Fachbeitrag

Nico Beyer, Felix Gericke und Alexander Hinze-Hüttl

EMILiA: Effiziente und nutzungsorientierte E-Mail-Archivierung mithilfe (teil-)automatisierter Prozesse

EMILiA: Efficent and user-oriented email archiving using (semi-)automated processes

https://doi.org/10.1515/abitech-2024-0028

Zusammenfassung: Angesichts der Flüchtigkeit digitaler Daten und der wachsenden Bedeutung der elektronischen Kommunikation sowohl im privaten als auch im dienstlichen Bereich, müssen zeitnah Lösungen für eine erfolgreiche Übernahme, langfristige Sicherung, Speicherung und rechtskonforme Zugänglichmachung von E-Mails gefunden werden. Das EMILiA-Projekt zielt darauf ab, Gedächtnisinstitutionen mithilfe der Entwicklung einer teilautomatisierten Software bei der Archivierung von E-Mails zu unterstützen. Im Anschluss an den Projektzeitraum ist die Ausgründung in Form eines Start-up-Unternehmens geplant.

Schlüsselwörter: E-Mail-Archivierung, Softwareentwicklung, Automatisierung

Abstract: Given the volatile nature of digital data and the growing importance of electronic communication in both private and professional spheres, timely solutions must be found for the successful acquisition, long-term storage and legally compliant accessibility of emails. The EMILiA project aims to support cultural memory institutions by developing semi-automated software for archiving emails. After the project, there are plans to establish a startup company based on the project's outcomes.

Keywords: email archiving, software development, automation

1 Einleitung

Ein großer Teil des archivwürdigen Informationsaustauschs hat sich in den letzten Jahren zunehmend in digitale Umgebungen verlagert, was Gedächtnisinstitutionen vor vielfältige organisatorische, rechtliche und technische Herausforderungen stellt. Als besonders wichtiger Kommunikationskanal sind E-Mails zu nennen, die inzwischen sowohl im dienstlichen als auch im privaten Bereich zunehmend Telefonate und analoge Schriftwechsel ersetzen. Laut dem Statistischen Amt der Europäischen Union nutzten im Jahr 2023 rund 79 Prozent aller in Europa ansässigen Personen regelmäßig E-Mails für die Kommunikation.[1] Einer Schätzung der Radicati Group zufolge sollen weltweit im gleichen Zeitraum mehr als 347 Milliarden Nachrichten gesendet und empfangen worden sein.[2] Trotz des potenziell großen Informationsgehalts von E-Mails gibt es bisher kaum konkrete Beispiele dafür, wie elektronische Nachrichten durch Gedächtnisinstitutionen effizient bewertet, sinnvoll aufbereitet und möglichst zeitnah zugänglich gemacht werden können. Besonders gravierend ist, dass E-Mail-Postfächer aufgrund der Flüchtigkeit digitaler Daten im Vergleich zu analogen Informationsträgern äußerst fragil sind. Ohne aktive Maßnahmen, die relativ zeitnah an ihrer Entstehung liegen, drohen sie in kürzester Zeit verloren zu gehen.[3] Aus diesem Grund müssen so schnell wie möglich Lösungen für ihre Übernahme und langfristige Erhaltung gefunden werden. Dieser Beitrag setzt sich damit auseinander, wie E-Mails effizient und ohne Informationsverlust bewahrt werden können.

1 Eurostat. Individuals using the internet for sending/receiving e-mails. https://doi.org/10.2908/TIN00094. Zuletzt geprüft am 17.05.2024.
2 The Radicati Group. Email Statistics Report, 2023–2027. https://www.radicati.com/wp/wp-content/uploads/2023/04/Email-Statistics-Report-2023-2027-Executive-Summary.pdf. Zuletzt geprüft am 17.05.2024.
3 Siehe hierzu z. B. auch Cornelio-Baker, Nikka. „Signed, Sealed, Delivered, I'm (Still Not) Yours: Challenges in Archiving Electronic Communication in the 21st Century." *The IJournal: Student Journal oft the Faculty of Information* 8,2 (2023): 85–97. https://theijournal.ca/index.php/ijournal/article/view/41036. Zuletzt geprüft am 17.05.2024.

