Readable data for many decades

Egg, January 29, 2025: *When Archivista GmbH first saw the light of day in 1998, nobody could have imagined the speed at which technology would develop. The only thing that was clear was that conventional paper filing was not such a good idea even back then, because handling analog documents was/is expensive, cumbersome and error-prone. Today, it should be noted: Paper documents have not completely disappeared, but digital data has largely replaced the analog world. However, digital data, the digital world, does not come for free. In this blog, this will be demonstrated using the example of emails with the new 2025/II version of ArchivistaBox.*

Not quite so trivial: emails as a replacement for letters

In digital correspondence, e-mail messages have become almost 100 percent accepted as a substitute for letters. And it may well be that e-mail messages currently seem outdated again. But whenever something needs to be recorded in writing, the e-mail message is used.

This also means that current e-mail messages are an important factor when it comes to keeping digital information securely available for the long term.

Mail archiving à la ArchivistaBox

Most solutions on the market are “content” with retrieving mail messages from the provider and saving them in their original format. The readability of the data depends on the ability to display the mails with current on-board tools. The rules of the game can change with every update. What is readable today may no longer be able to be displayed correctly tomorrow or even many years later.

ArchivistaBox is different. Here, all data is photographed virtually by default. This means that once recorded, mail messages can be read visually in the long term.

Such an approach may come as a surprise, but after almost three decades of dealing with scanned and digital data, it is worth noting that there are minor problems with the legibility of image files compared to information stored elsewhere (Office, PDF and/or text data) at a ratio of around 1:1000.

Image files consist of dots that are drawn as pixels on the screen. With all other formats, the information has to be “drawn” on the output device each time. This should not be so difficult, but it is. To illustrate why this is the case, the (simplified) structure of mail messages is briefly described.

What are mail messages?

In principle, mail messages correspond to simple text files:

```
From: xxx@mail.form
To: yyy@mail.to
Subject: Mail

Hello, I am an e-mail. It's more difficult with the
greetings...
Even more difficult, by the way, when it comes to
attachments (images, PDF files).
------=_20081201100242_51266
Content-Type: image/x-portable-anymap; name="eins1.pnm"
Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename="eins1.pnm"

/9j/hADbFiAAGBwYHCAUJBwaJiAiNFAwLCwwRmIwUDpKemZ0ZnJ4gG5wnLiQ
roiAcG6KotqgxL6u
ztDO4pp8yODyyrjwAcbOJCQiMCowNDRehMZexoRwxsbGxsbGxsbGxsbGxsbG
xsbGxsbGxsbGxsbG
------=_20081201100242_51266--
```

The simplified example here serves to give an “approximate” impression of a mail message. In a narrower sense, mail messages consist of three parts: 1) header, 2) text and 3) attachments.

1) Header

This contains the information from and to whom the message is to be sent. In order for a mail to get from the sender to the recipient, it is sent via providers (intermediaries). These usually leave a stamp in the form of a line of text:

```
Received: from ps15zhb (localhost [127.0.0.1])
by ps15zhb.bluewin.ch (Postfix) with ESMTP id B62575C0
```

This means that an email that is sent by the sender undergoes changes (at least) in the header section on its way to the recipient. This information is important in order to determine (later) whether a mail message took the usual route or whether it is almost certainly a “forgery”.

Side note: *Signatures and encryption can also help to improve or guarantee authenticity and data integrity. However, this does not change the fact that mail messages “travel” from the sender to the recipient via many points and the* *above rules do not apply.*

2) text part

This is where the actual message is located. What sounds simple has its pitfalls, because display problems can occur at the latest when it comes to displaying umlauts and special characters. This is why the desired character set is often specified. However, it remains unclear whether this information is correct and whether the recipient can do anything with it.

The matter is complicated by the fact that special characters can also occur in the header data or that these can contain different character sets from the main part.

Another challenge is that the main part can also exist in different variants. For example, in pure text form and further formatted (usually as web page fragments). Of course, an email is not a fully-fledged website and often it is just a matter of marking text passages in a special way (font, size or highlighting (e.g. bold). In this sense, there are almost no limits, but also almost no standards.

3) attachments

The attachments are stored with start and end markers at the end of the message. There are standards as to what the start and end sequence looks like, but not what data is saved in between.

The attachments often contain images and PDF files, but they can also contain (malicious) programs. Attachments from an Office manufacturer that run programs more or less without being asked are still particularly notorious. Irrespective of the problem of attachments that want to “harm” the recipient, the question arises as to what to do with the attachments when they are entered into a document management system (DMS) such as ArchivistaBox. Often the text part of the mail itself does not contain the relevant information (e.g. orders as PDF files).

