Lesbare Daten für viele Jahrzehnte

Egg, 29. Januar 2025: *Als die Firma Archivista GmbH im Jahre 1998 das Licht der Welt erblickte, konnte niemand erahnen, in welcher Geschwindigkeit sich die Technik entwickeln würde. Klar war einzig, die herkömmliche Ablage von Papier war schon damals keine so gute Idee, denn der Umgang mit analogem Schriftgut war/ist teuer, umständlich und fehleranfällig. Heute darf angemerkt werden: Belege in Papier sind zwar nicht ganz verschwunden, doch haben digitale Daten die analoge Welt in weiten Stücken abgelöst. Jedoch, digitale Daten, die digitale Welt, gibt es nicht umsonst. In diesem Blog soll dies am Beispiel der Mails mit der neuen Version 2025/II der ArchivistaBox aufgezeigt werden.*

Nicht ganz so trivial: Mails als Ersatz für Briefe

Im digitalen Schriftverkehr haben sich Mail-Nachrichten als Ersatz für Briefe fast zu 100 Prozent durchgesetzt. Und es mag sein, dass die Mail-Nachrichten aktuell auch schon wieder als veraltet erscheinen mögen. Doch immer dann, wenn es darum geht, etwas schriftlich festzuhalten, kommt die Mail-Nachricht zum Einsatz.

Damit ist auch gesagt, dass Mail-Nachrichten aktuell einen wichtigen Faktor darstellen, wenn es darum geht, digitale Informationen langfristig und sicher verfügbar zu halten.

Mail-Archivierung à la ArchivistaBox

Die meisten Lösungen auf dem Markt «begnügen» sich damit, die Mail-Nachrichten vom Provider abzuholen und im Originalformat zu speichern. Die Lesbarkeit der Daten hängt dabei von der Möglichkeit ab, die Mails mit aktuellen Bordmitteln darstellen zu können. Bei jedem Update können sich die Spielregeln ändern. Was heute lesbar ist, lässt sich vielleicht schon morgen oder auch erst viele Jahre später nicht mehr korrekt darstellen.

Anders bei der ArchivistaBox. Hier werden von Haus aus sämtliche Daten virtuell abfotografiert. Dies bedeutet, dass einmal aufgenommene Mail-Nachrichten visuell langfristig lesbar sind.

Eine solche Herangehensweise mag erstaunen, aber gerade nach mittlerweile fast drei Jahrzehnten im Umgang mit gescannten wie digitalen Daten darf angemerkt werden, dass bei Bild-Dateien im Vergleich zu anderweitig gespeicherten Informationen (Office-, PDF- und/oder Text-Daten) in etwa im Verhältnis 1:1000 kleinere Probleme bei der Lesbarkeit bestehen.

Bild-Dateien bestehen aus Punkten, die auf dem Bildschirm als Pixel gezeichnet

werden. Bei allen anderen Formaten müssen die Informationen jedes Mal neu auf das Ausgabegerät «gezeichnet» werden. Nun sollte dies nicht so schwierig sein, ist es aber. Um zu verdeutlichen, warum dem so ist, sei der (vereinfachte) Aufbau der Mail-Nachrichten kurz angeführt.

Was sind Mail-Nachrichten?
Im Prinzip entsprechen Mail-Nachrichten simplen Text-Dateien:

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

Hallo, ich bin eine Mail. Schwieriger wird es mit den
Grüssen...
Noch schwieriger im übrigen, wenn es um Anhänge (Bilder,
PDF-Dateien) geht.

------=_20081201100242_51266
Content-Type: image/x-portable-anymap; name="eins1.pnm"
Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename="eins1.pnm"

