White Paper: Securely archiving emails in the PDF/A format

25 years ago, Germany's first email was sent. At the time, few could guess that electronic messaging would become one of the world's most essential communications media. Even in the world of business, email has established itself as a fast and cost-effective means of exchanging information. As a result, more and more quotes, purchase orders, invoices and other business-relevant information are being sent electronically. However, legal requirements mean that these documents must be retained for long periods of time. This White Paper will discuss how you can use the PDF/A format to securely archive emails for the long term. It will discuss the various possible conversion scenarios and explain how to resolve any problems which may arise or already exist. This document will also cover background information, follow-up recommendations and a description of how you can use the LuraTech PDF Compressor to convert emails to PDF/A.

Contents:

Introduction: Archiving and Management

It is important to distinguish between email archiving and email management. Archiving can follow a hardware-based approach in which the complete data stream of all incoming and outgoing emails is stored. This will include spam and other junk email, which is why it is advisable to convert emails related to your business and your finances into a stable long-term format which is then used to archive them.

An email management system affects the entire organization and is often set up in a business' central mailroom or integrated into an existing DMS or ECM system. Intelligent search tools will make it easy to find relevant content, and predefined settings will specify how to handle spam or private emails. However, email management says nothing else about how to securely archive electronic messages. It is not enough to simply store them in your DMS or ECM system.

Legal Background*

"Which emails must I, or should I archive, and which should I not?" This is one of the first questions a business needs to ask itself when considering an email archiving system. The answer depends on a business' individual needs and is made up of a mixture of legal requirements and a business' own email archiving interests beyond what is legally required. For reasons of reliability, the VOI – Verband Organisations- und Informationssysteme (e.V.) – believes that archiving all emails is basically the best option. However, it is also important to take account of legal requirements for archiving specific emails – for data protection reasons, for example. Companies must also maintain confidentiality in line with telecommunication privacy laws (such as the Telekommunikationsgesetz or TKG in Germany) when handling private emails from employees as permitted by the employer. As for retention periods, emails are not fundamentally different from other documents: business law requires all traders to retain their business-related communications (all written content - including emails - relating to a business transaction, regardless of what form it was sent in) for six years. A message is related to a business transaction if its content deals with the transaction's preparation, closure, execution or cancellation. The retention periods also arise from regulations set by tax law and, again, apply to all business-related messages. While freelancers who are not technically defined as traders are theoretically unaffected by the legal data retention requirements described above, they are still fully subject to the corresponding tax regulations. This means that they, too, must still archive all their business-related correspondence and therefore all applicable emails. The retention period is six years here, too. However, tax regulations specify that retention periods do not expire as long as the documents are relevant to any taxes for which the appointment period has not yet run out. This means that in some cases the retention period can be longer than the six years specified here.

*Reproduced with the kind permission of the VOI – Verband Organisations- und Informationssysteme e. V., Bonn

Challenges

It is recommended that companies define which emails to archive, and in which format. As described above, this can be a data stream of all incoming and outgoing emails, or only business-relevant ones. However, if they choose to store only the data stream– all emails and attachments in their original format, in other words – then the long-term readability of the content is by no means guaranteed. This is because neither email formats such as .eml or .msg, nor attachments created by application software such as .doc, .xls or others, are suitable for archiving. It is highly questionable whether appropriate viewers for these files will still be available years from now. There is also no guarantee that emails ten years from now will be displayed in exactly the same manner they are received and sent today.

The Solution: A Dual Strategy

To meet email integrity, authenticity and readability requirements, businesses are recommended to archive both the complete data stream of all emails and specifically business-relevant information in the form of a "digital printout". This involves converting each email and its attachments to PDF/A. It makes sense to carry out this step promptly, so that if one or more attachments contain errors, or the format is unreadable, the sender can be asked to resend the attachment in another format.

PDF/A – The ISO Standard for Long-Term Archiving

PDF/A, ISO standard 19005 for long-term archiving using the PDF format, has found broad acceptance worldwide. The standard assesses and defines which PDF functions are considered suitable for archiving purposes. These specifications guarantee long-term document readability – no matter what application software or operating system was originally used to create them.

