enable disabled individuals to read and interact with web-pages, forms and electronic documents.

Governments are increasingly requiring their agencies and contractors to deliver accessible products and services. From websites to forms, regulations, product manuals and reports, documents in the US Federal government must comply with Section 508 accessibility regulations, in effect since 2001. Several state governments have similar laws, as do governments in Canada,

various EU member states, Australia and elsewhere. A number of organizations, including the retailer Target, have been found liable, with significant monetary damages, for their failure to provide equal access to content.

That said, accessibility isn't just about the needs of disabled users. Human beings are not the only "consumers" of electronic content; search and indexing engines are also "readers" of PDF files. There are several conventional business and operational reasons to ensure PDF files are tagged to high standards and thus achieve meaningful PDF/A-1a compliance.

Accessibility benefits every user

Blind users are prominent in calling for content accessibility; but the technology that makes documents readable by blind users is directly applicable to the mainstream business needs of civil servants, attorneys, archivists and others considering PDF/A. Properly tagged PDF files offer a series of functional effects with significant benefits for users of archival material.

After all, while visually reproducible pages are, obviously, critical, if you can't find the document in the first place because it's not tagged correctly, reproducibility becomes somewhat moot.

The key advantages of accessibility for the institutional or business archivist are:

- Searchability, because logical ordering of content ensures that words and phrases are made available to the search engine irrespective of page position, print order, or other, non-semantic factors. Additionally, well-tagged PDFs include alternate text for each semantically significant image, providing additional content to search engines.

- Search Engine Optimization (SEO), because tagging-aware search engines understand the logical structure elements (such as headings) in tagged PDF and can use them in their metrics.

- Content extraction (assuming your preferred PDF viewer is aware of PDF tags) is enhanced at two levels. First and foremost, proper tagging ensures that text is selected and extracted in the correct logical order. It's not OK to have page header text interrupting a sentence, or to mix up columns in a multiple-column document. Secondly, proper tagging ensures that complex logical structures such as tables may be exported to spreadsheets without error, while document text may be exported with key structural information such as headings and lists intact.

Of course, to gain the benefits of tagged PDF, your PDF software must process PDF tags!

Why PDF/A-1a is insufficient

In a PDF, just as in HTML, you must use as many tags as are required to correctly convey the logical structure of the content. Each paragraph, for example, needs a <P> tag. Headings get tags such as <H1> and <H2>, while lists consist of <LI> tags nested within an <L> tag. Tables (minimally) consist of a collection of <TR>, <TH> and <TD> elements grouped into a set of <TR> tags, themselves contained within a <Table> tag. There are many other such rules for tags, tag attributes, artifacts, images, languages, fonts and so on.

A full description of what accessibility means for PDF files was unavailable when PDF/A was first developed between 2001 and 2005. For this reason, PDF/A-1a offers only the broadest outlines of what's required for accessible PDF. Technically, it's possible to comply with PDF/A-1a using a single tag for each page, irrespective of the document's contents. That's the key reason why claims of conformance or validation of PDF/A-1a are, by themselves, essentially meaningless.

What's lacking from the standard is a technical description of PDF/A-1a's true intent; the preservation of not only a visually reproducible document, but an accessible one as well. This description is the subject of another ISO Standard – PDF/UA – which we'll discuss shortly.

Existing concepts of accessibility

From IBM's GML and SGML through to HTML and XML the need to mark up text with structure has led a steady march towards a more or less universally comprehensible, and thus accessible, set of concepts.

Large-scale authoring of structured content began with the birth of the Internet and the associated explosion in the use of HTML. NIMAS and DAISY provided important options for published materials, but not all content is formally published. To establish accessibility guidelines and to provide a baseline standard for consistent delivery of logical structure in web pages, the W3C's web Acces-

sibility Initiative published the first Web Content Accessibility Guidelines (WCAG) 1.0 in 1999. WCAG 1.0 has since been replaced by a far more advanced, less HTML-specific document, WCAG 2.0, in 2008.

The Federal regulations known as Section 508 have been in force since 2001. More recently, compliance has improved across most Federal agencies, with new websites and documents undergoing at least cursory examination for Section 508 compliance. While large volumes of content remain unvalidated, the trend is for new documents to be either created accessible or made accessible prior to release.

In late 2004, while PDF/A was preparing for its debut as ISO 19005-1, the industry's main standards development organization was gearing up an ambitious effort to produce an international standard for PDF accessibility: PDF/UA.

Introducing PDF/UA

Since PDF is a format for any document, not just published content, NIMAS and DAISY are fundamentally inapplicable. WCAG 1.0 was specific to HTML, and Section 508 leaves much to be desired. WCAG 2.0, while generally technology agnostic, doesn't specify technical requirements for accessible PDF files. Just as with PDF/X, PDF/A and then PDF/E, a new PDF standard was required to describe accessible PDF in technically complete terms.

Recognizing this need, AIIM, the ANSI-accredited organization leading electronic document standards development and education in the US, initiated the PDF/UA (Universal Accessibility) standards committee in 2004. The objective of PDF/UA: to set clear normative standards for developers seeking to create, manipulate or read accessible PDF files.

In 2009, PDF/UA became ISO/AWI 14289, a candidate International Standard. As of August, 2010, the document is a Committee Draft, with hopes to publish in 2011. Alongside the Standard itself, the Committee plans to publish an authoritative ***Developer's Guide to PDF/UA***, explaining core concepts for software developers, as well as ***Best Practices for PDF/UA***, a guide to tagging PDF files for end-users.

Creating PDF/A-1a (accessible) PDFs

The key thing to understand is that a really good PDF/A-1a file is one that also complies with PDF/UA.

Creating accessible PDF automatically directly from an authoring application is possible, but first and foremost requires the PDF creation software to be capable of generating PDF tags. A wide variety of applications, from Adobe Acrobat's plugin to Microsoft Word to Adobe's InDesign and FrameMaker, as well as free applications such as Open Office, can create tagged PDF.

However, it's not enough to simply use the right software and push the right buttons. Tags must correctly represent the logical structure of the document. Ensuring tags are correctly applied requires strict guidelines governing document authoring, layout and production. Styles must be appropriately named and/or role-mapped, and then employed consistently and correctly. Table structure must be well-considered and implemented; images need alternate text; heading tags should descend from H1 to H2 and H3 without skipping, and so on.

Manual validation work may be minimized or eliminated through authoring practices that are sensitive to accessibility requirements. Absent careful, accessibility-oriented authoring, alternate text for images, complex layouts, tables and forms will require human validation into the foreseeable future.

In principle, structured documents are just better. Teach the authors how to write documents with an eye for the concerns of accessibility and the problem is solved in the most cost-effective possible way.

Never

- Use color or contrast alone to indicate meaning
- Use design to convey meaning in a way that can't be expressed though the document's text

Avoid

- Spanning table cells, complex table structures in general
- Table tags without tabular data
- Illustrations comprised of many small vector graphics
- Overlapping elements
- Background images

Always

- Prefer simpler layouts
- Address the need for alternate text for graphics early in the authoring process

Making untagged PDFs accessible

PDFs can be created from any software that can print, one important reason why PDF is so successful. However, the ease of PDF creation poses a special challenge in terms of accessibility because today, most PDF creation software can't create a tagged PDF.

For this reason, most PDFs are untagged, and most tagged PDFs are unvalidated. If you're trying to achieve high quality PDF/A-1a conforming files from existing untagged PDFs, the only mainstream software currently capable of editing tags in a PDF file is Adobe's Acrobat Professional.

Acrobat includes automation triggered by the "Add Tags" function in the Advanced → Accessibility menu (Acrobat Professional 9). This feature scans the PDF and builds a tag tree for the document. You can get lucky on the simplest files, but on more complex content, the Add Tags function invariably makes mistakes, and results must always be checked.

If the software which created the PDF was relatively well-behaved, simple documents may require very few corrections. With more complex page layouts, especially when tables, multiple columns and graphics are involved, the more difficult it is to check and correct a tagged PDF to ensure accessibility.

The ten commandments at the 'core' of a PDF Accessibility Best Practices Workflow are as follows:

1. Identify and resolve low contrast and color-used-as-content situations
2. (If a scanned document) OCR and correct the output. OCR errors are not permitted in accessible documents.
3. Add hyperlinks as required, or check existing links for validity
4. Run "Add Tags" in Adobe Acrobat Professional, or other of PDF tagging software
5. Check and correct tag order (text and graphics tagged in correct logical order, artifacts marked)
6. Check and correct heading, list, and table structures and language attributes
7. Add alternative text to image tags
8. Ensure file metadata is correct and the document's language property is set
9. Add bookmarks (outlines) if the document is longer than ten or so pages
10. Quality control, optimize and deliver

How to Validate PDF/A-1a

Semantic content is the material of value, the significant text and graphics conveying the meaning of a document. Non-semantic content includes repeating page headers and numbers, image borders, lines separating columns and so on. These are artifacts of design and the layout, and must be marked as such rather than tagged so as not to interfere with the logical flow of the document.

The three basic questions when validating for PDF/A-1a conformance are:

1. Is all semantic content tagged in correct logical order?
2. Are the headings, lists and tables and other tags in the tags tree correctly structured?
3. Is the non-semantic content marked as artifact?

The answer to all three questions must be "yes".

At present, there is no way to automatically validate conformance with PDF/A-1a. Automated checkers are important, but can offer little more than verify tags are present, language is specified, that images include alternate text, and similar, limited validation functions. In many cases, validating the logical order, table or list structures is still a job for a human.

Conclusion

To ensure PDF/A-1a conformance is meaningful rather than notional, it is necessary to ensure the file's contents are accessible. PDF/UA will provide a clear set of file-format requirements to flesh out the details of conformance with the spirit of PDF/A-1a.

A PDF/A-1a conforming file is not just visually reproducible, the content is reproducible as well. As servers world-wide become into vast silos of PDF and other content, PDF/A-1a files will offer a key benchmark for the long-term preservation of document structure, ensuring high-quality search, reference and reuse for the lifetime of the file.

XMP Metadata Primer

Olaf Drümmer,

callas software, Managing Director; Board, PDF/A Competence Center

Extensible Metadata Platform (XMP) is a metadata framework developed around 2001 by Adobe, and the first version of its specification was published in January 2004. With one of its strengths being its seamless integration into various file formats, it has since conquered quite a number of user communities. While photographers still make use of EXIF (Exchangeable Image File Format, developed by digital camera makers) or IPTC (more strictly speaking IPTC IIM, or IPTC Information Interchange Model, where IPTC stands for the International Press Telecommunications Council), XMP is slowly taking its place as it is the more modern format, supporting for example Unicode. IPTC IIM itself is now more often used in its XMP flavour for the majority of digital photographs, rather than in its original binary format.

Also supporting the world of PDF files – the first incarnation saw the light as XMP support in Adobe Acrobat 5 – its flexibility made it an obvious choice for PDF/A: it appeared very well suited to keep track of metadata in the same reliable and structured way as PDF can keep track of the visual and semantic content of an e-paper document.

While PDF/A itself has turned into a huge success especially in European countries, the metadata aspect of PDF/A in the form of embedded XMP metadata seems to be more difficult to understand and to put to use than PDF/A itself.

XMP Metadata – the hidden gem

One of the problems with metadata in general, but more specifically so for XMP, is the fact that it is not looked at very often. In Adobe's Acrobat family of products, for example, it takes a number of steps to open the right dialogue to look at metadata. In addition, real world implementations that display XMP fail to attract users other than die hard engineers, and the presentation of the data and its structure is rawer then fresh sashimi. While – surprisingly enough – the majority of PDF files (and per requirement of the PDF/A standard all PDF/A files) do contain ample metadata, it is typically difficult to find the data fields of interest amid a number of computer generated sequences of seemingly arbitrary bytes.

At the same time, some vendors took it too far. They decided to hide the XMP-ness of metadata from the user and to present only those metadata fields they deemed worthy to meet the eye of the beholder. This took away all of XMP's flexibility, and even the tiniest need to also access custom metadata turned out to be difficult to nurture.

Intended as a means of counter balance, user interface definition languages to define custom views for custom metadata fields were invented, but their development was not for the faint hearted. Deploying them – and maintaining updates for them – was not easy at all, not even inside an organisation. And then, once those XMP custom panels were becoming more common and accepted by users interested in XMP, they were replaced by the next, even more potent user interface definition language, widely known as Flash. While it is unfair to always blame those who contributed something (most of the time it was Adobe developing XMP tools, and giving a lot of them away free of charge) rather than those who didn't bother to contribute anything at all, it has to be said that even the user interface implementations from the inventor and one of the most active supporters of XMP still have not achieved a state that lives up to the promise and potential of XMP.

Last but not least, the tools to manipulate metadata, or to interact with metadata beyond just reading it, often lack any degree of refinement: while everything seems possible, nothing turns out to be straightforward. There are free tools like the EXIFtool (whose XMP support is as good as that for EXIF) but the better they get the more it is likely that one needs to master the command line or shell scripting to be able to fully take advantage of them.

As a consequence, even users who tried to have a look at metadata for their files in numerous cases developed

tactics to avoid looking at metadata. This is a pity, since there is no more powerful and more flexible metadata framework around today than XMP. It just needs to be vivified (quite) a bit more by making looking at it a more pleasant and also a more useful experience.

XMP is not a list

The most basic approach to keeping track of metadata is to just have names for data fields and enter simple, unstructured information items into them: e.g. a number for the size of a book or the year of publication, a small piece of text for a description and so forth. This approach is already a very powerful one, fairly robust, and also easy to implement with regards to both storing the information in databases (or spreadsheets, or text files, or…) as well as presenting it to a user for reading or changing it.

It is however a relatively limited approach when it comes to slightly more advanced types of informational items, like a list of the authors of a publication, versions of the same piece of text in several languages, or the dimensions of a rectangle. Or for the more adventurous, ordered lists of data structures, like document change events, which might include the tool used, the type of modification applied and a time stamp.

XMP can do all of that, and probably more. Rooted in RDF (Resource Description Framework) and thus in XML, it has inherited most of XML's flexibility while not being bound by some of its limitations. In XMP it is no problem to use standard metadata fields alongside custom metadata fields. It also fully acceptable to use and combine individual entries from any number of XMP schemas – just as may be necessary on a case by case basis.

What is a blessing to some can easily turn into a curse for others. Some have abused the flexibility of XMP, changing their mind about what goes where every other day. This of course kills interoperability – one of the strongest parts of XMP when it's done correctly. Once it comes to extracting XMP from a file and storing it somewhere, e.g. for tracking an ingested document in the database of a DMS system, just having yet another table in the SQL database won't do. XMP can be way more complex than this, and the database design needs to reflect this. Similarly, the user interface for displaying and editing XMP needs to match XMP's structural flexibility. This can impose a substantial burden on developers, unless they decide to call it a day early on and try to get away with "structure by tabbed text" designs.

Today's world needs more powerful metadata structuring than lists of key value pairs, and someone will have to get the job of providing the right tools done. Those who care about metadata should make a review of metadata capabilities an integral part of their decision process when buying new tools or solutions.

XMP in PDF/A – even more complex than just XMP?

The PDF/A standard requires that any document metadata must be recorded inside PDF/A files in the form of XMP. But not only that: if some of the metadata fields are custom metadata fields (fields not yet specified in the original XMP specification), their syntax and meaning have to be documented inside the file's XMP metadata – using XMP.

While this looks like an unnecessary exercise to some, it actually follows the spirit of PDF/A. Whatever is inside a PDF/A file shall be self-contained and it must be possible, with reasonable effort, to retrieve the contents and their meaning in reasonable quality. The only prerequisites necessary to achieve this are defined in the PDF/A standard and the underlying specifications, like the PDF specification or the XMP specification. Data fields not specified in the XMP specification are considered custom metadata fields, and the meaning of such custom fields cannot be known unless some documentation is provided. As typically no repository exists that captures all custom fields ever used (and by whom), it happens to be a smart approach to put the documentation next to the data fields they describe.

He who will reap must sow

The previous paragraphs may give the impression that dealing with metadata is mostly painful and rarely ever rewarding. It must be admitted that a substantial part of the revenue from investing in metadata (and its careful storage in archived documents) may only become obvious in a number of years. Nevertheless, some advantages of its proper use may provide a return of investment relatively soon. It is important though to develop a good understanding of what is needed – what will the metadata be used for – and what the right tools are to implement the usage of metadata. Where metadata are to be archived inside PDF/A files, they will usually have a history of their own during the document's life cycle before being archived. Making (more) active use of a document's metadata while the document is created and used will increase the likelihood of good quality metadata worth archiving without a lot of headache.

The Knowledge Map: How to Take Advantage of XMP Metadata

Manuel Brunner,

Head of Projects and Services, Intrafind AG

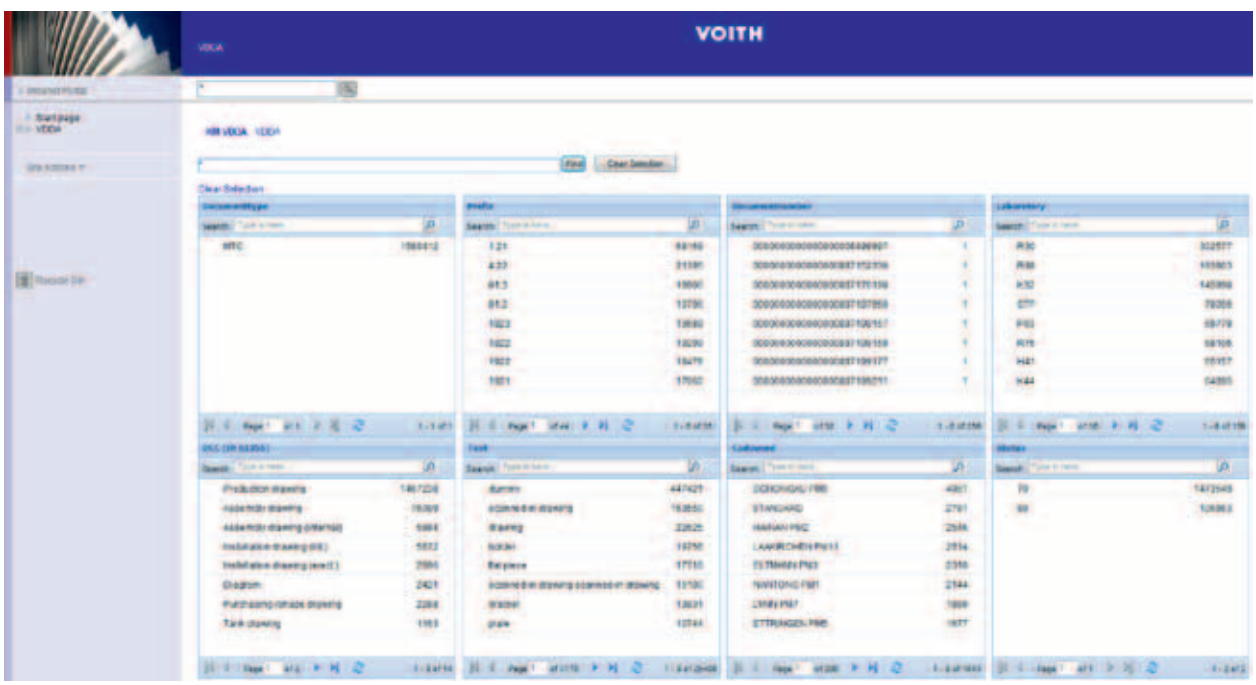
Adobe Systems Inc. is using the slogan "Adding intelligence to media" for their XMP standard. The question which will be discussed in this article is: "How can the metadata in XMP be reused in an easy way, and how can a corporation take advantage of this information?"