/9j/hADbFiAAGBwYHCAUJBwaJiAiNFAwLCwwRmIwUDpKemZ0ZnJ4gG5wnLiQ
roiAcG6KotqgxL6u
ztDO4pp8yODyyrjwAcbOJCQiMCowNDRehMZexoRwxsbGxsbGxsbGxsbG
xsbGxsbGxsbGxsbG
------=_20081201100242_51266--
```

Das vereinfachte Beispiel hier dient dazu, einen «ungefähren» Eindruck einer Mail-Nachricht zu vermitteln. Im engeren Sinn bestehen Mail-Nachrichten aus drei Teilen: 1) Kopf-Teil, 2) Text und 3) Anhänge.

1. Kopf-Teil (Header)
Hier finden sich die Informationen, von und zu wem die Nachricht verschickt werden soll. Damit eine Mail vom Absender zum Empfänger gelangen kann, wird sie über Provider (Mittler) gesandt. Meist hinterlassen diese in der Form einer Textzeile einen Stempel in Text-Form:

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

Dies bedeutet, dass eine Mail, welche beim Absender versandt wird, auf dem

Weg bis zum Empfänger (zumindest) im Kopf-Teil Veränderungen erfährt. Diese Informationen sind insofern wichtig, um (später) festzustellen, ob eine Mail-Nachricht den gewohnten Weg nahm oder ob es sich mit an Sicherheit grenzender Wahrscheinlichkeit um eine «Fälschung» handelt.
Nebenbemerkung: *Signaturen und Verschlüsselung können zusätzlich helfen, Echtheit und Datenintegrität zu verbessern bzw. zu gewährleisten. Dies ändert aber nichts an der Tatsache, dass Mail-Nachrichten vom Absender zum Empfänger über viele Stellen «wandern» und dabei die obigen Spielregeln nicht zur Anwendung gelangen.*

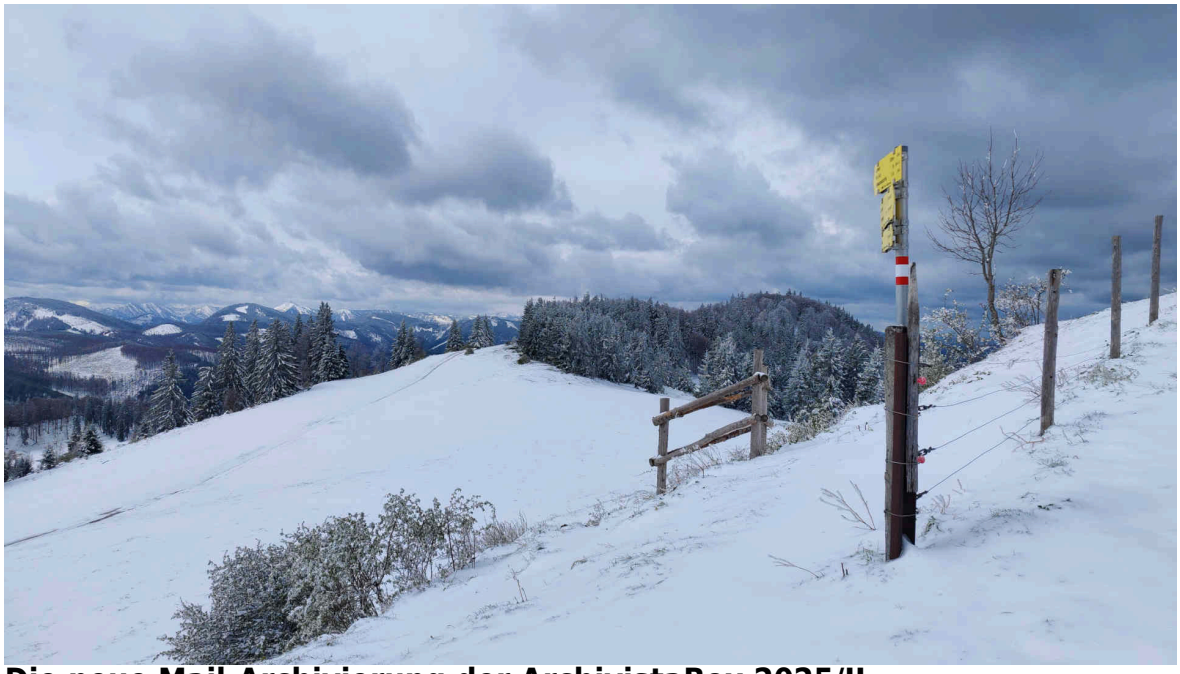
2. Text-Teil
Hier befindet sich die eigentliche Nachricht. Was einfach klingt, hat durchaus seine Tücken, denn spätestens wenn es darum geht, Umlaute und Sonderzeichen darzustellen, kann es zu Darstellungsproblemen kommen. Daher wird oft der gewünschte Zeichensatz mit angegeben. Ob diese Angabe stimmt, ob der Empfänger damit etwas anfangen kann, bleibt jedoch ungeklärt. Verkompliziert wird die Sache dadurch, dass auch bereits in den Kopf-Daten Sonderzeichen vorkommen können bzw. dass diese vom Hauptteil unterschiedliche Zeichensätze enthalten können.
Eine weitere Herausforderung besteht darin, dass der Hauptteil auch in verschiedenen Varianten vorliegen kann. Einmal z.B. in reiner Textform und weiter formatiert (meist als Webseiten-Fragmente). Nun ist eine Mail natürlich keine vollwertige Webseite und oft geht es auch nur darum, irgendwelche Textpassagen besonders zu kennzeichnen (Schrift, Grösse oder Auszeichnung (z.B. Fett). In diesem Sinne gibt es fast keine Grenzen, aber eben auch fast keine Normen.
3. Anhänge
Die Anhänge werden mit Start- und End-Markierungen am Ende der Nachricht abgelegt. Dabei gibt es zwar Normen, wie die Start und Endsequenz aussieht, nicht aber welche Daten dazwischen gesichert werden.
Oft finden sich bei den Anhängen Bilder und PDF-Dateien, sie können aber auch (schadhafte) Programme enthalten. Bis heute besonders berüchtigt sind die Anhänge eines Office-Herstellers, welche Programme mehr oder minder ungefragt ausführen.
Unabhängig von der Problematik der Anhänge, die dem Empfänger «übel» wollen, stellt sich die Frage, was mit den Anhängen bei der Einpflege in ein Dokumenten Management System (DMS) wie der ArchivistaBox zu tun ist. Oft

enthält der Text-Teil der Mail selber nicht die relevanten Informationen (z.B. Bestellungen als PDF-Datei).