Zunächst wird der Komplex E-Mail analysiert. Darauffolgend werden potentiell überlieferungswürdige Informationen gelistet und die daraus resultierenden Überlieferungsziele abgeleitet. Im Anschluss werden vorhandene Lösungsansätze der E-Mail-Archivierung vorgestellt und bewertet. Dann werden technische, inhaltliche und rechtliche Herausforderungen zusammengetragen, die bei der E-Mail-Archivierung berücksichtigt werden müssen. Abschließend wird der geplante Funktionsumfang der vom EMILiA-Projektteam konzipierten Softwarelösung präsentiert.

2 Ausgangslage und Überlieferungsziele

Um eine fundierte Grundlage für die nachfolgenden Ausführungen zu schaffen, muss zunächst darauf eingegangen werden, um was für ein Medium es sich bei E-Mails handelt und welche Arten von Informationen damit übermittelt werden. Darüber hinaus wird herausgearbeitet, welche Parameter aus Sicht des EMILiA-Projektteams bei der Ausarbeitung einer geeigneten Archivierungsstrategie beachtet werden sollten. Die inhaltliche Grundlage für das Projekt bilden Konzepte, Arbeitsergebnisse und Erfahrungen, die seit 2015 im Rahmen einer Kooperation des Fachbereichs Informatik der Freien Universität Berlin und des Archivs der Max-Planck-Gesellschaft entstanden sind. Zu Beginn der Kooperation gingen die Beteiligten davon aus, dass sich E-Mails ähnlich unkompliziert wie klassische Briefkorrespondenzen archivieren lassen und für die Übernahme lediglich einige wenige organisatorische und technische Fragen zu klären seien. Bereits erste praktische Übernahmeversuche zeigten jedoch, dass bei der E-Mail-Archivierung eine ganze Reihe von Herausforderungen bedacht werden muss.

E-Mails sind weit mehr als nur der moderne Nachfolger des traditionellen Briefs, da sie neben textuellen Inhalten auch die unterschiedlichsten Dateien übertragen können. Inhaltlich sind E-Mail-Postfächer ebenso vielfältig wie die sie nutzenden Personen oder Körperschaften und können von offiziellen Schreiben über private Nachrichten bis hin zu komplexen Forschungsdaten die verschiedensten archivwürdigen Informationen enthalten. Vor der Formulierung einer konkreten Archivierungsstrategie muss jedoch zuerst geklärt werden, welche signifikanten Eigenschaften eine E-Mail ausmachen und welche Überlieferungsziele hieraus abgeleitet werden können.

Grundsätzlich besteht jede E-Mail aus einem Header, der unterschiedliche Metadaten in Form von Schlüssel-Wert-Paaren (Key-Value-Pairs) enthält, und einem Body, in dem sich der eigentliche Inhalt befindet.[4] Der E-Mail-Body kann entweder aus reinem Text bestehen oder durch die Verwendung von Einbettungen auch audiovisuelle Elemente beinhalten. Optional können den elektronischen Nachrichten zudem Dateien angehängt werden.

Für die Archivierung ist der Header von zentraler Bedeutung, da er unter anderem Informationen über Sende- und Empfangspartei(en), eventuell vorhandene Dateianhänge sowie das Sendedatum enthält. Der E-Mail-Standard selbst definiert nur zwei obligatorische Felder. Darüber hinaus sind aber weit über 100 weitere Schlüssel-Wert-Paare bekannt, die beliebig vom E-Mail-Client und den kommunizierenden Mailservern gesetzt werden können.[5] Im Gegensatz zum Kopf eines traditionellen Briefs, ist beim E-Mail-Header ohne weitere Schritte nur ein Bruchteil der tatsächlich vorhandenen Metainformationen direkt im E-Mail-Client sichtbar. Neben Informationen, die für die erfolgreiche Übertragung einer Nachricht benötigt werden, können im Header beispielsweise auch Empfangs- und Lesebestätigungen, Berichte von Spam- und Virenscannern oder Termineinladungen enthalten sein. Für eine spätere Auswertung durch Forschende können diese Informationen durchaus von Bedeutung sein, weshalb sie erhalten werden sollten.