The first and most obvious way to use XMP data is in an information management system. This was also the initial idea of Voith Group when they began a project to replace their host systems via conversion to PDF/A, and save the PDF/A files in a document management system.

There was much emphasis placed on the urgent and contemporary removal of old host systems which had been spread over various sites of the Voith group.

In the course of the removal, all data and information stored on those host systems should be safely and cost-efficiently archived on new systems. Furthermore, the Voith employees should be able to intuitively access the information afterwards by using a new unique search solution. The introduction of the new solution also was intended to reduce the high running costs for the maintenance of the host systems.

After an evaluation of various document management system providers, the Voith IT Solutions GmbH & Co. KG in St. Pölten (Austria), the responsible Competence Center for archiving standards within the Voith Group, decided for a totally different approach to pre-



The Knowledge map at Voith using XMP Metadata for navigation

senting the host system content to their engineers. They decided for an enterprise search product, the iFinder from IntraFind AG.

But why use search technology instead of a DMS or ECM system for this task?

The answer is quite simple. The main goal of the host system replacement project was to make the data accessible and to find information in an easy and intuitive way, and not to manage the information. After the responsible managers of the IT Competence Center for archiving in St. Pölten attended at an international SharePoint Convention where the "Knowledge Map" idea was presented by Intrafind, a mind change started and the whole project deliveries were challenged again.

A simple way to get manageable content out of the host systems – use XMP metadata.

In the initial project, Voith and IntraFind indexed existing archive data which was stored on an old host system in St. Pölten and exported the required documents to PDF/A format. In addition to the full-text, the metadata of the documents, which is very important for the work of the Voith engineers, was extracted by a PDF/A conversion tool into XMP metadata. Then the information from the full-text and the XMP metadata of the PDF documents (e.g. order ID, customer ID, machine information etc.) was added to the search index of the iFinder.

In cooperation with Voith, IntraFind developed a user interface which significantly improves the search process. At a glance the user has access to all relevant information in terms of metadata (e.g. author, file format, creation date of a document, project-specific metadata fields). The result list can be quickly filtered by mouse click and can be visualised in terms of a so called "Knowledge Map". This way, the user gets a 360 degree view of "his enterprise knowledge" which is based on his individual user rights and can be intuitively used without any training.

Search like in an online shop

For the visualisation of the Knowledge Map, IntraFind designed a user interface that the user is already familiar with from the Internet. It can be compared to the navigation of leading modern online shops. In practice, this means that every Voith employee can limit the enormous amount of enterprise data via setting filters by mouse click until a manageable number of results is displayed. Then the search is activated and a list is created containing only the remaining results based on the filter criteria.

The following example demonstrates how the Knowledge Map works in practice:

A Voith engineer searches for information about a certain customer from the year 1994 and therefore clicks the customer's name in the Knowledge Map. With a second mouse click he limits the data set only to information from the year 1994. After that he gets a result list which contains all information about relevant projects affecting this customer, about items used in the course of the customer projects and even about the coat of paint of the individual components of the produced machines. With another mouse click he can then limit the data set to less than ten hits and gets the result list in a classical view which he is used to from the common Internet search engines or in list view with option to transfer these data into an excel sheet for further proceeding.

To accomplish this, it is important that the classical search entry field is combined with the functionality and possibilities of visualisation provided by the Knowledge Map.

The basic concept of the Knowledge Map: Find without search

As an alternative to the exclusive selection of filters, the Voith employee can also enter a term in the search entry field of the full-text search and combine this search with the filter elements of the Knowledge Map. Later on, he can remove each selected metadata from his search (e.g. the restriction to a certain period of time or author) and consequently enlarge the hit set again.

The difference with this solution in comparison to a classical database application is clearly obvious: by using the possibility to set filters, the error rate of the search process (e.g. typos of the user when entering the search term) is explicitly minimized. Furthermore, the Knowledge Map can be quickly implemented, is easy to manage and very high-performance.

The concept behind the Knowledge Map consists of the quick retrieval of already known information as well as in the possibility of getting a quick overview of all existing enterprise information about projects or products via just one aggregated access point. This supports users browsing through enterprise data.

All companies which already had the Knowledge Map in use had basically the same requirements for their new search solution:

- ***One*** search for the entire enterprise knowledge as well as search windows in their application or a dedicated set of documents.

- An intuitive handling via guided search which does not require any previous knowledge about how to use complex Boolean operators. Furthermore, the different search technologies in databases, intranet or document management systems should be harmonised.

- "Without the need to permanently reinvent the wheel": new colleagues should also be able to get quick and easy access to existing (old) information in order to avoid a repetition of previously made experiences (which are often very painful for the company) in their current projects.

Positive balance of the search project

The vision of Dipl. Ing. Erich Seher, managing director of the Voith IT Solutions GmbH Co. KG in St. Pölten to introduce an overall enterprise solution evoked very positive feedback among the Voith employees. Consequently, the initial project has been extended to the indexation of further host systems (extracted metadata + PDF documents), SharePoint data and file server data now. For the next step it is planned to index mailboxes and to connect SAP data and documents.

The benefit of the new solution is enormous, especially because the existing structural information is always integrated into the search.

During a search in the file system, for example, the folder where the hit documents are stored in is always displayed for fast filtering of huge result sets. Via browsing, like in the Windows Explorer, the user can quickly and easily find the relevant hit document and open it. This functionality complies with the search behaviour of the users who often remember that they have saved the required document to a certain folder or that it has been created by a certain author, but are unable to find it again without the help of an intelligent search.

In addition to the positive acceptance by the Voith employees, the innovative search via intelligent navigation is also financially a great success: in less than six months after shutdown of the old host systems the ROI in terms of the investment costs of the new overall system will be achieved.

This use case shows how to use essential information in a corporation, the XMP metadata, for a retrieval engine with enormous high usability and how to best meet customer, financial and legal requirements by using PDF/A documents.

Session Intro – Track A: Archives and Libraries

Session Chair:

Thomas Zellmann,
PDF/A Competence Center,
Managing Director

The articles in this chapter contain the content of the Conference Track A: Archives & Libraries. For archives and libraries, "long-term archiving" virtually means "forever"! Special measures are required for both hardware and software, and PDF/A can fulfil many of the requirements for software and file formats.

Hans-Joachim Hübner from the SRZ gives a very good introduction into "Relevance of PDF/A in Archives & Libraries – Digital Preservation". SRZ has been working for a long time in this sector, and he adds a lot of details and best practices from real world projects.

UBS as a large Swiss bank may not seem to belong in this track. But all large corporations have departments which are responsible for the historic company archive and also their own libraries. UBS had the requirement to archive their website for compliance and history purposes, and has chosen to do it as PDF/A files. UBS will now be able to answer questions in the future like: How did our website look 3 years ago?" or "Which special credit offer did we have in January?" for compliance and legal reasons.

Archives are typically receivers of documents, and from their point of view it would be optimal if the files they receive are already in PDF/A format. This can be achieved through proper communication and organisation with the suppliers. There are many PDF/A scanning solutions that can be applied when existing documents in an archive are converted into electronic format. For files that are delivered in other electronic formats, numerous tools are offered for converting to PDF/A. It may be desired to archive both the original format (e.g. Microsoft Word) as well as a PDF/A version of the document.

In general, libraries are both receivers and suppliers of documents. When existing files are converted into a new format, the digital original (e.g. uncompressed TIFF or JPEG2000 lossless) may still be retained. PDF/A can be created from the original file and serves as an ideal delivery format, enabling full-text search and supporting embedded metadata. With respect to digital documents, existing or newly delivered books can be converted to PDF/A. PDF/A can be used for archiving postgraduate work in university libraries as well, and would itself be an interesting topic for such work. There are an increasing number of recommendations and requirements for PDF/A in archives and libraries, for example with the German National Library.

Mrs Natascha Schumann from the German National Library and project leader of nestor, the German competence network for digital preservation, will present this long-term archiving initiative. nestor has been working together with the PDF/A Competence Center for several years and now we are happy to be an official cooperation partner of nestor.

I would like to thank the authors for their excellent contribution of articles which hopefully makes this chapter a worthwhile reading for you! If you want to speak with an international archive or library, please do not hesitate to contact the PDF/A Competence Center – we will be happy to connect you with them.

Relevance of PDF/A in Archives & Libraries/ Digital Preservation

Hans-Joachim Hübner,

Satz-Rechen-Zentrum Hartmann+ Heenemann GmbH & Co. KG

These days, many cultural institutions (scientific and public libraries as well as state, private and ecclesiastic archives) are digitizing valuable cultural assets such as books, prints and maps. Along with the aim of enabling a broad public or scientific use or to protect valuable originals from direct access, this process is used in order to preserve the historic originals and to securely store them in optimal environmental conditions.

In addition, the approach is to digitze these originals in a high quality and resolution, or, in the best case, the highest quality and resolution according to the current technical state. In Germany, this means that, in accordance with the regulations of the German Research Foundation (Deutsche Forschungsgemeinschaft (DFG) for the retrospective digitization, black and white originals must be scanned at a minimum resolution of 600 ppi and gray scale and colour originals must be scanned at a minimum resolution of 300 ppi.

Particularly in the case of very valuable originals, we must attempt to reach the highest possible resolution that corresponds to the technical state and therefore be able to offer a very broad range of usage options. As an example, we will use the Beethoven-Haus in Bonn. When digitizing documents on site, SRZ used a scanner that was of a particularly high-quality and high resolution. In accordance with the size of the original document, the inclusion head uses all of the available resolution capacity of the camera in relation to the size of the original. The digital masters that result from this were saved as uncompressed TIFF and may be several hundred megabytes in size. They are used as the source format for derivatives for various applications and solutions such as printouts and web display.

A lot of different information...

But what type of data is created during the digitization of documents from libraries and archives? Put simply, the initial type of data created is image data that you want to display in the social context, and you want to provide the users with data about the development, temporal and contextual relationships, creators and current location.

In addition to image data, a whole range of other information is therefore collated and gathered. This begins with the bibliographic metadata, in other words the descriptive data for the document, such as the author or creator, the place and date of publication, publisher, printer, edition, etc.

Then there is the metadata that relates to contents and structure. This involves, for example, recording an existing abstract or creating an abstract. In addition, today, it is also common to use OCR to process all of the documents that are suitable for this and to save the results in their uncorrected form. This provides the basis for conducting a fuzzy search in the text and to highlight search results in the facsimile for presentation on the web or in other applications. However, this only works if the positions of the identified words on the page are also saved in the application.

Structural metadata are created, for example, by recording tables of contents and their link to the physical start of the chapters in the image data or other parts of the work, such as the register, register of places, register of people or graphics, volumes and so on. For this, you must assign the existing pagination in the work to the physical files and you must also specify structure elements and contents such as titles, headers and similar elements. The creation of structure elements may go all the way down to page areas such as margins, images or footnotes.

It is also normal to gather the technical metadata for the creation and the physical attributes of the digital representation in order to prove the history of the digital documents in this case. Metadata includes, for example, resolution, bit depth, compression, date recorded, the institution that gathers and owns the information, scan software, scanner hardware and similar information.

Today, all of these descriptive data that relates to contents and structure is gathered into a specific XML schema and saved. The schema that is used mostly across the world is, in this case, the Metadata Encoding and Transmission Standard (METS) for libraries or Encoded Archival Description (EAD) in the world of archiving (see ***http://www.loc.gov/standards/mets/*** and ***http://www.loc.gov/ead/***).

Various storage formats and true colours

Various compression methods are used for image data. Bitonal images are usually saved in the TIFF format that has been compressed to lossless fax group IV. TIFF is also used for digital masters in gray scale and colour and these are stored as lossless uncompressed files or as compressed in accordance with LZW. In the case of derivatives for different purposes, formats such as JPEG, GIF and PNG and various resolutions are used. The JPEG 2000 format, which was approved as an ISO standard at the start of this decade, is becoming more and more popular as a compression method that allows considerably higher compression with much higher quality than the traditional JPEG. A 'lossless' (compression without loss) variant is also available for JPEG 2000.

You don't just want to archive the colour photographs of valuable originals in the highest possible resolution, you also want to be able to reproduce the colours for the screen and printing in such a way that the human eye identifies it as the original. You can attain so-called colour fastness using the colour management with colour profiles. Using colour charts, in which the RGB or CMYK colours are stored as numerical values in a reference, the variances in the colour devices are calculated and the differences to the standard are saved. This then becomes the so-called colour profile. These differences are then attached to the associated image, either as part of the image or as an attached file. In each case, the specific differences of the output devices to the standard, such as printers and monitors, are determined and saved in the same way and their display can therefore be adjusted using the comparison to the reference.

A colourful mix

Let's first summarize which data is involved in the digitization and should be taken into account during long-term archiving:

- Digital masters, image data in high or the highest-possible quality, compressed or uncompressed without loss
- Colour profiles for high-quality colour photographs
- Derivatives of the digital master that are created for various uses, such as printing, web display etc.
- Descriptive, technical, content and structural metadata in various XML and/or text formats

For each long-term archiving of a library or archive unit, the data, saved under different formats, are combined to form a data-technical unit, for example, a TAR archive, and then saved to a suitable archive medium.

To check the integrity of the data at a later point, an additional checksum file (designed with a suitable checksum algorithm) is usually saved.

We are dealing with an information package that is obviously a quite complex entity, contains formats that differ greatly, must include two information units and cannot necessarily be read again by each TAR program, particularly in the world of Windows.

And what is the case with PDF/A?

In contrast to all of the formats mentioned above, PDF/A is completely disclosed and is a defined ISO standard that, as the first ISO standard, does not have any time restrictions. PDF/A is a normal PDF that can be opened and read properly using any program that can display PDF. PDF/A does not depend on any operating system, because PDF readers exist for almost every operating system environment.

How does PDF/A behave with the mixed bag of information from a digital representation, as described before?

- During the conversion, PDF/A does not touch image data at all and the image data retains its original quality, resolution and size and these can be restored at any time.
- PDF/A stipulates that information about the colours that are used must be saved and PDF/A is able to integrate created colour profiles.
- If you want, you can also integrate the created derivatives into the same PDF/A file. This can also be exported without being touched by PDF/A.

- PDF/A has two completely documented and disclosed areas for metadata. One is the fields for the document description (title, author, topic, keywords). The other is the area of XMP data that consists of XML data and offers the option to incorporate user-defined XML descriptions in this area. All of the XML schemas that are used in the library and archiving environment can be included here.

An additional advantage is that the full text that was obtained through OCR is not only also saved in PDF/A but it can even be placed behind the text in a searchable format, so that search hits can be highlighted in the facsimile in a way that is user-friendly. The verification of the data integrity can be directly included in the PDF/A file using digital signatures of various levels of conclusiveness (from simple to qualified) and is not separate information.

What the ISO committee says

- **Independent of any device or operating system:** Can be reliably displayed on various systems and devices
- **Self-contained:** Contains all of the components that are required to display the data
- **Self-documenting:** Contains descriptions for the integrated data
- **Freely accessible:** Does not contain any technical access protection
- **Open source:** Authorized format definition is completely available
- **Wide distribution:** Wide usage is perhaps the best protection for the readability of long-term archives (see ***http://www.aiim.org/documents/standards/19005-1_FAQ.pdf***).

Examples of use

My speech about retrospective digitization using PDF/A mentions some examples:

- The German National Library of Science and Technology University Library Hanover performed the retrospective digitization of the research reports that were supported by the Federal Ministry for Education and Research (in Germany) and the long-term preparation for this digitization.
- The library of the Swiss Federal Institute of Technology Zurich and the retrospective digitization of dissertations.
- The German Broadcasting Archive and various projects, such as the digitization of documents regarding television programmes from the former GDR:
  - Scripts from the program "Der Schwarze Kanal" (The Black Channel)
  - Broadcasting schedules for "Aktuelle Kamera" (Current Camera)
  - The program guide "FF Dabei"
  - Design drawings for the extensive pool of vehicles in "Sandmännchen" (Little Sandman)

Conclusion

All evidence suggests that the considerable advantages of PDF/A that usually exist in contrast to all other data formats for long-term archiving will lead to further distribution of this standard. More and more applications whose data needs to be securely archived for a long time use PDF/A.

Website Archiving to PDF/A – Customer Story: UBS

Rolf Günter,

Head of Business Development, Sales and Marketing, PDF Tools AG

UBS AG is implementing a ground-breaking project together with PDF Tools AG for archiving business-critical web pages in a manner compliant with auditing requirements. The result: enhanced security for documentation and review of communications contents. And the corporate archivist is very happy.

UBS is a leading global financial institution, the market leader in Switzerland in the private and corporate client sector, with offices in all of the major financial centers around the world and more than 60,000 employees in over 50 countries.

Electronic archiving of business-relevant documentation according to compliance directives has long been standard procedure at UBS. The company secretariat at the UBS Corporate Center is responsible for archiving management process documentation. There, all management process documents generated are saved in PDF/A, TIFF or JPG format and placed in long-term archives that meet auditing requirements – reports from executive management and the board of directors, policy documents, company founding documentation and all governance-relevant information.

UBS AG headquarter in Zurich.