Bei der ArchivistaBox bzw. der Mail-Archivierung kann explizit festgelegt werden, ob die Anhänge «abfotografiert» (und auch in den Volltext aufgenommen) werden sollen oder nicht und wenn ja, bei welchen Dateien dies der Fall sein soll.

4. Spezialfälle

Nun gibt es für Mails an sich ein Regelwerk (rfc822). Ob und wie diese Normen umgesetzt werden, dies ist Aufgabe der Anbieter. Gerade grössere Player interpretieren bzw. erweitern die Spezifikation mitunter so, dass andere (zumindest kurzfristig) das Nachsehen heben. Positiver formuliert darf von Spezialfällen die Rede sein. Einige dieser «Sonderlinge» seien hier angeführt.
Mail-Nachrichten können sich selber als Anhang anpreisen und weiter hat irgendwann der Riese aus Redmond begonnen, Anhänge in einem eigenen Format (winmail.dat) zu versenden.

Mail-Dateien können Links auf externe Webseiten enthalten, um z.B. von dort zusätzliche Inhalte (meist Bilder, aber nicht nur) abzurufen. Diese Form wird aktuell häufig genutzt, lässt sich damit gut feststellen, ob bzw. wo Nachrichten ankommen bzw. geöffnet werden. Überdies können Inhalte je nach Empfänger angepasst werden. Beispiel: Ein z.B. als externer Link «getarnter» Preis in einer Mail wird je nach aufrufender IP-Adresse unterschiedlich «ausgespuckt».
Weiter kann es sein, dass Mail-Anhänge unter Umständen «kaputt» sind, denn die Umwandlung der originären Dateien nach ASCII-7Bit (bzw. Rückumwandlung beim Lesen) arbeitet ohne Fehlererkennung. Ein einziges «verschlucktes» (fehlerhaftes) Zeichen kann dazu führen, dass die Datei beim Empfänger nicht gelesen werden kann. Die gleiche Problematik besteht natürlich auch, wenn z.B. eine Bild-Datei bereits vor der Übermittlung «defekt» ist.



Die neue Mail-Archivierung der ArchivistaBox 2025/II

Die Mail-Archivierung der ArchivistaBox ist seit ca. 15 Jahren verfügbar. Damals enthielten Mail-Nachrichten meist nur kurze Anhänge und die graphische Gestaltung war (wenn überhaupt) schlicht gehalten. Vielleicht ein Firmenlogo oder eine kleine PDF-Datei, ein zwei Fotos, mehr wurde damals nicht per Mail übermittelt.

Vor etwa acht Jahren zeigte sich, dass die Mail-Nachrichten viel bunter dargestellt wurden. Daher wurde die Mail-Archivierung für HTML-Mails optimiert. Im Jahre 2023 erschien die Mail-Archivierung für Office365, weil bei diesen Diensten zwingend der IMAP-Standard deaktiviert wurde. Mit der Version 2025/II wird die Mail-Archivierung erneut erweitert. Folgende zentrale Punkte sind neu vorhanden:

Handhabung der nicht mehr existierenden Links

Immer häufiger werden Mail-Nachrichten mit externen Links versandt. Mittlerweile werden die extern vorgehaltenen Inhalte auch immer schneller wieder deaktiviert. Mit der neuen Mail-Archivierung in 2025/II kann gewählt werden, ob a) externe Links ignoriert werden, b) diese berücksichtigt werden bzw. c) die Mails wahlweise mit/ohne Links verarbeitet werden. Wahlweise bedeutet: Links werden verwendet, aber nur, wenn diese innert angemessener Zeit Resultate liefern.

Handhabung winmail.dat

Bislang wurden diese Anhänge als entsprechende Dateischnipsel gesichert. D.h. die eigentlichen Inhalte der Winmail-Fragmente wurden weder virtuell abfotografiert noch für die Volltexterkennung aufbereitet. Neu werden diese entpackt und die entsprechenden Inhalte korrekt erfasst.

Verschachtelte (kaskadierte) Mails