Ebenso überlieferungswürdig sind alle Informationen, die es erlauben, das Korrespondenznetzwerk einer Person zu rekonstruieren. Um die Authentizität von Nachrichten nachvollziehbar zu machen, können E-Mails digital signiert werden. Da es sich bei digitalen Signaturen gewissermaßen um das moderne Äquivalent zur analogen Unterschrift oder Paraphe handelt, sollten sie auf jeden Fall erhalten werden – und zwar in einer lesbaren Form.

Weiter ist es wünschenswert, bestimmte Funktionen eines klassischen E-Mail-Programms zu erhalten, um eine spätere Nutzung zu erleichtern. Gemeint sind hiermit vor allem Such- und Filterfunktionen, Ordnerstrukturen sowie Kennzeichnungen und Kategorisierungen. Um diese Informationen an Forschende weitergeben zu können, sollten E-Mails daher möglichst in ihrer ursprünglichen Form und Funktion erhalten werden.[6] Ein weiteres Argument

4 RFC 5322. https://datatracker.ietf.org/doc/html/rfc5322. Zuletzt geprüft am 17.05.2024.

5 Message headers – Permaent Message Header Field Names. https://www.iana.org/assignments/message-headers/message-headers.xhtml. Zuletzt geprüft am 17.05.2024.

6 Siehe hierzu z. B. auch Kupper, Beda. „E-Mail-Archivierung“. In *Actualité archivistique suisse/Archivwissenschaft Schweiz aktuell. Travaux du certificat en archivistique et sciences de l’information/Arbeiten aus dem Zertifikat in Archiv- und Informationswissenschaften.* Hrsg. von Gilbert Contaz, Nicole Meystre-Schaeren, Barbara Roth Lochner, Andreas, Steigmeier. Baden: Hier und Jetzt, 2008. 88–117.

für diese Strategie ist die Tatsache, dass in der Forschung neben der qualitativen Auswertung von Archivgut immer häufiger auch quantitative Untersuchungsmethoden zum Einsatz kommen.[7]

3 Vorhandene Lösungsansätze

Um herauszufinden, wie sich das soeben formulierte Überlieferungsziel effektiv erreichen lässt, erfolgte zunächst die Überprüfung vorhandener Lösungen auf ihre praktische Tauglichkeit für die professionelle archivische Überlieferung. Die Ergebnisse dieser Recherche werden im folgenden Abschnitt erörtert und mit den bisherigen Überlegungen in Beziehung gesetzt.

Im Internet finden sich zahlreiche Angebote, die „revisionssichere E-Mail-Archivierung“ anbieten. Bei genauerer Betrachtung zeigt sich, dass diese Werkzeuge eigentlich für den Einsatz im Geschäftsalltag ausgelegt sind und den Anforderungen einer OAIS-konformen Langzeitarchivierung[8] nicht gerecht werden. Auch von Seiten der Archive wurden bereits Vorschläge gemacht. Diese reichen vom Konvertieren der E-Mails in verschiedene PDF-Formate bis hin zum Ausdrucken ausgewählter Nachrichten auf Papier.[9] Einige Gedächtnisinstitutionen setzen ausschließlich auf die Verzeichnung archivwürdiger E-Mails und somit auf die Übernahme im Rahmen der E-Aktenarchivierung. Allerdings reduziert dieser Ansatz die Überlieferung ausschließlich auf die Datenproduktion öffentlicher Organe. Private Überlieferungsstränge bleiben dabei unberücksichtigt. Ohnehin bleibt trotz entsprechender Vorgaben fraglich, ob wirklich alle relevanten Inhalte in die Akten überführt werden. Zwar mag das Ziel, E-Mails samt Informationen zu erhalten, in den genannten Beispielen weitgehend gewährleistet sein, allerdings entsteht bei allen Ansätzen der Verlust zahlreicher Funktionen und Zusammenhänge. Hinzu kommt, dass sich bestimmte Anhänge, wie zum Beispiel audiovisuelle oder weitere dynamische Dateiformate, nicht ausdrucken lassen.