Risk and reputation issues

Until recently, only the company's web pages were excluded from this archiving process. "As part of our regular archive review, we noted to our surprise that the UBS website was not included in the archiving," said IT project manager Daniel Spichty. There is a detailed disclaimer on UBS.com and on all country-specific pages that excludes legal claims by third parties based on statements made on the websites. However, different national laws also apply in this case. Thus it was decided that prevention is better than cure. Primarily for reasons of risk management and reputation rather than legally binding issues, the company decided to archive particular pages of the website as PDF/A in accordance with auditing requirements. Then there is the historical aspect as well. The corporate archivist, a historian by profession, is pleased to be able to add the web activities to his company history.

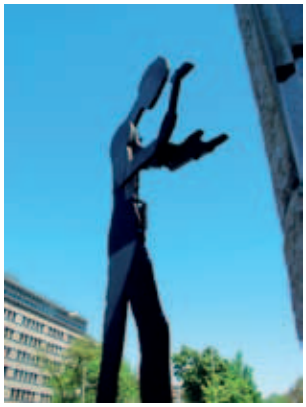
Daniel Spichty, IT project manager, UBS.

The content to be archived and what is technically possible were precisely defined in advance as part of a focused pre-selection process. Then the question of a suitable format arose. "We wanted to enable our employees to retrieve identified content in the archived pages via the URL or the date," explained Daniel Spichty – requirements that only the PDF/A format fulfills. With regard to long-term archiving, only TIFF, PDF/A and JPG make the shortlist of possible formats. Storing plain HTML files creates problems with displaying the files later, as particular browsers or operating systems are required for this. JPG or TIFF image files were not suitable due to their uncoded formats. In contrast, UBS can store the indexing

information such as URLs in the web pages when they are saved as PDF/A.

"Web page archiving is becoming a standard function!"

With PDF/A archiving, UBS can now definitively document which information was published on the web at a certain point in time, even 20 years from now. Furthermore, cases in which third parties take legal action against UBS based on information which was allegedly published on www.ubs.com and which UBS is not able to prove the opposite will now be impossible. For Daniel Spichty, it's only a matter of time until this becomes the standard. In the United States and the European Union, there are already pioneering judicial decisions according to which web pages must be treated just like other content with regard to legal obligations and archiving requirements.



"Hammering Man" by Jonathan Borofsky.

With regard to web content, all pages relevant to corporate governance will be filed in the long-term archives, including information about organizational structures, management staff, important corporate bodies and distribution of responsibilities as well as all information relating to investor relations such as share capital data, prices and the company in general. This also includes all link functions.

Software searches updated pages and automatically archives them as PDF/A

The solution has been in use since mid-2009; 150 documents were archived in the first six months. This takes place fully automatically and independently of the client. The document converter and an additional component which prepares content for archiving were installed as a preliminary process; these process the list of URLs to be archived on a daily basis on all UBS.com pages, including the country-specific pages worldwide. With each change from the archived version, the system automatically recognizes whether a new archive version needs to be generated or not. The sizes of the web pages to be archived are maintained one-to-one in PDF/A format. This creates a scrolling page in PDF/A, where the page numbering can be managed individually at a later point in time. The page is automatically saved in the UBS archive system, with full-text search functionality added.

Based on the experience of the web pages project, UBS has now introduced additional products from PDF Tools AG. Thus UBS also converts existing PDF documents to PDF/A format using technology from PDF Tools. "During the web project we saw that PDF Tools can do a lot more than we originally asked for and decided to expand our collaboration," remarked Daniel Spichty.

nestor – the German Network of Expertise for Digital Preservation

Natascha Schumann,

nestor – Kompetenznetzwerk Langzeitarchivierung

nestor is the German competence network for digital preservation. Its purpose is to bring together the existing know-how and the competencies with regard to digital preservation in Germany. Libraries, archives, museums and leading experts work together in the network to ensure the long-term preservation and accessibility of digital sources. After six years as a funded project, nestor transformed to a sustainable partner consortium in July 2009. Today, nestor is a cooperation association with 11 partners from different fields, all connected in some way with the subject of "digital preservation".

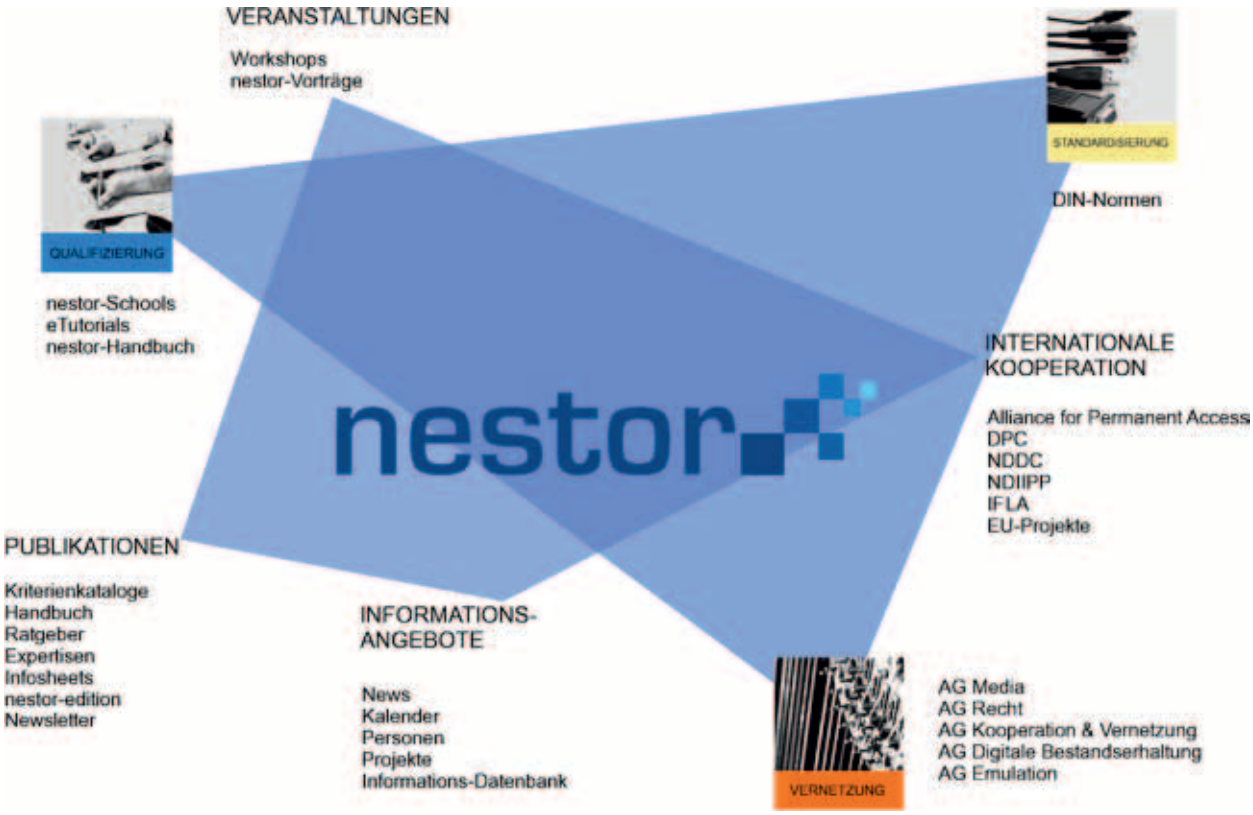
The partners are:

- Bayerische Staatsbibliothek (Bavarian State Library)
- Deutsche Nationalbibliothek (German National Library)
- Fernuniversität Hagen (Hagen Open University)
- Georg-August-Universität Göttingen / Niedersächsische Staats- und Universitätsbibliothek Göttingen (Georg-August University, Göttingen / Lower Saxony State and University Library, Göttingen)
- Humboldt-Universität zu Berlin (Humboldt University in Berlin)
- Landesarchiv Baden-Württemberg (Baden-Württemberg State Archive)
- Stiftung Preußischer Kulturbesitz / SMB – Institut für Museumsforschung (Prussian Cultural Heritage Foundation / SMB – Institute for Museum Research)
- Bibliotheksservice-Zentrum Baden-Württemberg (Baden-Württemberg Library Services Centre)
- Institut für Deutsche Sprache (German Language Institute)
- Computerspiele Museum Berlin (Computer Games Museum)
- Leibniz-Bibliotheksverbund Forschungsinformation Goportis (Leibniz Library Network for Research Information)

The nestor partners host five working groups, which are open to non-nestor members as well. The WG "Networking and Cooperation" provides a forum for identifying and addressing collective problems in digital preservation. The WG "Preservation of non-textual Media" gathers expertise and best practices from the area of AV and multimedia preservation. Legal experts of the WG "Legal Issues" points to passages of the copyright legislation which hinder digital preservation. The WG "Digital Preservation" discusses how the concept of significant properties can be pragmatically used in practical preservation processes. The just recently established WG "Emulation" acts as a network for all aspects concerning emulation and for exchange of best practices.

A group of eleven higher education institutions has signed a Memorandum of Understanding to work collaboratively and in association with nestor towards a digital preservation curriculum. They design several e-learning modules, for example an introduction to digital preservation, and modules on formats and data carriers, on metadata generation, and on web archiving. The university partners already deploy some of the developed modules in their academic teaching. The higher education partners also co-organise the yearly nestor summer school.

nestor provides an overview of existing standards in the field of digital preservation, bundles together standardisation activities and proposes new activities where



The nestor service portfolio.

required. User interests are represented at the national and international levels via the partnership with the DIN organisation. The standardization work takes place in the DIN Standardisation Committee for Libraries and Documentation (NABD), subcommittee Records Management and Preservation of Digital Information Objects (NABD 15). The subcommittee and nestor collaborate closely. The draft standard DIN 31644 – Kriterien für vertrauenswürdige digitale Langzeitarchive ("Criteria for trustworthy digital long-term archives"), based on the nestor-Catalogue of Criteria for Trusted Digital Repositories, has just been published.

nestor has been engaged with the topic of certification of trustworthy repositories for several years. Now progress on the European level is made: Following a series of European Commission sponsored workshops on audit and certification of trusted repositories, three European initiatives have signed a Memorandum of Understanding to harmonise their activities on the European level. The three initiatives are nestor/DIN NABD 15, the Data Seal of Approval (DSA), and ISO Repository Audit and Certification Working Group (RAC).

nestor also cooperates with institutions and initiatives in a European and international context. Even though the conditions may differ, there is a lively exchange and collaboration with the national coalitions of the Netherlands (NCDD), the UK (DPC), and the US (NDIIPP).

To provide a platform for sharing of knowledge and expertise, nestor organises workshops and events on different aspects of digital preservation and provides information and exchange about best practices. A publication series „nestor edition" is dedicated to scientific theses and scientific monographs with an explicit reference to digital preservation and analysis of new approaches.

Contact:
Natascha Schumann (n.schumann@d-nb.de)
nestor – Kompetenznetzwerk Langzeitarchivierung
www.langzeitarchivierung.de

Session Intro – Track B: Public Administration

Session Chair:

Bernd Wild,
Board, PDF/A Competence Center

Since 2006, the requirement for using the PDF/A ISO standard as a preferred document format has been increasing in proposals and projects of public authorities and administrations. There are even some national agencies that now explicitly require PDF/A as their standard archiving format. This follows the concept of relying on open standards, instead of company-specific formats which can lead to problems in handling the documents in the future.

A prominent example is in Denmark, where all governmental agencies are required to exclusively use ODF for working documents and PDF/A for archival documents beginning in 2011. A similar regulation was issued by the Ministry of Government of Norway, which relies on HTML, PDF, PDF/A and ODF. The French Direction Générale de la Modernisation de l'Etat regards PDF/A as the base format for archive documents.

In Germany, support for PDF/A is being demanded in most official proposals, as it conforms to the guidelines of the SAGA standard (open standards, open source implementations available) for public administration. And since 2008, Switzerland is also committed to using PDF/A for all documents which are exchanged between the authorities and the citizens, and thus have to be archived.

PDF/A is not only getting more and more popular in Europe, but also in countries like Dubai. A good example for adopting the ISO standard is presented by Sanat Kulkarni, who speaks about the use of PDF/A in the Road and Transport Authority of Dubai. There, the focus is on converting digitised paper into PDF/A-1 documents.

Another European country in the process of adopting PDF/A into their public administration is the Netherlands, which is following an approach of looking at the ISO standard as an integral part of their so-called records management. Dominique Hermans will give an idea on where the actual Dutch developments in the domain of records management are headed.

PDF/A at the Road and Transport Authority of Dubai

Sanat Kulkarni,

eDocuMAN Fz LLC

When the Road and Transport Authority (RTA) of Dubai was established in 2003, its main focus was to create and service a world-class infrastructure for Dubai. Handling the mammoth amount of information generated during the creation and servicing of this infrastructure was a big challenge.

The major problems faced after realization were:

1. Generation of more than 150,000 pages daily.
2. Categorization of the generated paper records.
3. Establishing a process for archiving 50 million pages of backlogged documents, which had to be finished within 18 months.
4. Setting up a day-to-day capture process for 150,000 pages. eDocuMAN provided a consultancy service as well as backlog conversion services.

Consultancy service

The following consultancy services were provided:

1. Explaining the difference between documents, records and archives.
2. Education on the importance of long-term electronic archiving.
3. Education on archiving paper information in PDF/A–1b format.
4. Establishing metadata (8-9 fields) for the different categories.

As part of the study it was determined that the 50 million pages would be divided into 168 categories, with each process being unique.

The document sizes ranged from B5 to A0, with paper thickness varying from as thin as onion skins to as thick as hard boards.

Backlog Services

The total manpower resources assigned were 89 for backlog conversion and 17 persons for day-to-day capture.

- Separate services were created using imaging libraries
- A process was defined covering reception to delivery of physical documents
- Synchronized physical document tracking vis-à-vis electronic documents was established

Conversion of the backlog has now been completed, with the data being archived in PDF/A-1b format for easy long-term retrieval and at the same time adhering to international standards.

PDF/A and Records Management in the Netherlands

Dominique Hermans,

Owner of DO Consultancy

As we all know, records management is "…the field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records." (Source: ISO 15489).

But how do we keep these records accessible and readable for the years to come? This is a difficult question not only for records managers, but also for governments as a whole. During the past couple of years PDF/A has become a well-known digital format within the archiving world in the Netherlands, but there are still a lot of grey areas.

Records management describes several important characteristics of a record. A record should be:

- **Authentic:** is the record what it is suppose to be
- **Reliable:** is it an accurate representation of the transaction
- **Incorruptible:** non-authorised changes aren't possible
- **Usable:** traceable, viewable and understandable within the original context
- **Performable:** data carriers should contain undamaged bit streams
- **Visually reproducible:** bits streams should be visualized correctly on the computer screen
- **Sustainable:** digital longevity of electronic files.

The Dutch Ministry of Education, Culture and Science has created a policy on how to keep digital records accessible, readable and usable for the future.

This policy includes the following requirements:

- The digital document format should be documented
- If possible the format should be an open standard
- Compression is not permitted – unless there is no information loss
- Encryption may not be used.

To raise awareness on the existence and use of open standards, an organisation was founded called NOiV (Nederland Open in Verbinding). They have published a document to assist the Dutch public body on the use of open standards and especially the use of PDF/A, PDF 1.7 and ODF (Open Document Format).

Within the Dutch government there is a "comply or explain" policy on the use of open standard software. Every public body in the Netherlands has to use open standard software when possible, but unfortunately not everyone is aware of this policy.

Two other organizations have been founded in the Netherlands for helping select the correct open standards: Forum Standardisation and Board of Standardisation. They have written articles on the use of PDF/A for example. Every standard has its own advantages and disadvantages, and a lot of users aren't able to "see the forest for the trees" anymore because of the number of different digital formats.

What digital format should/could be used, and when

When looking at the document lifecycle you can distinguish several stages:

- Document creation
- Collaboration (changes can and will be made)

- Document exchange (no changes)
- Publication
- Archiving

As can be seen in the diagram below, the NOiV recommends PDF/A as a digital format for archiving, but not for creation, collaboration or the exchange of documents.

	ODF	PDF/A-1	PDF 1.7
Creation			
Cooperation			
Exchange			
Publish			
Archive			

Green: completely usable
Orange: partly usable
Red: not usable

Source: Handreiking open documentstandaarden voor de overheid, NOiV (www.noiv.nl)

Before choosing the correct preservation file format, you should first determine which records ought to be kept for a long period of time. These should be records that are important due to their continuing administrative, informational, legal and historical value as evidence of the work of the creating organisation. The discussion about digital file formats is nothing new; it's part of an organisation's records management policy. But because of a lack of knowledge and awareness, or due to other priorities, these discussions have often been postponed.

Several questions still have to be answered before records managers can decide on the correct digital file format, like:

- How often are changes made to old(er) documents?
- What kind of digital format do the users within the organisation need?
- How often are digital documents shared with others?
- Which (kind of) documents should be archived?

One of the recommendations the Forum and Board of Standardisation has given is when to use PDF/A-1a and PDF/A-1b. PDF/A-1a can be used best for digital-born documents. On the other hand, the -1b version of the 19005-1 standard can be used for scanned documents and digital documents which cannot be converted to PDF/A-1a correctly.

Last but not least, the release of the upcoming standard PDF/A-2 will probably require the Dutch government to review their policies and seek advice from the PDF/A experts.

Session Intro – Track C: Business to Consumer

Session Chair:

Harald Grumser,
PDF/A Competence Center,
Chairman

In business-to-consumer relationships, the topic of PDF/A also crosses over to the consumer. When individual correspondence is used in communications between companies (or other organizations) and consumers, it is relevant for archiving. The correspondence is sent via e-mail or can be downloaded from a web portal, exposing the consumer to long-term archived documents. This usually concerns documents that are created individually for the end-user, such as invoices, quotations, order confirmations etc., and are more or less generated automatically in large quantities. In such a scenario, you must consider additional prerequisites not only on the side of the creator, but also consumer habits and options in this environment.

In the first presentation "PDF/VT – Overview and relation to PDF/A", Stewart Rogers from Crawford Technologies, UK describes PDF/VT. PDF/VT ("V" for "Variable" and "T" for "Transactional") defines a variable data printing (VDP) job exchange format and is a published ISO standard in the family of PDF standards. PDF/VT is the ISO 16612-2 (PDF/VT) standard on behalf of the variable and transactional printing industry.

