Pilotprojekt zur Langzeitarchivierung digitaler E-Mail-Korrespondenzen des Bundesvorstandes der Vereinten Dienstleistungsgewerkschaft ver.di

Von Mike Zuchet

Der Einsatz von E-Mails für die interne und externe Kommunikation ist heute so selbstverständlich wie die früher übliche Umlaufmappe oder der klassische Brief in Papierform. Glaubt man verschiedenen Studien, die sich mit der Relevanz der E-Mail als Kommunikationsmittel in Unternehmen und Organisationen auseinandergesetzt haben, liegen zwischen 35 und 75 % aller relevanten Informationen nur noch in dieser digitalen Form vor und finden keinen papierenen Ausdruck mehr.[1] Aus diesem Grund und weil der Gesetzgeber fordert, dass steuerlich relevante E-Mails sechs resp. zehn Jahre revisionssicher vorgehalten werden müssen, liegt das Thema der Langzeitarchivierung digitaler E-Mail-Korrespondenzen mehr als auf der Hand.[2] Überdies werden dezidierte E-Mail-Server durch die Auslagerung in entsprechende Langzeitspeicher entlastet und der in der Regel hohe Aufwand bei der Wiederherstellung gelöschter E-Mails wird vermieden.

Das Archiv der sozialen Demokratie (AdsD) der Friedrich-Ebert-Stiftung, das sich seit längerer Zeit mit der Langzeitarchivierung digitaler Überlieferungen befasst, nahm dies zum Anlass, sich intensiv mit diesem Thema auseinander zu setzen. Es handelte sich im wahrsten Sinne des Wortes um Pionierarbeit und das AdsD darf dabei für sich in Anspruch nehmen, als erstes

---

1 Vgl. dazu http://www.silicon.de/39193609/e-mails-sind-das-wichtigste-in-deutschen-unternehmen/ (alle Links in diesem Text wurden am 31.5.2012 geprüft).

2 Vgl. dazu Abgabenordnung (AO) § 146 *Ordnungsvorschriften für die Buchführung und für Aufzeichnungen*: http://www.gesetze-im-internet.de/ao_1977/__146.html und § 147 *Ordnungsvorschriften für die Aufbewahrung von Unterlagen*: http://www.gesetze-im-internet.de/ao_1977/__147.html; Grundsätze ordnungsgemäßer DV-gestützter Buchführungssysteme (GoBS), Abschnitte 5 Datensicherheit und 6 Dokumentation und Prüfbarkeit: http://www.bundesfinanzministerium.de/nn_314/DE/BMF__Startseite/Service/Downloads/Abt__IV/BMF__Schreiben/015,templateId=raw,property=publicationFile.pdf; Handelsgesetzbuch (HGB) § 257 *Aufbewahrung von Unterlagen/Aufbewahrungsfristen*: http://www.gesetze-im-internet.de/bundesrecht/hgb/gesamt.pdf; Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GDPdU), Abschnitte II *Prüfbarkeit digitaler Unterlagen* und III *Archivierung digitaler Unterlagen*: http://www.bundesfinanzministerium.de/nn_95356/DE/Wirtschaft__und__Verwaltung/Steuern/Veroeffentlichungen__zu__Steuerarten/Abgabenordnung/Datenzugriff__GDPdU/002,templateId=raw,property=publicationFile.pdf. Siehe dazu auch die Merksätze zur revisionssicheren elektronischen Archivierung des Verbands Organisations- und Informationssysteme e.V. (VOI): http://www.voi.de.

Archiv in Deutschland über seine Aktivität in diesem Bereich zu berichten.[3] Eine server- oder clientgesteuerte Archivierung[4] wurde von vorneherein ausgeschlossen. Stattdessen wurde im AdsD großen Wert darauf gelegt, die Archivierung außerhalb der Ursprungsumgebung durchzuführen sowie E-Mails und die dazugehörigen Metadaten für eine *systemunabhängige* Umgebung aufzubereiten und vorzuhalten. Neben der einseitigen Produktbindung sprechen noch weitere Gründe gegen die (Langzeit-)Archivierung mittels MS Outlook, das heißt als PST-Datei. Neben der Größenbeschränkung der zentral verwaltenden PST-Datei – näheres siehe weiter unten – sind große Dateien überproportional stark von einem Ausfallrisiko betroffen und die Datensicherung, die größenbedingt längere Zeit in Anspruch nimmt, blockiert das Produktivsystem. Hinzu kommt, dass die Recherche nach einzelnen E-Mails eine langwierige Angelegenheit werden kann, falls mehrere PST-Dateien durchsucht werden müssen, und dass der Speicherort auf der lokalen C-Partition des Anwenderrechners liegt, die in der Regel nicht redundant gesichert wird.