Ein möglicher Ansatz für die Erhaltung des Funktionsumfangs könnte die Speicherung von E-Mail-Konten in einem der gängigen E-Mail-Formate, wie zum Beispiel MBOX[10] oder PST,[11] sein. Beide erwiesen sich jedoch für die digitale Langzeitarchivierung als ungeeignet und bringen auch für die spätere Nutzung eine Vielzahl von Nachteilen mit sich. Die bisher einzige Alternative zu diesem Ansatz stellt die an der Universität Stanford entwickelte Software ePADD dar, mit der aktuell allerdings nur eingeschränkt über mehrere Postfächer hinweg recherchiert werden kann.[12] Da zum Zeitpunkt der Recherche keine der hier vorgestellten Lösungen vollständig für die Umsetzung der erläuterten Ziele geeignet war, kam letztlich nur die Entwicklung eines eigenständigen Konzepts in Frage. Zunächst musste hierfür genauer untersucht werden, wie real existierende E-Mail-Postfächer aufgebaut sind und welche Hindernisse bei ihrer archivfachlichen Aufbereitung überwunden werden müssen.

4 Technische Barrieren

Um bei der Entwicklung nicht auf „synthetische“ Daten vertrauen zu müssen, wurde frühzeitig auf die Analyse von echten E-Mail-Konten gesetzt. Hierfür konnte auf Postfächer des Archivs der Max-Planck-Gesellschaft zurückgegriffen werden. Um ein möglichst breites Spektrum abzudecken, wurden acht verschiedene E-Mail-Konten verarbeitet und untersucht. Die enthaltenen Nachrichten entstanden zwischen 1995 und 2023, also in einem verhältnismäßig langen Zeitraum. Die Postfächer umfassten durchschnittlich rund 40 000 E-Mails. Das größte Konto enthielt über 133 000 Nachrichten und über 258 000 Anhänge. Bei einer eingehenden Untersuchung der Postfächer zeigte sich, dass die Archivierung von E-Mails aus technischer Sicht komplex ist.

Schadsoftware: Im Zuge der Erweiterung des E-Mail-Standards im Jahr 1996 wurde es möglich, an E-Mails beliebige Dateien anzuhängen. Dies ermöglichte es jedoch auch, E-Mails für die Übertragung von Viren zu verwenden, die spätestens im Rahmen einer Archivierung erkannt und unschädlich gemacht werden müssen.

Verschlüsselung: Bei Bedarf können E-Mails in verschlüsselter Form übertragen werden.[13] Eine Entschlüsselung von E-Mail-Inhalten ist in diesem Fall nur mit einem

7 Siehe z. B. Prom, Christopher J. „Preserving Email.“ *Digital Preservation Coalition (DPC). Technology Watch Report* 11-01 (2011). 4. http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf. Zuletzt geprüft am 17.05.2024.

8 DIN ISO 14721:2012.

9 Für einen Überblick siehe z. B. Allegrezza, Stefano. „Recent developments on E-Mail preservation: Towards the ultimate solution.“ Paper IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy. https://ceur-ws.org/Vol-3160/short11.pdf. Zuletzt geprüft am 13.05.2024.

10 RFC 4155. https://datatracker.ietf.org/doc/html/rfc4155. Zuletzt geprüft am 13.05.2024

11 [MS-PST]: Outlook Personal Folders (.pst) File Format. https://learn.microsoft.com/en-us/openspecs/office_file_formats/ms-pst/. Zuletzt geprüft am 17.05.2024.

12 https://www.epaddproject.org/. Zuletzt geprüft am 17.05.2024.

13 RFC 8551. https://datatracker.ietf.org/doc/html/rfc8551. Zuletzt geprüft am 17.05.2024.

privaten Schlüssel möglich, der dem Archiv nicht ohne weiteres nicht zur Verfügung steht. Dies birgt die Gefahr, dass sich potenziell archivwürdige E-Mails nicht mehr rekonstruieren lassen.