Uwe Wächter, PDF product manager at SEAL Systems discusses "SAP and PDF/A – PDF/A in product life cycle"

What do ERP systems and PDF/A have in common? This question can be quickly answered: purchase vouchers, invoices and other business documents that are printed out of SAP must often be archived. This is an important field where PDF/A is of growing importance. If you use PDF/A for archiving, you will fulfil legal retention obligations and can guarantee that the archived documents will still be legible after many years. No special prerequisites are necessary for implementing PDF/A, and you won't have to forego on an elegant display or full-text search.

Harald Grumser from Compart AG concludes this section of the proceedings with an article entitled: "Optimising PDF/A documents for large archives" in which he reports on the characteristics of PDF/A in the environment of mass correspondence. Since the PDF/A standard does not allow you to work with external resources (as you would usually do in the area of high-volume printing) and since embedding resources such as fonts for each individual document may lead to an enormously inflated volume of storage, you must weigh up other techniques and approaches that allow you to work with PDF/A and its advantages.

SAP and PDF/A – PDF/A in Product Life Cycle

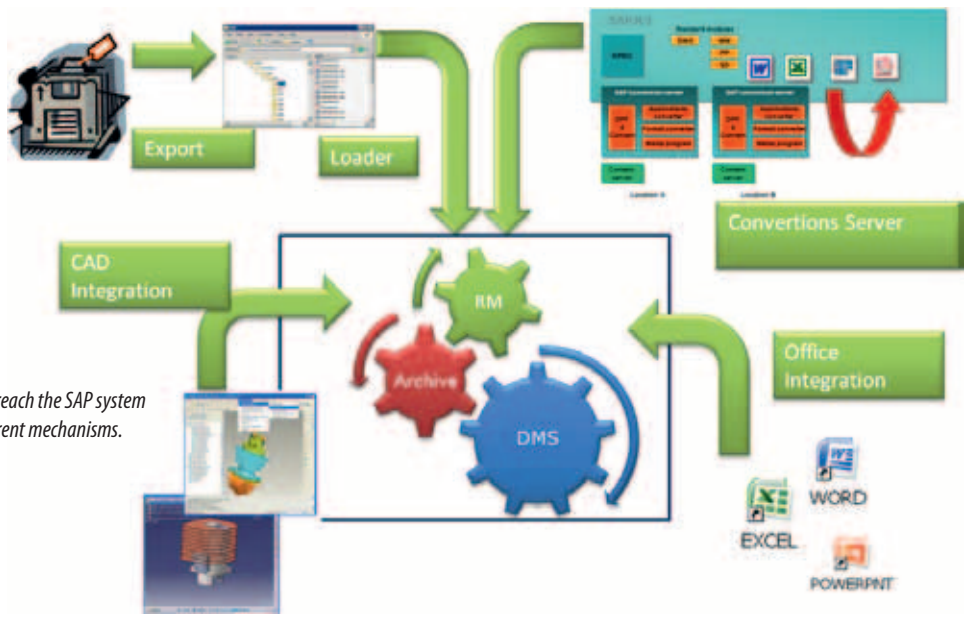
Dr. Uwe Wächter,

Product Manager PDF Technologies, SEAL Systems AG

With its integrated document management system, SAP enables a number of document-heavy business processes in industrial companies to be depicted. Documents are not only stored and found on the basis of their metadata – they are also directly linked with their usage in business processes. This procedure is based on practical experience, which tells us that the vast majority of documents in a company have a usage in ordering, production, maintenance, quality assurance etc. which is already defined when the document in question is created. In all cases, long-term reproducibility is required, so PDF/A recommends itself as the file format.

For the management of a company's documents, SAP offers an integrated document management system, an archive, and organizational functions via records management. Depending on the planned usage of a document, it is stored in one of these administrative structures and linked with the corresponding business objects. This keeps documents at the very heart of business processes.

The document management system DMS that is integrated into SAP has proved its worth for a large number of file types and document types. Due to the built-in version and status management and a large number of functions for authorization and classification, this system can be used in



Documents can reach the SAP system via various different mechanisms.

many application cases, for example, for product information and logistics information, quality data, manuals and operating instructions, catalogues and images.

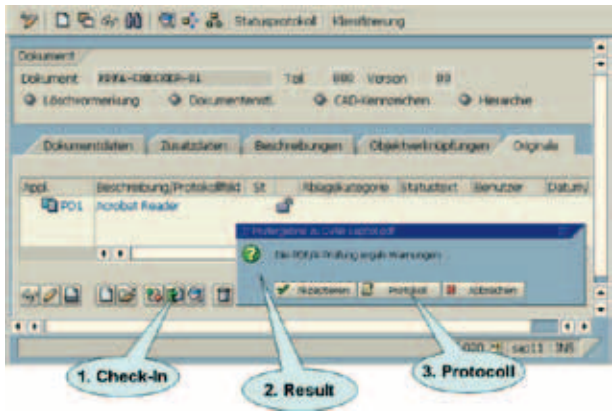
The documents that are managed using DMS can be used in a number of ways – for example, to display on screen or from printing and for electronic distribution. The joint preparation, together with other SAP forms for manufacturing, procurement and maintenance, establishes secure processes and saves a lot of time.

The built-in conversion interface to external SAP DMS conversion servers enables archiving and view formats to be created automatically or according to the status or user.

The SAP DMS manages documents for various SAP applications:

- Application data about connecting to external applications (e.g., Office, CAD, DTP),
- SAP PLM – Product Lifecycle Management,
- EasyDMS – the user-friendly DMS interface,
- Collaboration scenario with cFolders

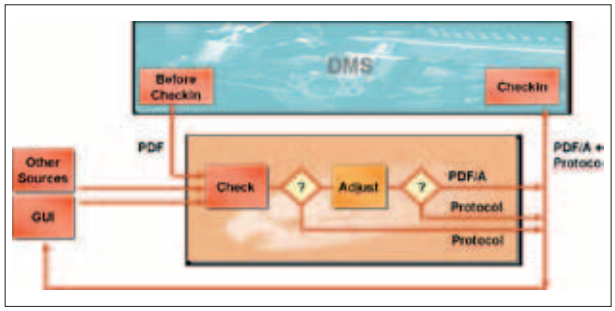
Therefore, in SAP DMS, data from very different sources and applications runs together. Since this product and process data is often subject to long record retention requirements, a quality check of the long-term formats is essential: The quality of new PDF files must already be ensured at the time when the files are stored or created in various applications and this quality must be ensured using tools for automatically checking and adjusting.



The SAP user controls the check process.

Using SAP interfaces to create and check PDF/A

Conversion interfaces already exist for SAP DMS and other processes. These conversion interfaces can be used for PDF/A processing. The SAP document management system DMS uses a built-in conversion interface that communicates with third-party conversion servers. This means that long-term formats such as PDF/A can be created and checked in again interactively or according to the status. This interface is also suitable for converting 3D CAD formats. For this, an archiving file must be generated from many individual files that are in the form of material document lists.



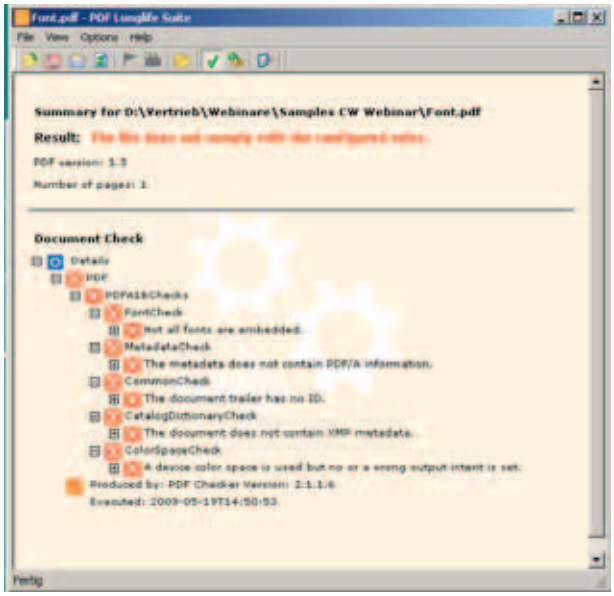
SAP DMS integration with PDF/A check.

The certifiable NetWeaver interface BC-XDC (External Document Converter) can also be used to create PDF/A. This interface enables a conversion procedure to be called from an SAP application. The type and number of application formats that can be converted depends on the functional range of the implementation of the BC-XDC server.

For the conversion procedure, you can use the PDF input and PDF/A output to perform a PDF/A check or adjustment. The BC-XDC interface is used by several SAP standard applications but is also available to all developers and solutions architects as of WAS 6.40. DPF4BCXDC is the name of the interface implementation from SEAL Systems Inc.

Central service for creating formats and quality management

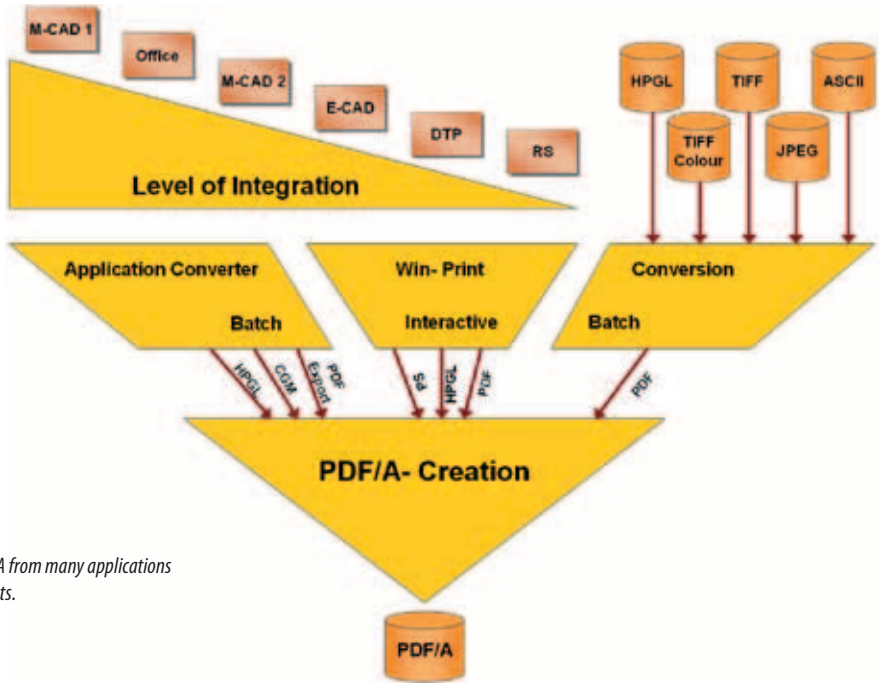
In practice, central server-based procedures have proven their worth. For all applications in a company, these procedures create and check PDF files and adjust PDF files from elsewhere to the quality guidelines. Companies such as SEAL Systems can deliver suitable solutions through the PDF Longlife Suite: PDF Checker, PDF Adjust and a standard integration for the SAP DMS.



PDF Checker checks that the standard is met – PDF Adjust adjusts the document to PDF/A.

Suitable converters are designed for the conversion of application data to PDF/A. There can and must not be a standard decision about which method is used for the application. In this case, it makes sense to classify the applications that are involved according to their level of integration in the PLM system. Usually, one or two CAD systems for mechanical construction and an Office package for managing the accompanying documents of all types are directly linked to the PLM system. These applications require a fully-integrated PDF/A generation. The interactive creation of PDF/A from all other applications (the other M-CAD systems, E-CAD, DTP, content management systems) seems to be sufficient. Other converters must be used for converting diverse legacy data (HPGL/CGM, PS/PDF, TIFF G4 and colour TIFF, ASCII, JPEG).

When several companies collaborate in the product development process, they exchange a large number of documents. When transferring the documents to their own data storage and storing them in the SAP DMS, these documents must be adjusted to the company's own quality standards. When doing this, the company must pay attention to legal aspects, such as the copyright protection of the original document. In this case, the company must not only store the document that has been adjusted to meet its own standards but, for legal reasons, they must also preserve the original document. For this, the integrated quality management should therefore also create a relevant protocol that includes test steps, their results and

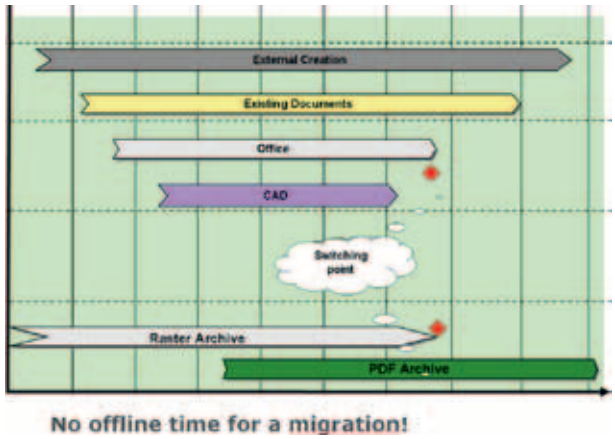


Creating PDF/A from many applications and file formats.

the changes that were made and should provide this protocol for storage in the SAP DMS.

Carefully planning the migration process

Project experiences have shown, that many PDF files can be retroactively changed to the stricter PDF/A standard. However, in principle, this is not possible. Migration projects show that about 10 – 30% of files cannot be made to conform to PDF/A without reusing the original application.



Migration concept for a company with PDF/A as the target.

When choosing a suitable set of software tools, it is useful to plan a step-by-step implementation. In this connection, it is advised to structure the volume of documents that exist in the company according to the processes that are responsible for their formation. In doing so, we can identify document groups for which a longer time is required for conversion the into qualitative, high-quality PDF files. This mostly concerns the large number of externally-created documents that, over years, were gathered in the company from various unsecure sources. The migration should begin with these documents.

However, the documents that are currently being received are often equally important to the activities of a company. If an automatic quality check is activated for the check-in to SAP DMS, the task of long-term archiving using PDF and SAP DMS can then be regarded as solved.

PDF/A establishes security through standardization

The PDF/A standard establishes mandatory rules for PDF documents that are intended for long-term archiving. In addition, standards are established for document exchange between companies.

As a result, the administration effort is reduced when using documents within a company and when exchanging documents with suppliers and customers. The processes run smoothly.

Today, tools exist for creating and checking the PDF/A standard. For the SAP internal document management system DMS and for other DMS and PDM systems, these tools ensure that the only documents that are checked in are those that conform to the standard.

Optimising PDF/A Documents for Large Archives

Harald Grumser,

CEO, Compart AG and Chairman of the PDF/A Competence Center

Nowadays, activities between enterprises and end-users are collectively referred to as B2C (Business to Consumer). This class of business, also commonly known as e-business, typically involves a high volume of communication in the form of offers, invoices, order confirmation, performance reports, policies or bank statements. While the volume of individual, physical documents (in the vernacular – letters) continues to fall in almost all countries, the percentage of electronic documents distributed as e-mail or via web portals increases disproportionately. When these documents have to be facsimiles of the original paper form there is no getting around PDF and as a consequence PDF/A.

Trade-Offs

From a technical perspective there are two types of IT systems which have to deal with these documents: The output management system (OMS) is used to create documents and provide the dispatch logic, while the classic document management system (DMS) is used to archive the same documents according to the relevant regulations for periods ranging from months (e.g. for itemized billing) to a number of years (life insurance documentation). In recent years a higher-level discipline, Enterprise Content Management (ECM) has come to be seen as uniting both requirements. At the first glance, both sets of needs can be fulfilled by PDF/A, but in practice differing technologies are used:

- High-volume printing, meaning for example industrial production and collection into an accordingly large spooling file containing up to one million letters in a single day. A typical format for doing this would be AFP, developed over 20 years ago, and particularly appropriate for resource optimization for large printing systems. Datastreams may also be PostScript or PCL.

- Selected individual documents may also be stored in an archiving system to secure customer documents or processes. Unfortunately, due to weaknesses in the software, TIFF raster format is often encountered here. More and more companies have come to recognize the advantages of PDF and in particular PDF/A, and changed their archiving system.

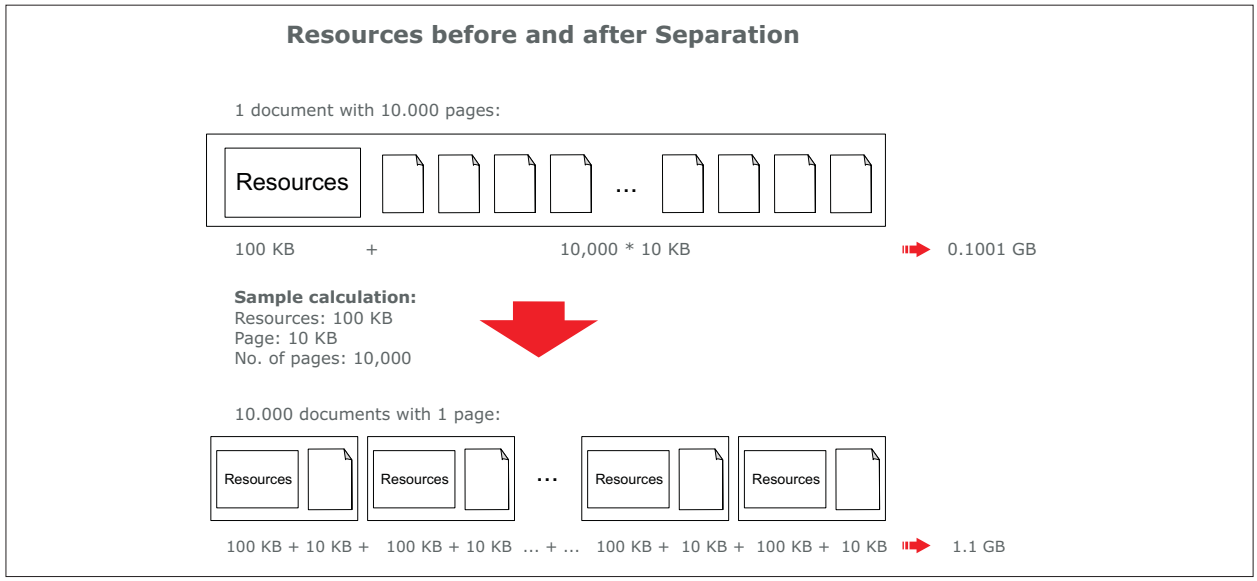
The conflict is readily apparent: While printing involves collecting a lot of documents, so that resources such as fonts or images are present only once in the datastream, for archiving documents have to be stored individually, with the result that relevant fonts and images always have to be embedded with each document. An often used argument against the use of PDF/A is the real need to have to embed all fonts, which from the perspective of long-term archiving should not really be a matter of contention. In any case, independently of the way in which documents are archived, they should be provided in PDF/A since the end-user will perhaps want to save them in a (perhaps smaller) archive.