Zusammen mit dem ver.di-Bundesvorstand wurde ein Pilotprojekt gestartet, das die Machbarkeit der Übernahme, Aufbereitung und des Zugriffs digitaler E-Mail-Korrespondenzen nach dem Open Archival Information System (OAIS, ISO 14721:2003[5]) im AdsD überprüfen sollte. Nach intensiven Gesprächen, die in erster Linie als 'vertrauensbildende Maßnahmen' dienten, und der Ausarbeitung einer speziellen Datenschutzerklärung seitens ver.di konnte das AdsD im Oktober 2010 die gesamte vorliegende E-Mail-Korrespondenz der Leitung der Abteilung Grundsatz und im November darauf des Büros des Bundesvorsitzenden übernehmen.[6] Bei dem verwendeten E-Mail-Programm handelte es sich um Microsoft Outlook 2007. Im ersten Fall wurden ca. 14 000 E-Mails übernommen, im zweiten Fall ca. 9 000 E-Mails. MS Outlook

---

3 Der Aufsatz von Beda Kupper ist dem Verfasser erst nach Fertigstellung dieses Texts bekannt geworden: Beda *Kupper*: E-Mail-Archivierung. In: Actualité archivistique Suisse. Archivwissenschaft Schweiz aktuell. Hg. von G. *Coutaz*, N. *Meystre-Schaeren*, B. *Roth-Lochner* und A. *Steigmeier*. Baden (CH) 2008. S. 88–117. In vielerlei Hinsicht stimmen die Auffassungen des Verfassers mit denen Kuppers überein, allerdings vertritt letzterer hinsichtlich des Punktes *Erscheinungsbild* (vgl. S. 101) einen diametralen Standpunkt, der der anerkannten Forderung nach Erhalt der strukturellen Authentizität nicht gerecht wird. Der konkrete äußere Anlass für die Pilotierung ist die Einführung eines Dokumentenmanagementsystems (DMS) beim Bundesvorstand der Vereinten Dienstleistungsgewerkschaft ver.di. Dieser trat an das AdsD mit der Frage heran, ob Möglichkeiten existieren, die bislang entstandenen E-Mail-Korrespondenzen außerhalb des zukünftigen DMS zu archivieren. Wie oben erwähnt, ergibt sich dabei auch der positive *Nebeneffekt*, dass die E-Mail-Server bei ver.di entlastet würden.

4 Beim serverseitigen Ansatz, dem Journaling, werden sämtliche eingehende wie ausgehende E-Mails sofort in das systemeigene Archivsystem transferiert, weshalb der Speicherplatz entsprechend hoch bemessen sein muss. Vorteil: Vollständigkeit, Nachteil: Speicherbedarf, digitaler *Beifang* (Spam). Beim clientseitigen Ansatz entscheidet der Anwender selbst, welche E-Mails im systemeigenen Archivsystem archiviert werden. Vorteil: Auswahl/Flexibilität, Nachteil: möglicher Datenverlust.

5 Siehe dazu http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_366.pdf.

6 Unmittelbar verknüpft mit der Übernahme der E-Mail-Korrespondenz ist die Frage der Bewertung. Die Bewertung wurde auf der Ebene des jeweiligen E-Mail-Ordners durchgeführt, nicht auf der Ebene der einzelnen E-Mail. Sämtliche als *privat* deklarierten E-Mail-Ordner wurden nicht übernommen. An dieser Stelle wurde wieder besonders deutlich, wie wichtig bereits die Arbeit im vorarchivischen Umfeld der Langzeitarchivierung digitaler Überlieferungen ist.

2007 – wie auch die Nachfolgeversion – legt sowohl die eigentliche Nachricht (E-Mail[7]) als auch die Anhänge in einer proprietären Personal Storage-Datei (PST) ab, die je nach Mail-Volumen mehrere Gigabyte (GB) groß werden kann.[8] Es sei an dieser Stelle angemerkt, dass das erarbeitete Konzept nicht nur dazu in der Lage ist, E-Mail-Korrespondenz aus der besagten Programmumgebung zu übernehmen und aufzubereiten, sondern aus nahezu allen gängigen E-Mail-Programmen.[9]