There are now three available versions of the PDF/A format. The most recently published version only has one more feature than PDF/A-2: the ability to embed non-PDF/A-compliant files such as CSV, MSG or Office files as attachments. This addition turns PDF/A into a "container" for any file type, which should be of significant value to any organization which archives its emails.

This is because emails and attachments can be bundled together in both PDF/A format and their native format, all in one file. In practice, this means that a PDF/A-3 file consists of an archive and a non-archive component. Each has different requirements: the archive component must be retained for a long time, which means it must meet the well-known conventions of PDF/A. Its life cycle is therefore a very long one. The embedded file, on the other hand, fulfils a different role: it contains additional information which will only be relevant for a specific period of time – namely, for as long as it takes to process or edit the information. Once processed, the embedded file is no longer needed. The result is a complete archived email which stores everything in one place.

Email Archiving with PDF/A

As previously discussed, three different versions of the PDF/A standard are available, each of which offers different options for embedding files.

The first version of the standard does not allow attachments within a PDF/A file. This means that emails containing any attachments must be converted to PDF/A, with the attachments inserted as a linear series of pages in a multi-page PDF.

The PDF/A-2 format allows attachments to be embedded as PDF/A files. These attachments can then be viewed by clicking on the Attachment symbol within the PDF viewer.

When converting to PDF/A-3, a multi-page PDF file will be created, and attachments will be embedded in their native format. The viewer will then show the email and its attachments as a linear series of pages. The PDFs can be found under Bookmarks, while the attachments can be found under the Attachment symbol, displayed just as they were in the original email.

Conversion Scenarios

There are a range of possible scenarios for converting emails. These are primarily differentiated by the level of automation, the level of user involvement, and the quality assurance (QA) stage.

Server-side conversion will achieve the highest level of automation. In this scenario, every email, including all attachments, will be converted to PDF/A. The tools used must be able to process a large volume of emails. Experience has shown that – depending on the specific application – automatic conversion rates of 95 to over 99 per cent can be reached.

There are two approaches to client-side conversion: direct conversion on the client machine, or conversion initiated by the client but performed on the server. Either scenario may or may not include a QA stage.

For direct client-side conversion, the client manually launches a conversion engine implemented as an add-in for their email software. A background process then hands over the PDF/A file to the archive interface. Any conversion errors are corrected there and then by the user. As this process is highly processor-intensive, low-end desktop computers will experience significant load. As a rule, therefore, it is necessary for the user to wait until the process is complete. Any subsequent QA is done using a GUI component. The user works with the GUI to visually inspect the PDF/A file and then release it for archiving, directly eliminating any errors. This does increase the user’s workload, however. One fundamental disadvantage of client-side conversion is that PDF/A files are produced on a distributed basis, making it difficult to ensure file conformity. One solution would be to implement a downstream server-based validation process, although this would of course cause additional overhead.

Alternatively, the user can trigger conversion using an add-in for their client, handing the conversion job itself over to the server. A quick test should be performed beforehand to ensure that the material can actually be converted. As with the purely server-side scenario, this approach converts emails using powerful, scalable, centralized tools. The user can then also perform a QA check if desired. As the conversion process is very complex and therefore time-intensive, precautions for asynchronous processing should be taken so that the user receives a response regarding the conversion job.

Server-side conversion is the better option for most projects. It is a highly scalable approach and runs in a secure environment, making it unnecessary to validate PDF/A compliance on a document-by-document level. When choosing a product, it is important to ensure that your conversion tool of choice can do the following:

- Log of each stage of processing (e.g. in line with the GoBS protocol),
- A quick test for input material,
- Assign input material to the correct stage of PDF/A processing,
- Convert PDF to PDF/A,
- Compress and convert images and scans, apply OCR where necessary,
- Validate incoming PDF/A files,
- If required, unpack attachments (ZIP, 7-ZIP, rar, …)
- Convert “born digital” documents (Word, Excel, email header & body …) to PDF/A
- Configure settings for handling exceptions and errors, e.g. videos or signed input material,
- Interface with email systems and DMS, ECM or BPM systems.