The following section details the options of how to overcome this dilemma.

Optimised File Size

When documents are saved as individual files together with their embedded resources, the simple question is – how large will each individual file be? In many cases the features available to create an efficient PDF/A are fully used, so that the PDF/A files are almost as large as the corresponding TIFF file size. By considering the points following, PDF/A file sizes should be smaller than the related TIFF files:

- Select the right compression: PDF/A offers a wide selection of compression options appropriate for each of the different data types. So, for example, JBIG2 com-



pression for black and white images is significantly better than FAX G4. And JPEG is usually worse for colour line drawings than flat compression.

- Re-use resources: PDF/A offers almost unlimited options to store a single instance of often used resources such as images or overlays. Unfortunately many applications, in particular print drivers, make little or no use of these options.

- Use font sets: PDF/A supports the use of font subsets. This means that only the actual characters used in a document are saved as a font. This has the advantage that large fonts containing hundreds of characters of many 100 kilobytes are reduced to just a few kilobytes.

- Use a limited number of fonts: A single font, even as a font subset, uses a multiple of the storage space required for the text contained on a page. Apart from this, using just a minimal number of fonts saves not only space, but also topographically looks better.

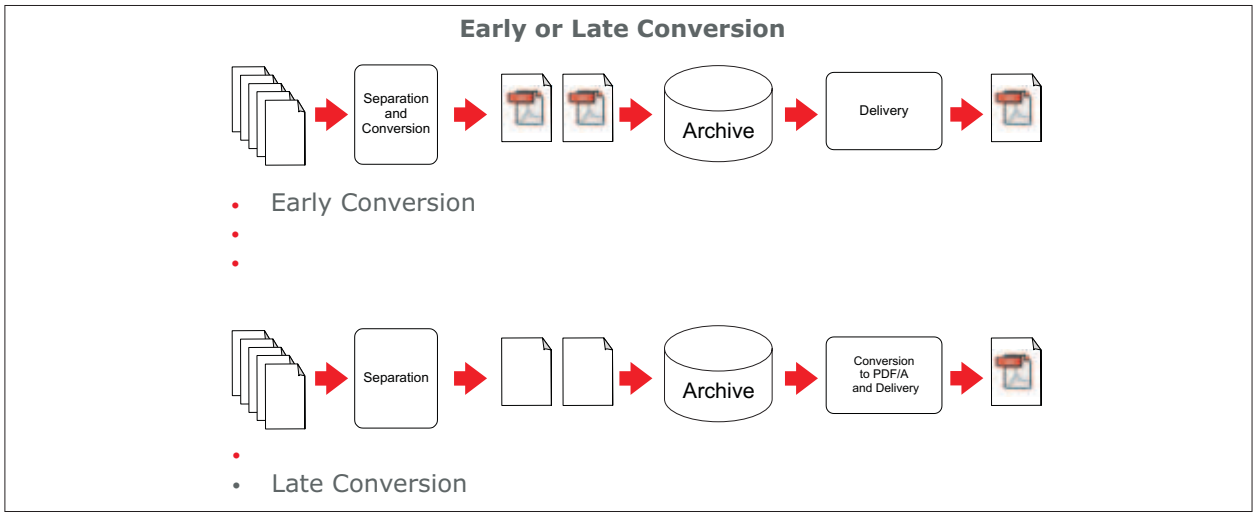
- Check the colour profile: PDF/A forces the use of device-independent colours such as RGB or CMYK colour profiles to ensure correct colour reproduction. For many archiving requirements the use, for example of CIELAB can be discontinued by specifying device-independent colour. If a colour profile has to be used, it does not have to always be more than 1 megabyte in size.

- Remove unwanted information: Metadata is wonderful, can however lead to very large files which are not required in the archive.

Optimised PDF/A files often require less than 100 kilobytes to save multi-paged documents, whereby the advantage of PDF/A compared to TIFF really comes to bear as the size of the document increases.

Keep Documents Together

Separating individual documents creates an unacceptable increase in storage requirements, it can be worthwhile to convert all the spool files to PDF/A, and to then archive them. In this case individual documents are extracted from the combined files when the archive is read. Converting arbitrary print data to PDF/A while retaining resource management is now-a-days no longer a significant technical challenge. On the archive side however, there is more work to be done because in the archive index the page number and size of the target document within the large spool file has to be managed, and in addition, for the purposes of retrieving an individual document, an active component has to be included so that the target page can be extracted. For performance reasons there will be no noticeable delay because extracting a number of pages



from a PDF/A file containing many thousand pages on a system with fast disks happens in fractions of a second. The PDF format has internal structures designed specifically to meet this requirement. These processes are by no means new, but have been included in archiving systems developed to do spool-based archiving since many years.

Late Conversion

If the effort to save a document in an archive in comparison to the effort necessary to retrieve a single document is too high, late conversion may be an appropriate alternative. This means the document will be converted into PDF/A format from its original print datastream only when it is read from the archive. If hundreds of thousand documents need to be archived daily, but only a few need to be taken from the archive, this method should be considered, taking into consideration the total cost per page. The work involved from the point of view of the archive is not significantly higher than the method described previously, but it should be noted that support for the datastream must to be guaranteed for many years to come and that a digital signature cannot be implemented with this approach.

In summary it is desirable to keep the number of formats in an archive to the absolute minimum. So for example, single documents coming from say office applications, or scanned input documents are to be immediately saved as PDF/A and that only the high-volume mass data is to be treated in a special way. If only AFP and PDF/A data are to be managed, there is a high level of certainty that both formats can continue to be supported.

Re-creating Print Data from PDF/A Documents

The reverse direction from archived data back to datastream is by no means an ordinary task. This can be an issue when documents are stored in a portal application directly as PDF/A, but nonetheless a printout request must be fulfilled. The conversion from PDF/A to a print datastream is no longer a noteworthy challenge; there are a number of products and printing systems available which can directly print PDF/A.

One difficulty may well be hidden in this approach: When many small PDF/A files with font subsets are concatenated or converted, it is not unusual that a print file is created containing hundred if not thousands of fonts, bring not just performance problems but which can also lead to unprintable files. This problem can be circumvented by creating a single font from those fonts which are the same but include different characters. This is referred to as a font super-setting, which is a minor challenge for all those involved.

Conclusion

There is no obvious silver bullet solution to choosing an archiving process for mass data. What is important is to carefully study the alternatives and to find a solution independent of the existing components being used. The cost associated with large documents in an archiving system may still play a role today, but within a few years this could be irrelevant. So, at this stage: better PDF/A today than tomorrow.

Promotional Contributions of our Sponsors and Exhibitors

Exhibitor: 4IT Group, Italy

Dialoguing with the market to gain competitive edge

4IT Group observes the ICT, Graphic Arts and Business Communication markets, talks about its trends, intercepts changes in it and is ahead of the game on its opportunities.

- It pulls together skills and reports, and creates networking opportunities.
- It publishes content that transforms culture and information.
- It cultivates its carefully selected contacts that have been gathered through 18 years work on behalf of associations, working groups and companies in High Tech and Information Technology.

Skill Areas

Document Management

Document life-cycle management and Enterprise Content Management. The correct structuring of information. Processes for greater efficiency and effectiveness for the taking of strategic decisions in communications, in dialogues with clients and towards the target market.

Green

Sustainable business models in the office and in graphics arts. The eco-compatible strategies that improve the company's finances. Green marketing tools to improve company running and stimulate choices in line with an ecological conscience and responsible consumer behaviour.

Printing

Digital printing in office and production settings. Products and solutions to make the treatment of printed documents more effective, business communication applications, high volume printing projects and implementations, optimization of the use of consumables and printing supports.

Educational Projects

Document Management Academy

Held in cooperation with SDA Bocconi School of Management, DMA is a Post-Graduate course in Document Management to prepare the professional figure of the Document Manager to work in enterprises. It is a one year course made up of about 40 study sessions, workshops on topics integrated with document management and business management and on internship carried out in the field.

Master Course in Document Legal Archiving and management

The course aims to train those who are in charge of legal archiving within a company. It addresses legal, administrative and fiscal regulations and all topics related to document storage in accountancy, public administration and health system, as well as matters related to documents security

Editorial Products

Digital Document magazine puts into the foreground the data, trends and prospects for the Italian and international market by way of exclusive agreements with the most important sector associations and the most prestigious research companies, like InfoTrends.

All of this goes towards supplying a constantly updated picture on development in the sector and a privileged position on current and expected market dynamics.

The magazine addresses company strategies, successful application experiences, case studies, technological news

on products and services, the most innovative solutions in the document management, content management and printing area.

DDm Europe is the bi-monthly publication which collects all information relevant to the European digital document management markets.

- It comes with the contribution of central figures and journalists in the industry and targets enterprises' management, printers and end users of technology and technology's applications.
- DDm Europe is the publication with European distribution which gathers the international experience, relations and contents of 4IT Group, a publishing company already producing publications and events at national level in different European countries.
- It is posted only to a selected mailing list, an exclusive group of profiled managers who make up the industry's leadership today.
- The magazine addresses the topics of TransPromo, Enterprise Content Management, Book On demand, Web to Print, Mail and Postal services, Cross Media Communication.

Solutions Directory and Guides
Monographs on subjects of particular interest, each of which presents a picture of the Italian market, relating to the topic dealt with and placed in the broadest European setting. This is done through the presentation of one or more analyses from the most internationally respected research institutes.

- Opinions from the sector with the central figures' comments.
- Structured list of vendors, services and products on the market

Trade Fairs

DocuBusiness
The exhibition-convention dedicated to Document Management solutions, focussed on technologies and processes that are part of content management.

At DocuBusiness there is a conference area which focuses on topics such as electronic invoicing and legal archiving, document life-cycle management, and the laws on electronic invoicing.

Inprinting
Inprinting is a four days exhibition on the vertical segments of digital printing in high volume applications (Transpromo, Direct Mail, digital printing and book on demand, graphic arts and visual communication).

As well as this, it represents the segment of office printing: Managed Print Service, Pay per Page, TCO, without overlooking Green Printing and the topics of environmental sustainability.

DM Expo
The exhibition dedicated to marketing and commercial managers, advertising and promotion managers, press office, general managers, managing directors and business owners, buying managers and all kind of communications people. The objective for the companies at the show is to build business in the area of direct marketing by presenting to the attendees the true potential of direct marketing and the results it can achieve.

Conferences

Digital Printing Forum
It has been held in Milan for 13 years and it's considered the privileged place for the main market players to exchange knowledge and best practices, but also the opportunity for manufacturers and solution providers to meet companies acting in the graphic arts, digital printing, communication and marketing sectors.

The digital equipment sales in the last three years involved 70% of the digital community attending the Forum, in other words those companies, professional figures and decision makers who, by attending, increase the value of this event.

MailForum
MailForum is an international Forum dedicated to business communication, Direct Marketing and postal regulations and liberalisation and Data Protection Act

BookForum
BookForum analyses the dynamics which entwine the digital printing market with the book market.

It observes the development and economic impact in various evolving markets and shows all the ways of conceiving a book, either as a publishing or photographic product, or as a business communication tool.

Having content available in digital format opens up new possibilities that were previously unthinkable. Digital printing has the chance to manage the profound changes that the book is going through.

European Marketing Dynamics Summit
The European Marketing Dynamics Summit is an educational forum located in Bruxelles for marketers, advertisers, service providers, and vendors on cross-media direct marketing opportunities and strategies.

Challenges related to new media, changes in consumer behaviour, the market environment, and global expansion have all contributed to reasons why businesses are seeking new ways to market to, communicate with, and reach customers. Marketing service providers and print service providers must re-evaluate the services they are offering to customers and the role that they want to play. Transpromo and other cross-media marketing campaigns that blend print and other media offer ways to accomplish this. In is organised in cooperation with InfoTrends.

callas software: A Core Provider of PDF/A Technology and Solutions

callas software is a leading vendor of PDF/A technology and has been actively involved in the development and shaping of PDF related ISO standards for many years. When Olaf Drümmer founded the company in Berlin in 1995, the Portable Document Format (PDF) had existed for only two years and was just growing out of its infancy. Today a world without PDF can hardly be imagined anymore, and callas software continues to play an important role when it comes to ensuring reliable use, exchange and preservation of documents in the form of PDF.

About callas software

In the world of printing and publishing, PDF very quickly gained importance as a file format for the exchange of digital masters, and the first correction tools from callas software were geared towards the high quality requirements of this sector. As the German representative in the ISO, Olaf Drümmer has been deeply involved in the adoption of the PDF/X standard for exchanging master copies. When, in the following years, PDF was also standardized in ISO for other application areas, Olaf Drümmer again contributed to the development of the PDF/A standard for long-term archiving as well as other PDF related standards, like PDF/E, PDF/UA or PDF/VT. He was also deeply involved in the work on the new PDF/A-2 part of the PDF/A standard. Through the active exchange with all of the participating specialists during the standardization, newly-gained knowledge is always swiftly incorporated into the PDF validation, optimization and conversion technology from callas software.

Today callas software is the first port of call internationally for PDF validation and the development of high-quality optimization and correction tools for PDF of virtually any form and origin. For more than ten years, the software developer and PDF specialist has worked on the continuous further development of their PDF technology, which is based on an uncompromising concept of in-depth analysis and therefore does not set any restrictions regarding special or less common constructions in PDF documents. This comprehensive approach emanates from a complete analysis of the PDF and also includes embedded formats: the various font types, XMP metadata, ICC profiles for colour display, the various image compression algorithms and many more. Only this approach gives you the option to prevent, from the outset, hidden problems that occur in a complex file format like PDF. This is the basis for a versatile conversion engine that can solve the widest range of PDF/A problems. Unique in this area is the ability to flatten transparent objects. Whether in print publishing, document processing or for archiving, the PDF validation, optimization and correction technology from callas software offers reliable checks for the various application areas of PDF while any user can very easily tailor the technology to their specific requirements.

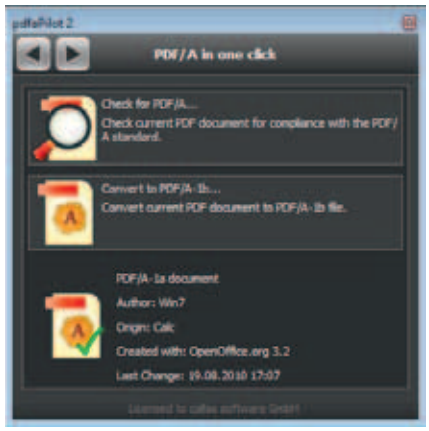
callas software technology for checking, optimising and correcting PDF documents

The PDF technology from callas software provides the most thorough PDF analysis and conversion that is currently possible. It is also possible to convert native office files from Microsoft Office or Open Office directly to PDF/A. That allows you to maintain additional information like the structure information (Tagging) that is important for PDF/A-1a, links or metadata and to combine that with the additional improvements.

For interactive use, callas software offers products as stand-alone programs or as Adobe Acrobat Plug-Ins. In addition, there are server products and command-line versions that can easily be integrated as well as programming libraries (SDK). As a result, integrators, inhouse developers and OEM partners can take advantage of the technology and functionality of the products from callas software in their own projects and solutions. For this reason, even vendors that operate globally (such as Adobe, Xerox, Hewlett Packard, Mitsubishi Paper Mills and Quark) have licensed this technology to analyse and process PDF files.

The inventor of the PDF format, Adobe Systems, has used the validation technology from callas software in Adobe Acrobat Professional since 2003, where it is integrated into Adobe Acrobat under the name Preflight.

Many vendors world-wide also rely on PDF/A technology from callas software in solutions for document management and archiving, whenever they provide their customers with PDF/A support for their document processes and archiving requirements. Argus, COI, Optimal Systems and SER, and many more well known Enterprise Content Management and Document Management System vendors have entered into a PDF/A technology partnership with callas software.



callas pdfaPilot is the professional tool to validate and/or create PDF/A-files for long-term archiving.

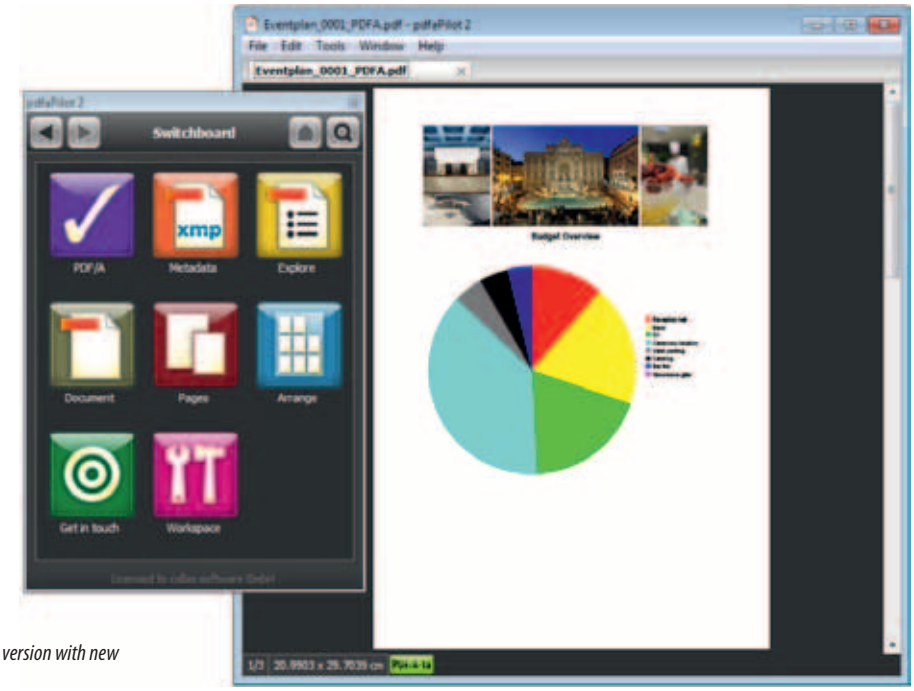
callas pdfaPilot – desktop version

For archive users, government agencies and enterprises, callas pdfaPilot Desktop 2 is the professional tool to validate and/or create PDF/A-1 or PDF/A-2 files for long-term archiving. It turns all kinds of PDF or office files like interactive forms, construction drawings, presentations, newspapers, magazines or books, into PDF/A documents that conform to ISO 19005.

pdfaPilot is an easy-to-use, quick and secure tool to create PDF/A documents. Non-compliant PDF or office files can be converted and saved as valid PDF/A documents with the touch of a button. callas pdfaPilot uses the same validation technology as the Preflight function licensed by callas software to Adobe for their Adobe Acrobat 9 Pro application.

callas pdfaPilot features an intuitive user interface with easily understandable instructions and detailed explanations for the user.