With the ArchivistaBox or mail archiving, it can be explicitly defined whether the attachments should be “photographed” (and also included in the full text) or not and if so, for which files this should be the case.

4) special cases

There is now a set of rules for mails (rfc822). Whether and how these standards are implemented is a matter for the providers. Larger players in particular sometimes interpret or extend the specification in such a way that others are left behind (at least in the short term). To put it more positively, we can talk about special cases. Some of these “special cases” are listed here.

Mail messages can advertise themselves as attachments and at some point the giant from Redmond started to send attachments in its own format (winmail.dat).

Mail files can contain links to external websites, e.g. to retrieve additional content (usually images, but not only) from there. This form is currently used frequently, as it makes it easy to determine whether and where messages arrive or are opened. In addition, content can be customized depending on the recipient. Example: A price “disguised” as an external link in an email, for example, is “spit out” differently depending on the calling IP address.

It is also possible that mail attachments may be “broken”, as the conversion of the original files to ASCII-7Bit (or reverse conversion when reading) works without error detection. A single “swallowed” (incorrect) character can mean that the file cannot be read by the recipient. The same problem naturally also exists if, for example, an image file is already “defective” before transmission.

The new ArchivistaBox 2025/II mail archiving solution

ArchivistaBox mail archiving has been available for around 15 years. Back then, mail messages usually only contained short attachments and the graphic design was kept simple (if at all). Perhaps a company logo or a small PDF file, a couple of photos – that was all that was sent by email back then.

About eight years ago, it became apparent that email messages were much more colorful. Mail archiving was therefore optimized for HTML mails. In 2023, mail archiving for Office365 appeared because the IMAP standard had to be deactivated for these services. With version 2025/II, mail archiving is being expanded again. The following key points are new:

Handling links that no longer exist

Email messages with external links are being sent more and more frequently. In the meantime, external content is being deactivated more and more quickly. With the new mail archiving in 2025/II, you can choose whether a) external links are ignored, b) they are taken into account or c) the mails are processed with/without links. Optional means: Links are used, but only if they deliver results within a reasonable time.

Handling winmail.dat

Previously, these attachments were saved as corresponding file snippets. This means that the actual contents of the Winmail fragments were neither virtually photographed nor prepared for full-text recognition. They are now unpacked and the corresponding content is captured correctly.

Nested (cascaded) mails

Mail messages can themselves appear as attachments as mail messages. Previously, mail messages that occurred as attachments were not processed. These are now recorded as far as is reasonable (maximum of nine levels).

Check for many (un)known file types

The new mail archiving system has been tested with tens of thousands of mail messages that have been received by Archivista GmbH since it was founded. Many new file types previously unknown to mail archiving have been added. Of course, unwanted file types still cannot be completely ruled out. However, the ArchivistaBox 2025/II recognizes many file formats in the basic configuration that were previously not specifically or optimally processed.

Check for incorrect data

With version 2025/II, attachments are better checked for incorrect transmission or corresponding origin. The previous check program recognized the data as correct, which meant that the import process started but failed in the end. The integrity of the data is now checked better and the import job is not even started if the data is incorrect.

This raises the question of how to deal with incorrect data. Should the transmitted data be “cleaned” or should the original remain unchanged, even if the content is incorrect? Or, more heretically, should any viruses be “backed up” in the long term?

Original data is of such central importance for a later “taking of evidence” or analysis that it should always be saved in its delivered state. However, the entire preparation of the visual representation or for the search takes place without starting programs contained in files, 100% on the ArchivistaBox.

Character sets and character sets

The ArchivistaBox now always works with the UTF8 character set when processing emails. If the mail messages to be processed contain other character sets (e.g. ISO-8859-1), the data is converted to UTF8. This means that special characters from almost any character set can be displayed correctly.

Faster processing

Mail messages have to be split into fragments (as shown above) before they can be processed. This process is now carried out using the open source tool ‘ripmime’. This works faster and is better able to recognize faulty mails than the old ‘mha-decode’ solution. This leads to much faster processing of the mails.

Uniform processing of all incoming data

Previously, mail archiving worked separately from the process responsible for importing digital data. The same rules now apply to all files, regardless of whether they are processed as email attachments or via the Office folder.

Conclusion: Continuous further development

This blog has shown what emails are and how they can be processed more securely and efficiently with the ArchivistaBox using the new version 2025/II. An update is recommended for all customers who already use mail archiving.
Mail archiving with the ArchivistaBox can be recommended to all those who have not yet used mail archiving. Regardless of whether they already have an ArchivistaBox in use or not. Enjoy!