Policies

As a rule, businesses tend to encounter a wide range of file formats: varying types of office documents, PDFs of all kinds, scanned documents, CAD files, industry-specific formats or perhaps even PDF/A files that still need to be checked to see if they actually do comply with the ISO standard. In other words, businesses face an all-but-irrepressible “format menagerie”. However, experience has shown that recommendations to business partners regarding which file formats will be accepted can actually curtail this proliferation of format types. Additional guidelines, such as requests not to password-protect PDF files, will also reduce the number of conversion problems encountered.

Troubleshooting

Problems most often arise during email conversion when PDF files are password-protected on some level. Poor quality can also hinder conversion. Typical examples include missing fonts or unacceptable content such as videos or JavaScript. To avoid interrupting the entire process, it is important to define how the system should proceed in such cases. For example, you can instruct it to replace a missing font with another available one. Forbidden content in a document can be replaced with an image before converting the file to PDF/A, and a video can be stored as a separate file in its original format. The more rules you define, the closer you will get to one hundred per cent automation. If any documents crop up which still cannot be converted to PDF/A even after taking every rule into account, an administrator can be consulted or the document can be returned to its sender. Experience has shown that the learning curve for projects such as these tends to be rapid, and the number of files which can be automatically converted to PDF/A will continue to rise rapidly over time.

Email Archiving with the PDF Compressor

With Version 7.1 of our PDF Compressor, the production-oriented application for compressing files and converting them to PDF(/A) with optical character recognition (OCR), you can easily convert your emails and any attachments to PDF/A, ensuring their long-term readability. To do so, the software will automatically select the best possible processing mode for each file. This means emails (including a range of attachments), scanned documents and born-digital files can be processed in a single step. There is no need to install or run Microsoft Outlook to do so. This is important because Outlook can only run in one instance at a time, making parallel processing of multiple emails impossible. In addition, the PDF Compressor converts email files in .eml format into the ISO standard for long-term archiving. Additional functions, such as compression, conversion and OCR extraction for scanned files, expand the software's range of potential applications.

Converting emails to PDF/A takes only a few steps:

- Store your emails in MSG or EML format in a predefined folder which is regularly monitored by the PDF Compressor.
- Specify what to do with the file after the conversion process, such as the location to save it to. Select the PDF format to be created.
- Optionally, you can choose to exclude certain formats, such as image or audio files, from being embedded in a PDF/A-3-format file. It is also possible to select formats to always or never convert to PDF.
- Finally, start the job, and the PDF Compressor will securely and reliably convert your emails to the desired format.

Our PDF Compressor is available in three different variants:

Basic	**Advanced**	**Server**
• One-time processing for a fixed number of pages • Page quotas (cartridges) from 100,000 pages to several million available	• Recurring quota of pages per year	• License with no time or volume restrictions
• Use all CPU cores on a workstation for fastest possible processing times	• Use all CPU cores on a workstation for fastest possible processing times	• Use the total number of licensed CPU cores
Ideal for: • Project-based document processing • Migration projects • Converting old paper archives • Clearing production backlogs	**Ideal for:** • Ongoing small and medium-sized projects • Invoice processing • Ongoing document compression projects	**Ideal for:** • Scan service providers, businesses and organizations looking to invest just once in their document conversion workflow in order to reduce costs significantly
Information No time limit on license. Rechargeable at any time, compatible with latest software version.	**Information** If a company comes close to exceeding its annual page quota, they can combine the Advanced model with an additional Basic model in reserve.	**Information** By licensing additional CPU cores, the PDF Compressor Server can process millions of pages every month.

About LuraTech:

LuraTech is one of the leading providers of ISO-compliant document and image compression solutions for scan service providers, businesses and a wide range of organizations and institutions. LuraTech delivers software, services and outstanding support for converting digital documents worldwide. Founded in 1995, the company has four offices across Germany, England and the USA.

Contact:

Tel: +49 2191 58960-0

Fax: +49 2191 58960-29

info@luratech.com

www.luratech.com