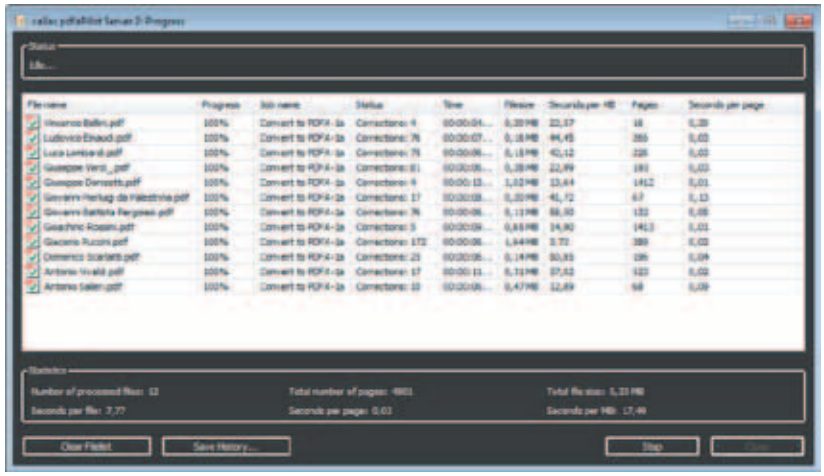
Any organization in both the private sector as well as the public sector can minimize their risks related to the long-term archiving of electronic documents, and



callas pdfaPilot stand alone version with new intuitive user interface.

callas pdfaPilot enables organizations to comply with regulatory requirements. callas pdfaPilot allows them reliable long-term archiving of their valuable documents, regardless of the software they were originally produced with. It avoids dependence on a particular product, manufacturer, or operating system.

In addition, suitable PDF files can be saved as PDF/A-1a or PDF/A-2a. This means that the logical structure of such a file, with its carefully marked-up or "tagged" content, will still be made available in the long-term. As a consequence, the text in the PDF/A-2a document can be extracted as Unicode, and the content of the PDF/A-2a document can be converted to other semantic representations, whether to support intelligent search and retrieval or conversion at a later time for migration purposes.



callas pdfaPilot Server automates and centralizes the validation and correction of large volumes of PDF/A files.

Overview:

- Checks PDF files for compliance with any part or conformance level of the PDF/A standard (ISO 19005): PDF/A-1b, PDF/A-1a, PDF/A-2b, PDF/A-2u, PDF/A-2a
- Converts PDF files to PDF/A and implements all necessary corrections
- Embeds and/or replaces missing fonts and handles missing glyphs
- Optimises all colour data for compliance with the PDF/A standard
- Adjusts comments and form fields to the defaults required by the PDF/A standard
- Removes unwanted attributes such as layers or interactive content like movies
- Makes image compressions compliant with the PDF/A standard
- Creates PDF/A documents that are web optimised for easier access and viewing
- Brings document metadata in line with PDF/A requirements
- Converts newer PDF features in those ones that are compatible with the respective PDF/A provision
- Implements adjustments and corrections without loss of information
- Delivers clear reports to document all test and correction procedures
- Improves overall accessibility of PDF/A
- It is available in English, German, French, Italian, Spanish, Japanese and Chinese language versions

callas pdfaPilot Server 2, pdfaPilot Command Line 2 and pdfaPilot SDK 2 (Programming Library)

callas pdfaPilot Server automates and centralizes the validation and correction of large volumes of PDF/A files. Whatever your sector – financial, legal, manufacturing, public or other – pdfaPilot Server is the ideal tool for producing high quantities of PDF/A files easily for long-term archiving. A wide variety of types of PDF documents, from straightforward memos to complex forms and documentation can be made PDF/A compliant, with pdfaPilot Server embedding fonts, adjusting dynamic elements, making colours device-independent and performing many other required changes.

pdfaPilot Server also features an intuitive user-interface that enables quick set-up. And if you want to integrate PDF/A compliance into existing workflows, integration is easy through hot folders or a command line interface. Even a software library (for C, .NET or Java) is

available. HTML- or XML-format reporting further support seamless integration.

The solutions are available on the widest range of platforms: Windows, Linux, Sun Solaris (x86), Sun Solaris (SPARC), Macintosh and IBM AIX.

Overview:

- Corrects and converts PDF files to PDF/A within seconds, implementing all necessary corrections
- Large amounts of PDF files can be quickly converted according to PDF/A-1
- Validates PDF files according to the PDF/A ISO standard (ISO 19005)
- Supports all parts and conformance levels of PDF/A: PDF/A-1b, PDF/A-1a, PDF/A-2b, PDF/A-2u, PDF/A-2a
- Provides comprehensive reports on all test and correction procedures
- Supports archiving accessible PDF documents according to PDF/A-2a or PDF/A-1a
- All adjustments and corrections are implemented without the loss of information
- Embeds missing or incorrect font characters into PDF files, with full support for PostScript, TrueType and CID fonts
- Offers optional font substitution in case of font embedding restrictions
- Optimises all colour data for compliance to the PDF/A standard
- Makes comments and form fields compliant with the PDF/A standard
- Flattens transparent objects for PDF/A-1
- Adapts image compression according to the PDF/A standard
- Uses an intuitive user-interface for configuring and managing hot folder automation
- Provides clear HTML reports with extensive configuration options
- Machine-readable XML reports offer optimal integration into automated processes
- Easy and flexible integration into existing DMS solutions and archive workflows achieved through a command-line interface
- Performs additional custom checks upon verification if needed
- More than 100 additional correction or modification features allow for adapting any PDF to individual needs
- Available in German, English, French, Italian, Spanish, Japanese and Chinese on Windows, Linux, Sun Solaris, AIX and Mac OS X.

Further information is available at:
www.callassoftware.com

Content and Document Management Europe Limited (cdm Europe)

cdm Europe is a community of European associations and association chapter representatives focused on Enterprise Content Management technology and services.

cdm Europe was founded during CeBIT 2010 as a co-operation of equal rights member associations with an annually rotating secretary. VOI e.V. (Germany) is hosting the secretary in 2010 and DIMO (Denmark) will be the hosting association in 2011.

Current members include:

AEDOC Digital – www.aedocdigital.org

Aproged – www.aproged.org

DIMO – www.dimo.dk

Document@Work – www.documentatwork.be

PDF/A Competence Center – www.pdfa.org

VOI – www.voi.de

The initiative´s intention is

- to bring together individuals and organizations from both the public and private sectors involved in Enterprise Content Management.
- to support the member associations in fulfilling their roles and responsibilities and to raise awareness in the field of Enterprise Content Management.
- to provide technology and knowledge transfer and information services, appropriate practice guidelines, benchmark indicators and information, educational, skills, development and research opportunities.
- to support the cooperation between local associations, local solutions suppliers and partners in Europe.

For further information, please contact:

Content and Document Management Europe Limited (cdm Europe)
c/o VOI eV

Tel: +49 (0) 171 / 2217975
Mail: info@cdm-europe.org
www.cdm-europe.org

Crawford Technologies – High-Volume Transactional Output Software

Since 1995, Crawford Technologies' award-winning solutions have helped over 700 companies around the world reduce costs associated with mission-critical transactional customer communications. Our products, solutions, services and expertise help our clients deliver key documents to their customers in the format they need, when they need it including Braille, e-text, audio and large-print formats for the visually impaired and print disabled.

Famous for print-stream transformation, document re-engineering, Transpromo, archive & retrieval and customer communications software, CrawfordTech's products and expert services simplify, automate and extend the value of document delivery cost-effectively.

Our innovative and flexible solutions enable meaningful process improvements irrespective of our clients' current, legacy or future standards in infrastructure or document output achieving operational savings.

Companies look to Crawford Technologies for our platform-independent approach, leading system performance and superior output fidelity. CrawfordTech's quality software, proven support and extensive document industry knowledge help our clients meet operational goals, reduce costs, support compliance and bolster the effectiveness of their document processing infrastructures.

Our platform and vendor independent architecture insures our products and solutions seamlessly interface with existing in-house, legacy and third party applications and workflows. CrawfordTech software is available for Windows, Linux, UNIX, z/OS, z/Linux and most other operating systems.

PRO Transform Family

CrawfordTech's PRO Transform Family consists of a suite of flexible, fast and accurate print stream conversion products. These affordable enterprise utilities are rapidly implemented to convert existing print streams, delivering superior output fidelity and leading system performance.

Convert from and to these formats and more:

- AFP
- LCDS/DJDE
- PostScript
- PDF and PDF/A
- PCL
- Image (TIFF, PNG etc.)
- Text (ASCII, EBCDIC, HTML, XML etc.)

Our PRO transform technology is widely recognised for its processing speed, small output file size, fidelity, single pass processing, platform independence and cost effective implementation.

As a result, our customers extend the value of legacy document streams, bring increased flexibility and efficiency to their production operations, and maximise the latest technology developments without the need for document reprogramming or redesign.

PRO Document Enhancer Family

The PRO Document Enhancer Family of products allows you to re-engineer and re-purpose print applications by making rapid, focused changes to legacy applications. Since these enhancements are done at the print file level and not at the applications system level your operational efficiency and customer communication effectiveness projects are completed rapidly and with a minimal IT resources, therefore enhancing the financial results of your projects goals.

These enhancements happen in a single pass during the conversion of your input file format to your choice of output via our PRO Transform Family of products. Imagine converting from AFP to PDF/A while adding additional functionality to the output on-the-fly.

- With **Transpromo Express** you can quickly and affordably insert custom marketing messages and graphics into existing white space on your customer bills, statements and invoices – without the need for complicated programming. Remove the barriers to adding promotional messages to your transactional documents (Transpromo) and open the way to a new marketing channel to your existing customers.

- **QR Express**, an optional module for CrawfordTech's Transpromo Express, enables the rapid and inexpensive application of QR Codes to bills, statements and other transactional customer communications without a lengthy and expensive IT project. By enhancing the print files of your bills, statements and other transactional customer communications, CrawfordTech's

QR Express allows you to apply QR Codes to those documents. Via your customers' smart phones, QR codes take users to a personalized URL for immediate connection to a promotional offer, to learn more about a new service or for other communications purpose.

- With **Operations Express** you can quickly and cost-effectively leverage your production print streams to optimize print output and automate mail processing across your existing printer and inserter investments, consolidate your work to fewer devices and reduce operational costs without having to make expensive and time consuming re-composition or application changes upstream. Operations Express adds flexibility to print and mail room operations, optimizes equipment utilisation, increases operator productivity and reduces your operational costs.

PRO Workflow Family

CrawfordTech's PRO WorkFlow family builds on our proven PRO WorkFlow JES and PRO WorkFlow Server products by adding our powerful PRO WorkFlow Connector software to provide a single workflow solution.

Our customers automate their document development, application testing, print production, mail processing, and archiving processes to achieve significant savings and efficiencies. This powerful product family, unique to the transactional document market place, frees data files, resources, metadata and related production processes from any dependence on computing platforms throughout the workflow of the processes in your document's life cycle.

The PRO WorkFlow product family includes the following products:

- **PRO WorkFlow Connector** is a fully automated, high performance solution focused on the specific workflow needs of your transactional document production workflows. It is used to move output related data streams and resources freely between platforms to rationalize cost and load factors. You can also initiate data manipulation applications, such as transformations, postal sorting, print stream re-engineering, 'ACIFing' your AFP print files, composition processes and more, based on your needs.

- **PRO WorkFlow Server** is a queue manager with e-mail notifications that is used to automate data manipulation, application transformations and print stream re-engineering while concurrently routing the output files to dynamically assigned destinations such as printers, folders and FTP sites.

- **PRO WorkFlow JES** is a software tool used for automating the extraction of input files from the JES spool for a variety of enhancements and improvements. Utilizing TCP/IP and FTP, users manipulate data and applications right from the JES queue and route output files to the specified destinations, applications and post processing systems.

Use the power of the PRO WorkFlow family on the widest range of input and output formats in the industry, including LCDS/Metacode, PostScript, PDF, AFP, PCL, HTML, TIFF, Flat files, Line data, ASCII, EBCDIC, SCS and PDF/A.

Contact your local Crawford Technologies office now:

Click 'Contact Us' at www.crawfordtech.com
E-mail us at sales@crawfordtech.com
Follow us on Twitter at twitter.com/CrawfordTechInc

Exhibitor: Digital Planet, Turkey

Digital Planet is a leading technology company, delivering customer communication management systems based on related sector needs for nearly 10 years. We are proud to say that our customer communication management and hybrid mail solutions are used as a standard by leading banks, financial and telco companies. Every month approximately 100 million pages of documents are processed through our solutions and millions of document deliveries are executed via web, e-mail and fax.

Digital Planet, founded in 2001, has a 95% share of the market in Turkey in bulk document production and output management, and is growing internationally with new customers in Russia and the Middle East. Our products, covering document production, delivery, and electronic content management, are used in the financial, telecommunications, postal, service bureaus, and utility sectors.

We also have a proprietary e-invoice application that is completely integrated into our solutions and includes digital signatures. NetVault is unique in Turkey, enabling digital archiving and delivery of documents like invoices, bank statements, etc. with a high compression ratio (90:1).

The NetVault enterprise stream archiving solution was recently enhanced to enable PDF/A documents to be archived with a 90-99% compression ratio without any performance overhead. This revolutionary and unique technology eliminates the major concern of high storage demands for PDF/A archival.

In addition to PDF/A and stream archival, Digital Planet offer a wide selection of products in the document delivery and document management fields. Netgateway, our leading product in Turkey for e-messaging, is pure Digital Planet technology. With its high scalable architecture, Netgateway enables customers to send up to 1.5 million e-mail messages attached with digitally signed PDF and PDF/A documents in 24 hours.

NetECM, released in 2009 from Digital Planet, allows customers to manage documents in electronic content repositories with workflow management capabilities. Powered with NetVault core architecture, NetECM changes your understanding of document management.

We at Digital Planet will continue to present innovative and reliable solutions and set bigger targets for foreign markets with our young, dynamic and professional team.

Internet: www.dtp.com.tr

intarsys – Electronic Signature and PDF/A

intarsys is a leading provider of electronic signature software and solutions for electronic forms-based business workflows. An outstanding feature of the Sign Live! product line is the support of various types of electronic signatures, starting with advanced electronic signatures and biometrical signatures (e.g. pen-pad) through to qualified electronic signatures. intarsys has a Common Criteria certification for Sign Live! which is compliant with EAL3+, thus satisfying the highest security needs. As Sign Live! is based on the CABAReT Stage software platform it inherits all PDF features from this powerful base application. Therefore, PDF/A forms an integral part of Sign Live!, resulting in a unique offer to combine the potential of both technologies in a coherent way. Customers of Sign Live! can be sure to get ISO standards compliant PDF/A documents with embedded advanced or qualified electronic signatures, multi-signed PDF/A forms and even hybrid signed documents with biometrical and digital signatures.

Since 1996 when the company was founded, intarsys has concentrated on optimising business workflows through the use of electronic documents and forms technology. The focus is on enabling straight-through processes without media breaks, thus allowing for complete electronic processing.

An important use case of PDF/A together with electronic signatures is eInvoicing, a fast growing application domain. Here intarsys not only offers the technological components to implement the generation and reception of tax-compliant eInvoices, but also provides consulting for the individual invoice business processes on a European level. intarsys is a core member of the German Forum on Electronic Invoices (FeRD), an association promoted by the German Ministry of Economic Affairs, and is responsible for the technical standards for electronic invoicing within this organization.

Customers of intarsys can be found in the areas of healthcare, aerospace industry, insurance companies, banks and e-commerce companies.

The solutions we provide are categorised into 3 product lines:

- CABAReT Stage for PDF-based document and forms solutions
- PDF/A Live! for the generation, validation and correction of PDF/A compliant files
- Sign Live! for electronic signing of documents in single, batch and mass signing modes

Reference customers of intarsys include Deutsche Post Com Signtrust, Areva, STILL AG, Lufthansa Systems, Deutsche Bank, KIA Motors, Swisscom, Concordia AG, HOLCIM AG and the City of Zug. In addition, intarsys software is used at Stadtsparkasse Köln-Bonn, Sparkasse Krefeld, NASPA, Frankfurter Sparkasse and more than 10 other savings banks in Germany.

intarsys' technological partners include, amongst others, Giesecke & Devrient, ReinerSCT, T-Systems Austria and Compart.

intarsys participates in various associations like VOI (DMS/ECM), GMDS (healthcare), bwcon eHealth (healthcare), GFaR (einvoice), FeRD (einvoice) and T7 (trust centres). In addition, intarsys is member of the board in the PDF/A Competence Center, the international association of businesses who deal with PDF/A. Within the PDF/A Competence Center intarsys chairs the Technical Working Group and takes responsibility for the electronic signature domain and the healthcare industry.

intarsys products

Sign Live! CC client

Sign Live! CC is an application that offers document functions with the highest level of security: it is certified in accordance with Common Criteria and is approved as secure in accordance with SigG/SigV (German Digital Signature Act/German Digital Signature Ordinance) by the German Federal Office for Infor-

mation Security (BSI). Sign Live! lets you edit, sign and encrypt your documents with a single application. Qualified signatures created with Sign Live! CC turn your files into legal electronic documents – regardless of their format. Sign Live! CC can be easily integrated into existing system environments and controlled through various interfaces. In addition, you can give the application your own corporate design. Sign Live! CC plug-ins are available for PDF editing and PDF form designing. Make use of PDF/A, the long-term archiving solution of the future, log document editing steps or introduce electronic document workflows free of media breaks – Sign Live! CC is your solution for secure electronic business processes.

Sign Live! CC server

Sign Live! CC signature server opens up a new dimension in speed and performance for signature workflows, realising maximum through-put volumes for the creation of advanced and qualified electronic signatures on your server. Sign Live! CC signature server is a multi-tenant, high-performance and reliable server application that operates in the background or as an embedded system. The product well fits computing centre needs where an automatic and fault-tolerant operation mode and scalable performance are required. Sign Live! CC signature server gives you a cost-efficient signature server application not only for service providers, but also for medium-scale users. User-specific document operations like dynamic add-on of signature fields, extraction of archive index information or the embedding of XML data (e.g. eInvoicing) can be handled as additional steps within one signature workflow.