Mail-Nachrichten können selber als Anhang wieder als Mail-Nachricht auftreten. Bislang wurden die als Anhang anfallende Mail-Nachrichten nicht verarbeitet. Neu werden diese so weit wie sinnvoll (maximal über neun Ebenen) erfasst.

Prüfen auf viele (un)bekannte Dateitypen

Die neue Mail-Archivierung wurde mit zehntausenden von Mail-Nachrichten, die bei der Firma Archivista GmbH seit der Gründung anfielen, getestet. Daraus wurden viele bislang der Mail-Archivierung unbekannte Dateitypen neu hinzugefügt. Selbstverständlich können ungewollte Dateitypen nach wie vor ausgeschlossen werden. Die ArchivistaBox 2025/II erkennt aber viele Dateiformate in der Grundkonfiguration, die bislang nicht spezifisch bzw.

optimal verarbeitet wurden.

Prüfung auf fehlerhafte Daten

Mit der Version 2025/II werden Anhänge besser auf fehlerhafte Übermittlung bzw. entsprechendem Ursprung geprüft. Das bisherige Prüfprogramm erkannte die Daten als korrekt an, was dazu führte, dass der Importvorgang startete, aber am Ende scheiterte. Neu wird die Integrität der Daten besser geprüft und bei fehlerhaften Daten der Importjob gar nicht erst gestartet.

Dabei darf die Frage gestellt werden, wie mit fehlerhaften Daten umzugehen ist. Sollen die übermittelten Daten «gesäubert» werden oder soll das Original unverändert bleiben, auch wenn die Inhalte fehlerhaft sind? Oder ketzerischer gefragt: Sollen gar allfällige Viren langfristig «gesichert» werden?

Originäre Daten sind für eine spätere «Beweisaufnahme» bzw. Analyse von derart zentraler Bedeutung, dass diese immer im angelieferten Zustand gesichert werden. Jedoch, die gesamte Aufbereitung der visuellen Darstellung bzw. für die Suche erfolgt ohne das Starten von in Dateien enthaltenen Programmen, und zwar zu 100% auf der ArchivistaBox.

Zeichensätze und Zeichensätze

Die ArchivistaBox arbeitet beim Verarbeiten der Mails neu immer mit dem UTF8-Zeichensatz. Enthalten zu verarbeitende Mail-Nachrichten andere Zeichensätze (z.B. ISO-8859-1), so werden die Daten nach UTF8 umgewandelt. Damit können Sonderzeichen aus fast beliebigen Zeichensätzen korrekt dargestellt werden.

Schnellere Verarbeitung

Mail-Nachrichten müssen (wie oben aufgezeigt) in Fragmente aufgeteilt werden, ehe sie verarbeitet werden können. Neu erfolgt dieser Prozess mit dem Open Source Tool ‹ripmime›. Dieses arbeitet schneller und ist besser in der Lage, fehlerhafte Mails zu erkennen als die alte Lösung ‹mha-decode›. Dies führt zu einer weit schnelleren Verarbeitung der Mails.

Einheitliches Verarbeiten aller eingehenden Daten

Die Mail-Archivierung arbeitete bislang losgelöst vom Prozess, welcher für den Import von digitalen Daten zuständig ist. Neu gelten die gleichen Regeln für alle Dateien, ganz egal, ob diese als Mail-Anhang oder über den Office-Ordner verarbeitet werden.

Fazit: Stetige Weiterentwicklung

In diesem Blog wurde aufgezeigt, was Mails sind und wie sie mit der neuen Version 2025/II sicherer und effizienter mit der ArchivistaBox verarbeitet werden können. Ein Update darf allen Kunden bzw. Kundinnen empfohlen werden,

welche die Mail-Archivierung bereits einsetzen.
All jene, welche bislang keine Mail-Archivierung im Einsatz haben, darf die Mail-Archivierung mit der ArchivistaBox empfohlen werden. Und zwar unabhängig davon, ob sie bereits eine ArchivistaBox im Einsatz haben oder nicht. Enjoy!


Facebook


Twitter