Nach der Überführung der PST-Dateien in das AdsD wurden zuerst Kopien der Ursprungsdateien auf mehreren voneinander unabhängigen Datenträgern erstellt. Um die Datenintegrität bei den Kopierschritten nachhalten zu können, wurde ein Prüfsummenverfahren (SHA-1; *secure hash algorithm*) eingesetzt. Da sämtliche Transferschritte fehlerfrei verliefen, wurden die PST-Dateien in ihrer Entstehungsumgebung (gemeint sind sowohl Programm als auch Betriebssystem) geöffnet, mit anschließendem inkrementellem Abgleich.[10] Danach wurde jede zu archivierende E-Mail mithilfe eines eigens im AdsD entwickelten Makros aus MS Outlook 2007 exportiert und als MSG-Datei abgelegt.[11] In Anbetracht der Funktionen und Importmöglichkeiten des Datenbanksystems Faust 6.0 Professional, das derzeit im AdsD eingesetzt wird, wurde das Makro speziell dahin gehend programmiert, dass der Name jeder MSG-Datei aus einer fortlaufenden Nummer, dem Datum und der Uhrzeit des Eingangs resp. des Ausgangs in das E-Mail-Programm generiert wurde.[12] Als Nächstes wurde mithilfe des Tools Solid PDF Tools

---

7 Hauptbestandteile einer E-Mail sind der Header, Body (Nachrichtentext) und Anhang.

8 In der PST-Datei werden sämtliche E-Mail-Vorgänge (Eingang, Ausgang, Gesendet, Entwürfe), Termine, Aufgaben, Notizen, Journaleinträge, Kontakte (ausgenommen: Adressen des Persönlichen Adressbuches (PAB)) sowie alle Ordner und deren strukturelle Gliederungen gespeichert. In MS Outlook 2003 und 2007 liegt die Standardgrößenbeschränkung bei PST-Dateien bei 20 GB. Vorkonfiguriert ist MS Outlook 2010 für eine Größe der PST-Datei von 50 GB, die allerdings durch Änderungen in der Windows-Registry erhöht oder reduziert werden kann. Siehe dazu http://support.microsoft.com/kb/982577/de.

9 Derzeit wird die Übernahme digitaler E-Mail-Korrespondenzen eines Gewerkschaftssekretärs der ehemaligen Gewerkschaft Öffentliche Dienste, Transport und Verkehr (ÖTV) resp. der Vereinten Dienstleistungsgewerkschaft ver.di vorbereitet, der mit einem proprietären E-Mail-Programm eines Internetproviders arbeitet. Nach dem Transfer der E-Mail-Korrespondenz des Hinterlegers an den Mail-Server der FES werden diese an MS Outlook 2010 übergeben und wie im Text beschrieben aufbereitet. Da es bei diesem Ablauf zu entsprechenden Einträgen in den jeweiligen E-Mail-Headern kommt, werden diese entsprechend gekennzeichnet, um den letzten Transferschritt ins AdsD von den vorherigen differenzieren zu können.

10 Die bisherige Praxis hat gezeigt, dass E-Mails bereitgestellt werden, die bereits bei einer vorherigen Übernahme übernommen wurden. Da auch diesem Wege der Grad der Redundanzen sukzessive zunähme, verfolgt das AdsD den Ansatz des inkrementellen Archivierens. Dabei werden die zu übernehmenden E-Mails mit den bereits übernommenen E-Mails automatisch abgeglichen und nur der Zuwachs kommt für die Langzeitarchivierung infrage.

11 Derzeit wird daran gearbeitet, den Zwischenschritt der MSG-Datei zu überspringen und die E-Mails direkt als PDF resp. PDF/A-Datei zu generieren.

12 Der Aufbau des Dateinamens entspricht folgendem Schema: nn mm.dd.yyyy hh.mm.ss, z.B. 01 12.12.2004 07.35.55.

V6 jede MSG-Datei in eine PDF/A-Datei (PDF/A-1b)[13] konvertiert, um Aufbau und Struktur der E-Mails *einzufrieren* und somit die informationelle und strukturelle Authentizität der digitalen Quelle zu gewährleisten.[14] In einigen wenigen Fällen war die PDF/A-Konvertierung nicht möglich, da der Inhalt der Ausgangsdatei nicht konform war mit den Spezifikationen der ISO-Norm. In diesem Falle fand lediglich eine Konvertierung nach PDF (1.4) statt. Diese Ausnahmen wurden vom o.g. Konvertierungsprogramm protokolliert und fanden Eingang in die dazugehörigen Metadaten.