Signing Functions

- Mass signatures and batch signatures with unlimited batches (PDF, PKCS#7, XML-DSig)
- Compliant with CAdES and XAdES
- Creates multiple signatures and certification signatures
- Integrates attribute certificates in signatures
- Supports time-stamp signatures
- Customisable graphical signature representation (stamps, images, etc.) in documents
- Highly-scalable by multiple smartcards or HSMs
- Multi-tenant operation supported through signature pools
- Hot replacement of smartcards

Additional Functions

- Smartcard tools (e.g. unlock/change of PINs)
- Decryption and encryption (AES-128, -192, -256, 3DES)
- Sending of e-mails and POP3 support
- Extensive scripting and reporting functions via JavaScript
- Installation check routine allows Trusted Mode operation

Sign Live! CC secure mail gateway

Sign Live! smg combines the power of the Sign Live! CC server with a flexible and robust mail gateway platform. This allows for the transparent signing of the complete outgoing e-mail traffic of your company as well as the validation of signed e-mails without any installation or modification on the client computers or e-mail applications. In addition, you can encrypt your e-mails or only the attachments by using advanced certificates either as software certificates or on smartcard. An outstanding feature is the transparent signature of outgoing electronic invoices, even with PDF/A conversion, which doesn't require local smartcard readers or other precautions at the client. Sign Live! smg is a highly scalable and cost-efficient signing gateway solution which can be used in SMEs as well as in large companies.

PDF/A Live!

PDF/A Live! allows you to comfortably validate PDF documents with respect to PDF/A-1 conformability. You can create and archive a validation log for each file, attach the log or summarise and archive multiple validations. The validation results are displayed in accordance with ISO-19005 specification chapters in order to resolve problem easier. PDF/A Live! can be enhanced with PDF operations by using the powerful JavaScript interface of the underlying CABAReT Stage platform. This allows for the full access of every PDF object within the document. You can implement PDF/A Live! in various client or server scenarios. In client mode you can review the docu-

ments on your screen and convert them immediately, turning documents over to the archive with the push of a button. In server operation you can use PDF/A Live! as a high performance server process which, for example, monitors entry lists. Batch processes can be designed to regularly convert any number documents. PDF/A Live! can be configured as a web server that uses SOAP or another interface to receive and archive documents. Or control the process directly using the intarsys JAVA-API, which offers you the absolute maximum freedom when working with your documents.

CABAReT Stage

CABAReT Stage is a software for filling out and editing PDF documents. The level of functionality depends on the type of license. CABAReT Stage offers:

- Extended mark-up functions:
  - highlight text
  - add notes anywhere
  - attach files
- Text extraction from PDF
- Exporting images / graphics from the PDF
- Creation of bookmarks
- TIFF to PDF conversion
- Pictures on buttons
- Password encryption
- Stamp tool
- Seamless integration:
  - “headless” operation
  - calling via command line
  - advanced scripting capabilities in JavaScript
  - direct access to Java routines

Further information can be found at www.intarsys.de

IntraFind AG – Specialist for Information Retrieval

The software publisher IntraFind AG located in Planegg near Munich is specialist for information retrieval from unstructured and structured enterprise data.

IntraFind provides its customers with software products, solutions and consultancy services for enterprise search, information access & text mining. The company, which has been founded in October 2000, employs 20 specialists.

The range of high-quality and standardized IntraFind products includes software products for efficient and scalable search, automated text classification, as well as Named Entity Recognition (NER). Additionally, the search solution can be enhanced by modules for linguistic textual analysis, semantic-associative search, thesaurus-based search, similarity search, text clustering, and knowledge maps. The IntraFind software products enable the user to efficiently search, find, filter, and sort textual information from various data sources. Furthermore, the software can be easily integrated into the customer' existing IT environment.

Besides quickly installable and applicable standard search software, IntraFind also provides tailor-made solutions for realizing individual customer requirements. The IntraFind team of experts is characterized by deep and comprehensive know-how as well as 10 years experience in the special topic "search". Furthermore, IntraFind does research in the fields of natural language query and relation mining.

In the course of search & retrieval projects IntraFind provides overall solutions which contain qualified consulting and project planning, software implementation and maintenance, as well as support in English and German.

IntraFind´s customer base includes numerous well-known companies and organizations in Germany, Austria, and Switzerland.

As IntraFind offers standard as well as specialist software and solutions, its customers belong to various economic sectors, as an abstract of the company´s references shows:

- **Automotive Industry:** Robert Bosch GmbH, Phoenix Contact GmbH & Co. KG, Audi AG
- **Chemical and Pharmaceutical Industry:** Beiersdorf AG, Merck KGaA, Evonik Degussa GmbH, Nycomed Deutschland GmbH, Boehringer-Ingelheim GmbH, DyStar Textilfarben GmbH & Co. Deutschland KG, Wacker Chemie AG

- **Trade and Industry:** Deutsche Post AG, Rohde & Schwarz GmbH & Co. KG, Tech Data GmbH & Co. OHG, Infineon AG, AMD, GfK Global Custom Research, Krauss-Maffei Wegmann GmbH & Co. KG, Voith AG
- **Publishers:** Haufe Verlagsgruppe, Herder Verlag GmbH, Gruner+Jahr AG & Co., Langenscheidt KG, Hexaglot Holding GmbH, Weka Media Publishing GmbH, Motor Presse Stuttgart GmbH & Co. KG
- **Finance and Banking:** BHF-Bank AG, Helvetia Schweizerische Versicherungsgesellschaft AG, HypoVereinsbank, Auxilia AG, Bank Burgenland, Landesbank Berlin AG
- **Public Sector:** State of Lower Saxony, German Army, Senate of Berlin, REEEP/UN, Police of Bavaria, Police of Baden-Württemberg, European Patent Office (EPO)

Service Spectrum:

- Information Logistics (crawler, advanced crawler, connection and analysis of external contents)
- Efficient and scalable full text search / indexing
- High-quality linguistic (→ leads to completeness and precision!)
- Associative search, search supported by taxonomies and thesauri
- Similarity search
- Named Entity Recognition (NER): automated recognition and extraction of entities (people, companies, locations, products, brands, company-specific entities)
- Information extraction
- Topic recognition
- Text clustering
- Guided search via ConteXtors and facets
- Intelligent navigation through existing metadata and through metadata generated by text mining methods (text classification and NER)
- Search "as easy as in an online shop"

Range of Products:

- **iFinder:** Enterprise Search Solution
- **TopicFinder:** Text Classification Solution
- **Namer:** Named Entity Extraction & Recognition
- **Linguistic:** Tool for morphological analyses
- **Advanced Crawler:** Intelligent Information Logistics
- **Content Grabber:** Rule-based extraction of content out of websites
- **Search+:** Collection of WebParts for improving the MOSS 2007 search

Range of Solutions

IntraFind´s range of solutions covers the whole bandwidth from simple full text search (application-specific) to enterprise search to highly specialized text mining solutions. This means, IntraFind provides the infrastructure for all search, retrieval, and text mining scenarios.

- Enterprise Search
- Intranet-& On-site Search
- Compliance & E-Discovery Solutions
- Semantic Linking for media portals
- Expert Identification
- Knowledge Maps
- Business Intelligence Solutions
- Information Logistics Solutions
- Newsletter Management / Newswatch
- Intelligent search for online shops
- Topic-based Internet search engines

For further information about IntraFind please refer to **www.intrafind.de**

LuraTech – Document Conversion Software

LuraTech provides software, services and outstanding support for document conversion with the objective of widely automated preparation of these documents for long-term archiving and processing in ERP and other systems. This includes LuraDocument PDF Compressor Enterprise, a production-level application for compression, conversion to PDF and PDF/A, OCR, classification and form data extraction. DocYard is a complete, centrally managed platform that integrates all functions of document conversion into seamless workflows. LuraTech's solutions developed for these complex tasks can be put into operation quickly and offer the stability and scalability to be used in any project. These solutions pay for themselves through flexible licensing models, regardless of whether the number of documents processed is small or very large.

Since the company was founded in 1995, LuraTech has been a leading provider of document and image compression solutions based on open and ISO standards. Based on the experience with the successful PDF, PDF/A and JPEG2000 products, LuraTech now also offers comprehensive document conversion products together with flexible services and outstanding support.

LuraTech's solutions are used worldwide, primarily by scan service providers, banks, insurance companies, energy providers, public institutions and in the healthcare sector.

Our products are used for various applications in numerous sectors, including:

- **Scan Service Companies**
- **Financial and Professional**
  - Banks
  - Insurance Companies
- **Government**
- **Energy Sector**
- **Health Care**
  - Medical Information System Providers (EHR)
  - Pharmaceutical Companies
  - PACS Vendors
  - Medical Imagery Systems
- **Cultural Heritage**
  - Libraries
  - Museums
  - Archives

Our available solutions are organized in three product lines:

- Document Conversion Solutions
- Imaging Solutions
- Developer Tool Kits

LuraTech's reference customers include scan service providers Arvato (Bertelsmann) and Ratiodata, German health insurance provider DAK, the German state bank of Hesse and Thuringia (Helaba), the Kreissparkasse Ludwigsburg savings bank and other savings banks, the city of Stuttgart and numerous other cities and communities, publisher Heinrich Bauer Verlag, and energy providers Vattenfall, RWE and E.ON. International reference customers include Harvard University, the U.S. Library of Congress, the Dutch Royal Library, the Internet Archive and the U.S. Air Force.

LuraTech strengthens its leadership position with strategic partnerships, such as with ABBYY, and close collaboration with research institutions such as the Technical University of Berlin. LuraTech was a voting delegate for German standards institute DIN for the ISO standardization of JPEG2000 and works for the ISO committee for PDF/A.

LuraTech participates actively in various associations, including the working committee "Standards" and in regional groups of the Association of Organization and Information Systems (VOI). In addition, LuraTech is the initiator and founding member of the PDF/A Compe-

tence Center – a global association with more than 100 members. LuraTech also is a member in the associations AIIM, ARMA, NIRMA and TAWPI and contributes to the consortium for economic administration (Arbeitsgemeinschaft für wirtschaftliche Verwaltung, AWV).

The company headquarters are in Berlin, with additional locations in Remscheid, Germany, and San Jose, California, in the United States and Swindon, UK

Products:

LuraTech's solutions for document conversion enable organizations to optimize both their scanned and digital born documents throughout the conversion process. Users can integrate and centrally manage all steps of document transformation, including image optimization, document structuring and information extraction, such as using OCR or classification. These high-performance products are easy to use and are equipped with our industry leading document compression and popular features, including character recognition (OCR) and batch processing.

Our range of products meets all document conversion requirements:

LuraDocument PDF Compressor:

Easily integrated into document imaging workflow, this product family enables conversion of colour, greyscale or black-and-white scanned documents into high-quality, highly compressed PDF and PDF/A files using our best-in-class, MRC layered compression technology. Our solutions also offer an integrated ABBYY OCR engine for full text search capabilities in all PDF and PDF/A files. PDF Compressor can be enhanced by adding optional modules – such as Born Digital, Forms Recognition, and Digital Signature – while taking advantage of its unparalleled automated, batch conversion.

- **LuraDocument PDF Compressor** is a document conversion engine that can meet much more than the compression needs of a wide variety of organizations. Many of the largest scan service providers use the PDF Compressor to process millions of pages a month.

Thousands of individuals also use the PDF Compressor to occasionally compress and convert to PDF and PDF/A making it easy to e-mail and store scanned documents.

The PDF Compressor is available in three versions to meet both communities' unique needs:

- **PDF Compressor Enterprise** supports production-level batch MRC compression and conversion, integrates the ABBYY FineReader OCR engine to deliver full-text searchable PDF documents, and is extensible with add-on modules including Form Recognition and Data Extraction, Born Digital document conversion and Digital Signature.

- **PDF Compressor Desktop** supports manual document conversion and is ideal for individuals who have occasional compression and PDF/A conversion needs.

- **LuraDocument PDF Compressor for InputAccel:** The PDF Compressor for InputAccel enables InputAccel users to create colour PDF and PDF/A files of high quality that require little storage space. This version of the PDF Compressor Enterprise includes a complete integrated interface for InputAccel. This means that only minor adaptation of the existing InputAccel processes is necessary.

DocYard:

DocYard is LuraTech's new and comprehensive platform for managing custom document conversion workflows, through which all process steps can be integrated. DocYard enables companies and organizations to create a production environment for document processing that can be centrally managed. The modular architecture enables existing components to be integrated in DocYard with little effort and thus allows central management and monitoring of the existing infrastructure. During ongoing operation, workflows and jobs can be generated, retrieved or adapted on a graphical user interface without programming. Users can centrally monitor all current jobs, running distributed in the DocYard system, in real time. Its monitoring and reporting capabilities allow us-

ers to ensure reliable, fast and cost-effective conversion of all documents.

LuraDocument PDF/A:
We offer a number of tools designed to support creation and validation of PDF/A documents.

- **LuraDocument PDF Validator:** This tool tests and certifies that all PDF files meet PDF/A standards, to ensure files are structured properly and the visual appearance will reliably reproduce.

- **LuraDocument PDF Compressor:** Integrating easily into document imaging workflow, this product family enables conversion of colour, greyscale or black-and-white scanned documents into high-quality, highly compressed PDF and PDF/A files using LuraTech's best-in-class MRC technology. They also offer an integrated ABBYY OCR engine for full text search capabilities in all PDF and PDF/A files.

- **Additional Document Solutions:** A comprehensive set of complimentary plug-ins and tool kits enable the implementation of best-in-class document compression and optimization within an organization's document capture workflow and applications.

Users will benefit from:

- **Designed for Bulk Processing, yet Flexible Solution:** LuraTech's solutions for document conversion are so flexible that they can meet almost any requirement for document processing in DMS/ECM departments of companies and scan service providers. LuraTech products have been used successfully for many years for processing document volumes ranging from small to extremely large.

- **Outstanding Document Compression:** Using our award winning mixed raster content (MRC) compression technology, high resolution black-and-white, greyscale and colour documents can be reduced up to 100 times smaller than their original size, while maintaining superior image quality and text legibility. Optimising compression results in lower storage costs and bandwidth requirements.

- **Ease of Integration:** LuraTech's document conversion solutions are especially designed to easily integrate into existing workflows and offer users a variety of models, from comprehensive server-based solutions to tool kits that can be incorporated into customized systems.

- **Standardization:** By building solutions that output standard file formats, PDF, PDF/A and JPEG2000 Part 6 (. JPM), LuraTech's document solutions adhere to the principle of interoperability thus giving customers vendor independence.

- **Strong Technical Partnerships:** LuraTech integrates the world's best technologies to complete its document conversion solutions. For example, the ABBYY OCR engine creates highly accurate, full-text searchable PDF and PDF/A files.

- Outstanding Technical Support: LuraTech provides wide-ranging and timely support for all its solutions, including operation, integration into existing environments, customization and system architecture planning.

Further information is available at www.luratech.com.

SEAL Systems – The Digital Paper Factory

Solutions from SEAL Systems simplify and speed up the generation, administration, and distribution of documents and technical papers.

SEALSYSTEMS
THE DIGITAL PAPER FACTORY

SEAL Systems is the international leading developer of information and document distribution solutions. The company has four offices in Germany. Main subsidiary offices are located in North America, Australia and France. External marketing is directed by partners located all over the world in every major region. Within the fiscal year 2009 the group achieved a turnover of more than 11 Million Euro. As of December 2009, 100 staff members were employed worldwide.

Solutions from SEAL Systems are designed to help organize business processes more quickly, economically, and reliably. Many large enterprises and medium sized companies are using products from SEAL Systems. At present there are more than 1.000 installations in 30 countries worldwide.

Users of solutions from SEAL Systems are located in various industries including automotive, machinery and plant construction, processing, healthcare, manufacturing, aerospace and defense, high tech, supplying, extractive, electrical, ship and railway building.

Product Range

SEAL Systems has four main product lines:

- **Output Management:** Business, Engineering and Network Printing, print and electronic distribution for Office/CAD/PDM/ERP.
- **Solutions for SAP:** Document Input Management, Document Output Management, Conversion server, process-oriented document distribution, SAP Records Management.
- **Conversion:** PDF and TIFF generation, check of PDF files according conformity to PDF/A standard, conversion of graphical file formats, application converters.
- **Direct Publishing:** Automatic generation of documentation, manuals and product information.
- SEAL Systems is market leader in the field of **Document Output Management**.
- **PLOSSYSnetdome** and **gXnetplot** are installed at almost 1.100 locations worldwide as output management solutions for Office, CAD, PLM, DMS and ERP.
- Especially in the scope and implementation of **SAP PLM**, SEAL Systems has gained professional experience in more than 800 projects.
- With SEAL´s **Digital Process Factory** (DPF) processes of information structure and distribution can easily be designed, managed and controlled.
- **PLOSSYS@rchive** – a complete solution for digital document archiving rounds out the offering.

SEAL Systems provides a special platform for enterprise wide print output in all ranges:

- **Business Printing** supplies solutions for printing from ERP applications like SAP to finishing devices of all manufacturers, brands, and models.
- **Engineering Printing** arranges document distribution from CAD, DMS, and PDM/PLM.
- **Network Printing** provides services for unique printer control from desktop, office and in terminal server environments.

Seal Systems Solutions Map for Document Conversions Applications

	SEAL System Product				
Feature	**ConvertWIZ**	**PDF Longlife Suite**	**SAP DMS Conversion Suite**	**SAP BC XDC Conversion Suite**	**gXconvert**
Interactive Tool	✓	✓			
Background Operation		✓	✓	✓	✓
SAP Integration		✓	✓	✓	
PLM/ECM Integration		✓			
PDF & PDF/A Output	✓	✓	✓	✓	✓
TIFF Output	✓		✓	✓	✓
Printable Format Output/PCL & PostScript				✓	✓
PDF/A Quality Assurance		✓			
Graphical Fileformat Output			✓	✓	✓
Graphical Fileformat Input			✓	✓	✓
CAD Input	✓		✓	✓	
Office Input	✓		✓	✓	

- Alternatively to paper output, the documents can be distributed also electronically by **Direct Publishing** solutions from SEAL Systems: as file, e-mail, fax, CD/DVD, or automated compositions of complete manuals and documentation.

PDF/A and Conversion

SEAL Systems is the leading supplier of conversion tools for users and integrators and has developed solutions for every environment in this range.