Jede PDF/A-Datei wurde mit einer Prüfsumme versehen, um wie im Falle der PST-Datei Änderungen festzustellen und die Datenintegrität langfristig zu gewährleisten. Da es nicht nur um die E-Mail ging, sondern auch um den jeweiligen Anhang – oftmals diente die E-Mail nur als *Transportmittel* für eine zu übermittelnde Datei –, wurden die Anhänge aus der PST-Datei extrahiert und in das Langzeitarchivierungsformat PDF/A konvertiert, soweit dies technisch möglich und sinnvoll war. War die Level B-Konformität nicht zu erreichen, wurde zumindest versucht, die Dateianhänge in das PDF-Format (1.4) zu konvertieren. Schlug auch dieser Schritt aus inhaltlich-technischen Gründen fehl, wurden die Dateianhänge im Ausgangsformat vorgehalten. Redundanzen innerhalb der identische Dateianhänge wurden mit Hilfe von Prüfsummen erkannt, Dubletten werden vor der Formatkonvertierung gelöscht.[15] Da jede Langzeitarchivierung digitaler Überlieferungen mit der quantitativen und qualitativen Existenz von Metadaten steht oder fällt, war es nun notwendig, auf die Metadaten der einzelnen E-Mails zuzugreifen, sie aus der proprietären PST-Hülle zu extrahieren und sie in ein programmunabhängiges Dateiformat zu überführen.[16] Überdies werden auch die Metadaten des E-Mail-Headers (Internetkopfzeilen) übernommen, da sie Aufschluss über Absender und Weg einer E-Mail geben und bei der Über-

---

[13] PDF/A ist ein Subset des verbreiteten Portable Document Format (PDF) und liegt seit 2005 als ISO-Norm vor (ISO 19005-1:2005). Die ISO-Norm beschreibt zwei Konformitätsebenen: A. PDF/A-1b – Level B (Basic) conformance, Ziel: eindeutige visuelle Reproduzierbarkeit. B. PDF/A-1a – Level A – (Accessible) conformance, Ziel: eindeutige visuelle Reproduzierbarkeit sowie Textabbildung nach Unicode und inhaltliche Strukturierung des Dokuments. Im Verlauf des Pilotprojekts hat sich Level B als ausreichende und zuverlässige Variante erwiesen, die auch nach der Pilotphase verwendet wird.

[14] Falls der Text aus den E-Mails lediglich extrahiert und in das Faust-Datenbanksystem überführt worden wäre – auch über den Zwischenschritt einer entsprechenden XML-Metadatendatei –, wären sämtliche Textformatierungen und -strukturierungen verloren gegangen. Die jeweilige Nachricht hätte sich als unübersichtlicher Fließtext dargestellt, was einem wesentlichem Verlust der sog. *significant properties* entspricht.

[15] Handelt es sich dabei zweifellos um einen wichtigen Aspekt der Bewertung, ist dieser Vorgang auch unter dem Gesichtspunkt der Speicherökonomie interessant. Sind mehrere E-Mails mit Dateien verknüpft, die ausweislich ihrer Prüfsumme identisch sind, so werden die Dateien in der Datenhaltung auf eine reduziert. Die Verknüpfung der gelöschten Dateien wird umgelenkt, sodass der Nutzer der archivierten E-Mails keinen Unterschied bemerkt, Administratoren können den Vorgang aber nachvollziehen. Bei Dateien, die unterschiedliche Namen tragen, aber den identischen Inhalt aufweisen, hat dies jedoch Auswirkungen auf die Metadaten, d. h. auf die Referenz. In diesem Falle trägt die Referenz als letzte Pfadangabe nicht den ursprünglichen Dateinamen, sondern den Namen der jeweiligen übrig gebliebenen Datei. Auch in diesem Falle ist ein entsprechender Vermerk in der Datenbank zu finden, der den Namen der gelöschten Datei beinhaltet.

[16] Folgende Metadaten wurden extrahiert: Dringlichkeit, Betreff, Absender, Empfänger, Kopie an, Erhalten am, Erstellt am, Nachrichtentext, Name(n) der Anlage(n).

prüfung der Authentizität gegebenenfalls zusätzlich herangezogen werden können.[17] Die Wahl für das Dateiformat fiel auf das XML-Format, nicht zuletzt deshalb, weil Faust 6.0 professional über entsprechend konfigurierbare Importfilter verfügt und es sich bereits bei früheren Datenimporten als zuverlässiges Austauschformat bewährt hatte.

Sowohl die überlieferten E-Mails der erwähnten Abteilung Grundsatz als auch die des Büros des ver.di-Bundesvorsitzenden lagen größtenteils strukturiert in entsprechenden E-Mail-Ordnern vor, die eine Mischung aus Provenienz- und Pertinenzmerkmalen aufwiesen. Die jeweiligen (Ablage-)Strukturen wurden übernommen und mit Hilfe von Thesauri in Faust-Datenbanken nachgebildet, was der Ordnung und dem schnellen Zugriff diente.

Nach Abschluss der aufgeführten Arbeitsschritte lagen sämtliche Komponenten vor, um sie in das Archivdatenbanksystem Faust 6.0 professional zu überführen. Als Erstes wurden die PDF/A-Dateien, die als authentische E-Mail-Abbilder fungierten, sukzessive importiert, wobei nicht die PDF/A-Datei selbst Bestandteil der Datenbank wurde, sondern nur die Referenz zum jeweiligen Speicherort. Als Nächstes wurden die dabei generierten Datensätze um die dazugehörigen Metadaten ergänzt. Damit dieser Schritt reibungslos funktionierte und es zu keinen falschen Verknüpfungen von E-Mail-Abbildern und Metadaten kam, war es wichtig, dass das oben erwähnte Makro an besagter Stelle eingesetzt wurde. Die vorliegenden Dateianhänge wurden von einer separaten Faust-Datenbank erfasst. Beide Datenbanken – a) Referenz zum E-Mail-Abbild samt Metadaten und b) technische Metadaten und Referenz zu den Dateianhängen – wurden über Assoziativ-Referenzen miteinander verknüpft.[18] Sie ermöglichen es einerseits, dass der jeweilige Dateianhang vom Faust-Datensatz aus, der das E-Mail-Abbild repräsentiert, angesteuert und anschließend angesehen werden kann. Andererseits ist es ebenso möglich, von einem recherchierten Anhang zur entsprechenden E-Mail zu gelangen. Das beschriebene Prinzip – ein zentraler Datensatz, von dem entsprechende Verweise ausgehen – ist auf eine Vielzahl von Datenbanksystemen übertragbar und deshalb nicht auf das im AdsD verwendete beschränkt – eine konsequente und zukunftsweisende Umsetzung des OAIS-Moduls *Preservation Planning*. Da sowohl der textuelle Inhalt der PDF/A-Dateien als auch der der Dateianhänge im Datenbanksystem invertiert wurden, kann direkt in den E-Mails und v.a. in den Dateianhängen recherchiert werden. Damit existieren nun Recherchemöglichkeiten, die deutlich über die Möglichkeiten im Ursprungsprogramm hinausgehen.

---

17 ver.di nutzt den Microsoft Exchange Server für die E-Mail-Kommunikation. Sofern E-Mails intern auf einem Exchange Server zugestellt werden, werden keine Metadaten in den Header geschrieben und die Internetkopfzeile bleibt entsprechend leer. Aus diesem Grund können aus internen E-Mails der ver.di keine entsprechenden Header-Metadaten extrahiert werden. Sobald E-Mails einen externen Server passieren, werden entsprechende Metadaten in den Header geschrieben, anhand derer die jeweilige Route rekonstruiert werden kann.

18 Die Praxis hat gezeigt, dass in zahlreichen Fällen die identischen Dateien als Anhänge verschickt wurden. Wie im Text erwähnt, wurden diese Redundanzen gelöscht. Hinzu kam der Umstand, dass unterschiedliche Dateien identische Dateinamen trugen und die Assoziativ-Referenzen nicht mehr eindeutig waren und es falschen Referenzierungen gekommen ist. Um dieser Fehlerquelle zukünftig zu beheben, wurden die bisherigen Metadaten in den Faust-Datenbanken um das Metadatum *Prüfsumme Anhang* ergänzt, über das nun die Assoziativ-Referenzierung läuft.

Erfreulicherweise kann festgestellt werden, dass das vom AdsD entwickelte Verfahren auf eine breite Akzeptanz innerhalb der ver.di-Bundesverwaltung gestoßen ist. Dies wird allein dadurch deutlich, dass sich ver.di aus eigenem Antrieb an das AdsD wendet, um digitale Korrespondenzen für die Langzeitarchivierung abzugeben – für private und politische Archive kein alltäglicher Vorgang.

Die Pilotierung wurde erfolgreich abgeschlossen und die Langzeitarchivierung digitaler Korrespondenzen läuft mittlerweile im Produktivbetrieb. Die letzte Übernahme fand Anfang April 2012 statt, die nächste steht Ende Mai des Jahres bevor.