PDF Longlife Suite

PDF/A check and adjust to standard

- PDF Longlife Suite from SEAL Systems automatically evaluates PDF files for compliance with ISO standards (PDF Checker).
- If required, it can generate ISO compliant files of configurable quality (PDF Adjust).
- PDF Longlife Suite can be used interactively and by server operations.

PDF Longlife Suite – SAP Integration

With the SAP DMS Integration of the PDF Longlife Suite, the ways of archiving a PDF file within SAP can be controlled.

Check of existing data

According to configuration and to difference of the file to the selectable PDF standard, the process generates

- an adapted PDF file
- an error code
- a detailed error message file in TXT format

The corrected files can instead of or additionally to the original PDF file be assigned to the appropriate DIS. Altogether a log file is created about all checked files. This file can be used as a statistics basis.

A transaction for the interactive selection of documents, which are to be proofed and adapted, can be offered optionally (product code DMS-XSA).

A 30-days-trial-version of the PDF Longlife Suite is available for download on **www.pdfforever.de**.

About the Authors

Hans Bärfuss,

Vice-Chairman of the
PDF/A Competence Center;
PDF Tools AG, Winkel, Switzerland
www.pdf-tools.com

Hans Bärfuss is the CEO of PDF Tools AG, an internationally active software development and marketing company. He studied Electrical Engineering at the Swiss Federal Institute of Technology in Zurich (ETH), earned a doctorate in Computer Science and completed a postgraduate study in Business Administration at the University of St. Gallen. Since 1985 he has founded and developed successful pioneer enterprises including: GLANCE AG, CCS Creative Computer Software AG, and Medica Medizinal-Informatik AG among others.

Since the mid-nineties Dr. Hans Bärfuss, an internationally active PDF expert and member of the ISO working group for the standardization of PDF/A, has dealt with PDF technologies. He founded PDF Tools AG in 2001 in order to offer high quality and superior performance PDF products. Within the scope of the PDF/A Competence Center, he strives for a uniform interpretation and implementation of the standard in order that PDF documents can be better exchanged and archived.

Domenico Barile,

E-Mission S.r.l, S. Maria di Zevio, VR, Italy
www.emission.it

Domenico Barile is 39 years old, married with 1 child, and holds a degree in computer science. His thesis was "The automation of passive cycles using document optical recognition". He has been with Archiva S.r.l. since April 2005 as development team leader, heading the team that enables Archiva to be autonomous in business development. He is responsible for creating and maintaining all software for optical document storing, electronic document creating and transferring. Actually his main goal is to integrate standards for data interchange, data representing and handling for business language. Domenico Barile was one of the architects of the E-Mission structure and distribution services. He was also the driving force behind Weaver, which was created based on his past experiences in production and industrial processes management, transfer standards analysis, business consulting and freeform recognition. He currently leads the development and consulting teams for both companies.

E-Mission S.r.l. started its business in October 2008 as a split of Archiva S.r.l., which has been on the market since 1979 with substitutive optical storage of documents. All document distribution services were entrusted to E-Mission with the target of completely automating and improving them. Currently the company handles the transmission of documents of every type and nature (commercial, tax, administrative, technical, etc.) in hardcopy and through electronic mailing services. With Weaver it has completed its range of services being offered with the transmission of the data contained in these documents, creating a new transmission engine open to all new technologies and all present and future communications standards (XML, EDI, etc.).

Alessandro Beltrami,

Consultant, TechnoSolutions srl, Italy
www.technosolutions.it

Alessandro Beltrami, a prepress consultant for TechnoSolutions srl,

was one of the pioneers of multimedia communication technologies, beginning in the early 90's first with. He worked as a software architect and developer on the first distributed multimedia catalogues that combined images and documents stored on CD. His in-depth knowledge of colorimetry has been applied both in improving the colour seen on screen as well as for producing image processing software. In 2004 he began working as a consultant, moving from software and industry to graphic arts. In recent years he has managed numerous projects in document management, colour standardisation, prepress workflows, soft-proof systems and restoration of digital archives. He is the founder of cmyQ: the first ISO 12647 certification project in Italy. Alessandro Beltrami is also a member of ISO TC130, an advisor to TAGA Italy (Technical Association of the Graphic Arts), a consultant for Associazione delle Arti Grafiche di Bologna, and acts as technology evangelist at seminars and courses

Raffaele Bernardinello,

CMT Group, Roma, Italy

www.gruppocmtrading.it

Raffaele Bernardinello is 45 years old, married with 1 child and holds a degree in Computer Science. Mr. Bernardinello has been working in the document management field since 2001. At Postel he was responsible for the Business Unit named "Printel", with the main focus on Web-to-Print solutions. He was the project leader of the first PDF-based "Print on Demand" platform and also for the first PDF-based platform for mass printing. In 2005 he became responsible for the production platform and development of Postel, including the solutions for printing and archiving.

In 2007 Mr. Bernardinello was appointed CTO of the CMT Group, responsible for all Group technologies. In this role he founded PostaJet srl., a new company working in the hybrid mail area for small and medium entities. Many solutions of the CMT Group and of PostaJet are based on PDF workflows. Raffaele Bernardinello was appointed CEO of PostaJet in 2009, a position which he currently holds.

In 2010 Mr. Bernadinello helped found the Italian Chapter of the PDF/A Competence Center. All PDF-based workflows in the CMT group have now been migrated to PDF/A.

Manuel Brunner,

Head of Projects and Services,
Intrafind AG, Planegg, Germany

www.intrafind.de

Manuel Brunner, head of projects & services at IntraFind AG, is a specialist in search and retrieval technologies. He supports numerous customer projects in project management, conception and planning activities. Manuel Brunner is an IPMA certified project manager.

The IntraFind Software AG specialises in software solutions for knowledge retrieval and text mining.

The company is the specialist for information retrieval in unstructured and structured data. It offers solutions for searching and finding information, and processing it with methods and procedures consisting of a combination of linguistic and associative-semantic methods, using state-of-the-art information theory techniques.

Olaf Drümmer,

Board, PDF/A Competence Center.
callas software and axaio software,
Berlin, Germany

www.callassoftware.com

Olaf Drümmer, president of both callas software GmbH and axaio software GmbH in Berlin, has been involved in the print, publishing and document management industries since 1990. He is recognized worldwide as an expert in PDF and color management. As a participant in the ISO he was heavily involved in the standardization of PDF/X and PDF/A. Furthermore, Olaf Drümmer is chairman of the European Color Initiative (ECI), which has ensured the continual applica-

tion of ISO standards for color management and printouts in daily business far beyond the European borders since 1996.

François Fernandes,

levigo solutions GmbH,
Holzgerlingen, Germany

www.levigo.de

François Fernandes studied computer science and economics at the Reutlingen University. Then he started as a developer and consultant at the levigo solutions GmbH. He is responsible for the PDF support of the jadice document platform. Instead of the typical "Hello World" program, he learned the Java programming language by writing a PDF parser.

Harald Grumser,

Chairman of the PDF/A Competence Center;
CEO Compart AG, Böblingen, Germany

www.compart.com

Harald Grumser is one of the co-founders of Compart AG and has been CEO of the company since its inception in 1992.

His move to a career in IT first began in 1984 following his time at the University of Karlsruhe, where he read Physics. In the first few years after leaving university, he worked both in software development and publishing and also wrote a number of specialist magazine articles on the relatively new subject area of PCs and microprocessors. He then moved to IBM, where he managed a Europe-wide PC project for several years before becoming PC application manager for a Böblingen-based software company and subsequently co-founding Compart.

As CEO of the PDF/A Competence Center, Harald Grumser is responsible for overall coordination of the association's activities and the acquisition of new members.

Rolf Günter,

Head of Business Development,
Sales & Marketing,
PDF Tools AG,
Winkel, Switzerland

www.pdf-tools.com

Rolf Günter is Head Business Development, Sales and Marketing of PDF Tools AG. He studied at the University of St. Gallen and Bern and is Attorney-at-Law. He completed a postgraduate study for his Executive MBA at the University of Rochester (New York) and Bern. After his studies he started his career in the Insurance Industry in Switzerland in several positions and worked as an Attorney-at-Law and as a Project Manager. Starting to work self-employed, he was founding partner of reflecta ag, a specialist for legal, management and project consulting and co-owner of a medium law firm in Bern.

After some years, he get the chance to become CEO of the Protekta Legal Protection Insurance AG, in which he worked several years before changing to the working place Zurich, where he started his own business as a legal and management consultant. Records Management from a legal and an organizational point of view was one of his special services to medium and big enterprises in Switzerland.

Engaged in a bigger project in the biggest Health Care Insurance Company in Switzerland, he get the offer to join the company and to build up the Key Accounting. Since September 2009, Rolf Günter join PDF Tools AG as a Board Member.

Dominique Hermans,

Owner of DO Consultancy,
Landgraaf, Netherlands

www.doconsultancy.com

Dominique Hermans is owner and manager of the consulting and

training organisation DO Consultancy which was established in 2008. Dominique studied Library and Information Science in Maastricht (Netherlands), graduating in 1995. He is fully active in the records management field, having worked for several years as an information specialist and manager of a library and company archive. For the past couple of years Dominique has trained numerous people in subjects like PDF/A and digital longevity. In addition to training and consulting on records management, DO Consultancy has organised several successful seminars in the Netherlands on digital archiving.

Johannes Hesel,

**PDF/A Competence Center,
Member of the Board;
SEAL Systems,
Roßdorf, Germany**

www.sealsystems.com

Johannes Hesel is Vice President for Business Development in the executive board of SEAL Systems. After finishing his studies in computer science with a focus in computer graphics, he was working in several positions at GTS-GRAL and SEAL Systems. He was mainly concentrating in the development and consulting of conversion and output management solutions for technical documents. He is an expert for document management solutions for technical documents, especially for SAP DMS and long term archiving.

In the board of the PDF/A Competence Center Johannes Hesel is the sponsor for the activities in France and he supports the marketing working group.

David Hook,

**Director Product Management,
Crawford Technologies,
Canada**

www.crawfordtech.com

Dave Hook is the Director of Product Management for Crawford Technologies and has been in the transactional customer communications document industry for 23 years. He has held a broad range of management positions in organizations such as Xerox Canada Ltd., Symcor Inc., and Davis + Henderson. He has led the introduction of many products such as check image statements, electronic statement archives, statement composition services, Braille and large format printing services, statement marketing (transpromo), and many others. Through these initiatives he has gained a broad range of experience within the transactional printing industry both as a vendor and as a services provider to the banking, telecommunications, loyalty, retail and other industries.

Dave sits on the US PDF/UA committee as well as various other standards committees. Dave speaks at XPLOR, (an international association that promotes education and networking for its members) and has been an invited speaker to other user and vendor conferences to discuss industry trends and developments. As Director of Product Management, Dave is responsible for directing and developing the portfolio of software products and solutions for Crawford Technologies Inc.

Dave is married and has 5 children.

Hans-Joachim Hübner,

**Satz-Rechen-Zentrum Hartmann+Heenemann GmbH & Co. KG,
Berlin, Germany**

www.srz.de

Hans-Joachim Hübner is the Head of the System Integration and Sales division at Berlin-based ECM and digitalization specialist Satz-Rechen-Zentrum (SRZ). Mr. Hübner has more than 20 years of experience in the successful execution of projects relating to the digitalization of large document collections, the design and realization of capturing solutions, and workflows ranging from the delivery of documents from libraries to mass digitalization with high-performance scanners in the field of ECM. His specialist areas also include the presentation of digital collections on the web and the creation of archive formats for long-

term archiving. He is a major player in the shaping and promotion of software development at SRZ in the fields of capturing, ECM, and content management.

Carsten Heiermann,

Managing Director, LuraTech Europe GmbH, Berlin, Germany

www.luratech.com

Carsten Heiermann is a shareholder and executive director of the LuraTech group which has headquarters in Berlin, Germany and a branch in San Jose, USA. After studying communications engineering and undertaking work in various IT companies, since 1995, Carsten Heiermann has concerned himself with topics connected with compression and standardization.

Duff Johnson,

CEO Appligent Document Solutions, United States of America

www.appligent.com

Duff Johnson is the CEO of Appligent Document Solutions, based in Lansdowne, Pennsylvania in the USA. The company invented PDF redaction and form-flattening and was first to market with a wide variety of PDF-specific server applications, all available on the leading server OS platforms.

Duff Johnson founded the first business document service bureau for PDF files in 1996. Now part of Appligent Document Solutions, the company offers PDF forms, accessibility, automation and imaging services, among others. With dozens of articles on electronic document management subjects, Duff Johnson serves in leadership roles in PDF International Standards development. He is Vice Chair of the US TAG for ISO 32000 (PDF), Chair of the US TAG for ISO/CD 14289 (PDF/UA) and a member of the US TAG for ISO 19005 (PDF/A). Johnson serves the PDF/A Competence Center as Chair of the North American chapter.

Stephen P. Levenson,

US District Courts and Convener of the ISO PDF/A Committee, United States of America

www.uscourts.gov

Stephen resides in Virginia outside of Washington D.C. where he works for the U.S. Federal Judiciary. He began work for the Judiciary after ten years in the State Courts in Florida. He has been involved in large microfilm, automation, records and Internet projects.

Stephen's current role with the Judiciary includes his work with policy issues regarding electronic filing and long-term record retention. Stephen is the International Conveyor for PDF/A, Member of AIIM Board of Directors, Chair of AIIM Standards Board and AIIM Educational Advisory Committee.

Local Washington groups include Federal Information and Records Management (FIRM) Council and the Capitol Chapter of AIIM.

Stephen is also involved in many local civic and professional associations, including his role as president of both the Springvale Homeowners Association and the Central Springfield Redevelopment Authority.

Natascha Schumann,

nestor – Kompetenznetzwerk Langzeitarchivierung, Frankfurt am Main, Germany

www.langzeitarchivierung.de

Natascha Schumann is currently coordinator in nestor, the German network of expertise in digital long-term preservation. She has a degree in social sciences and further training as an information specialist.

Since November 2005 she is affiliate at the German National Library.

Key activities in electronic publications, particularly in electronic theses and dissertations within the coordination agency DissOnline.

From 2002 until 2005 she was affiliated at Darmstadt University of Technology, department of Sociology. She was involved in the development of a portal for social sciences resources which were freely available via the Internet.

Dr. Uwe Wächter,

SEAL Systems AG,
Roßdorf, Germany

www.sealsystems.com

Dr. rer. nat. Uwe Wächter studied Physics at the Technical University of Rostock and finished with his PhD in 1990. From 1990 to 1996 he was working as external consultant in the software development department from Siemens PG.

Since 1996 Dr. Uwe Wächter is working for SEAL Systems AG. As product manager for the PDF solutions he was involved into PDF/A long before the official norm was published. Dr. Wächter is an admitted expert for PDF/A. As an active member of the PDF/A Competence Center he is doing presentations about PDF/A since many years.

Dr. Bernd Wild,

PDF/A Competence Center,
Member of the Board;
intarsys consulting GmbH,
Karlsruhe, Germany

www.intarsys.de

Dr. Bernd Wild is originally a graduate physicist. After completing his studies, his worked for several years at a computer science research center in the field of artificial intelligence and its applications in industrial processes.

Upon obtaining his PhD, Bernd Wild was responsible for the organization and management of C/S software development at an IT service provider in the banking sector.

Together with some partners, Dr. Wild founded intarsys consulting GmbH in Karlsruhe in 1996. He now concentrates on consulting and providing assistance for complex system integration projects. Document technology has increasingly become a focus point during the past few years. This includes not only the creation of documents from source data, but also the entire documentation life cycle through to archiving. Technologies like electronic signatures, intelligent forms and document standards are at the core of his activities. In addition, intarsys offers products and software components like a BSI-certified electronic signature solution which allows for standard-compliant signing of PDF/A documents.

Thomas Zellmann,

PDF/A Competence Center,
Managing Director;
LuraTech Europe GmbH,
Berlin, Germany

www.luratech.com

Thomas Zellmann has been working in EDP for almost 20 years and has extensive experience with classic and modern IT solutions. He started his job at LuraTech in 2001. Prior to joining LuraTech he worked for Softmatic AG, Software AG and Nixdorf among others.

Thomas Zellmann is working in the banking/insurance and archives/libraries segments and is one of LuraTech's shareholders.

As board member he coordinates all marketing activities of the PDF/A Competence Center. His work focuses on visibility of the PDF/A Competence Center through marketing, public relations, Internet presence and seminars.

About the PDF/A Competence Center

Association for Digital Document Standards – ADDS

The PDF/A Competence Center is an initiative of the Association for Digital Document Standards (ADDS) e.V., founded in September 2006. A particularly important aim of the association is to promote the exchange of information and experience in the area of long-term archiving in accordance with ISO 19005 (PDF/A).

The ISO standard for long-term archiving, PDF/A, is generating considerable interest in the market. In order to encourage the high demand for information and exchange of ideas concerning PDF/A, callas software GmbH, Compart AG, LuraTech Europe GmbH, PDF Tools AG and PDFlib GmbH have founded the PDF/A Competence Center.

The executive chairman is Harald Grumser, who founded, together with other shareholders, Compart AG in 1992 and has headed it since its inception. Dr. Hans Baerfuss, CEO of PDF Tools AG, Switzerland, is the executive vice-chairman.

The association is geared towards developers of PDF solutions, companies that work with PDF/A in the area of DMS/ECM, interested individuals, and also users who want to implement PDF/A in their organizations. Although the months directly after the founding saw new members predominantly from German speaking regions, the executive committee has expanded their activities internationally beginning in 2007. Country Chapters for several regions – at present BeNeLux, France, Germany, Italy and the US – are established to simplify the contact with interested companies and users.

Interested parties can thus benefit from the combined knowledge of competent PDF/A suppliers. The association offers numerous services including conducting events, working on further standardizations and serving

as a central competent point of contact for answering all questions about PDF/A.

Work on the ISO Standard

Several members of the PDF/A Competence Center are technically oriented and actively participate in the development of the PDF/A standard as members of the responsible ISO committee (ISO TC 171 – Document management applications).

PDF/A-2 was technically approved this summer in Paris and will be published early 2011 by the International Organization for Standardization. In Paris the committee already laid groundwork for a further part to the standard, PDF/A-3.

Events around the PDF/A Standard

In order to meet the high informational needs around PDF/A in the market, the PDF/A Competence Center organizes seminars and events in different locations on a regular basis.

For details about current activities, please check the Events page at **pdfa.org** on the Internet.