Signierte E-Mails: Wie zuvor angedeutet, können E-Mails signiert werden, um ihre Authentizität sicherzustellen.[14] Die technische Umsetzung erfolgt durch das Anhängen einer Zertifikatsdatei, die Informationen zur sendenden Person enthält. Damit Zertifikate auch in Zukunft noch verifizierbar sind, müssen sie so schnell wie möglich verarbeitet werden. Die Funktionsweise von Zertifikaten beruht auf dem Prinzip der Vertrauenskette. Ob die entsprechende Infrastruktur für die Überprüfung in 50 Jahren noch existiert, ist fraglich.

E-Mail-Formate: Die verbreitetsten Formate zur Speicherung von E-Mail-Postfächern sind die bereits genannten Container-Formate PST und MBOX sowie die dazugehörigen Formate für Einzelnachrichten, EML und MSG. PST ist ein proprietäres Dateiformat der Firma Microsoft, das ohne Hilfsmittel nicht menschenlesbar ist. Für die langfristige Archivierung ist dieses Format somit ungeeignet. Beim MBOX-Format werden hingegen alle Nachrichten aus einem Ordner innerhalb einer einzigen Datei gespeichert. Die Abgrenzung der einzelnen E-Mails erfolgt durch die Verwendung einer Leerzeile, gefolgt von dem Schlüsselwort „From“ und einem Zeilenumbruch. Aufgrund dieser Struktur ist für jede archivfachliche Bearbeitung oder Nutzung das Einlesen der gesamten MBOX-Datei notwendig, was in hohem Maße ineffizient ist und bei größeren Postfächern schnell sehr ressourcenintensiv werden kann. Ein zusätzliches Problem besteht darin, dass sich nach jedem Bearbeitungsschritt die Prüfsumme der Datei ändert. Der eigentliche Zweck einer Prüfsumme, nämlich der nahtlose Nachweis der Authentizität und Integrität, wird dadurch kompromittiert.

Besonderheiten von PST-Dateien: Das gängigste E-Mail-Programm für die Verwendung von PST-Dateien ist Microsoft Outlook. Bei der Verwendung einer E-Mail-Adresse in Outlook existieren zwei verschiedene Varianten der lokalen Speicherung und der Einbettung des E-Mail-Kontos in eine Exchange-Umgebung. Die lokale Speicherung ermöglicht das Extrahieren aller Informationen aus dem E-Mail-Postfach, insbesondere der E-Mail-Adressen. Sollte das Postfach in eine Exchange-Umgebung eingebunden sein, enthält die PST-Datei Anweisungen für den Exchange-Server anstelle der eigentlichen Informationen über die Kommunikationsparteien. Der Exchange-Server ist Teil einer Internetdomäne. In dieser Domäne übernimmt ein sogenannter Controller der Domäne die Übersetzung zwischen Personen der Domäne und deren E-Mail-Adressen. Dies hat zur Folge, dass PST-Dateien möglicherweise nur Namen und keine konkreten E-Mail-Adressen enthalten, was eine nachträgliche Rekonstruktion und Zuordnung erfordert.

Klarnamen: Eine E-Mail-Adresse besteht aus einem lokalen Teil und der Internetdomäne, diese werden durch das @-Symbol getrennt.[15] Der hintere Teil wird durch die Domäne festgelegt. Der lokale Teil kann entweder von der verwaltenden Organisation für diesen bestimmt werden oder bei einer privaten E-Mail-Adresse frei gewählt werden. Aus diesem Grund lässt sich die Identität der dazugehörigen Person nicht immer zweifelsfrei feststellen, was entsprechende technische Lösungen erforderlich macht.

Zeichenkodierung: Die Vielzahl von möglichen Betriebssystemen und Spracheinstellungen kann dazu führen, dass Texte mit Sonderzeichen und Symbolen in den unterschiedlichsten Zeichenkodierungen vorliegen. Da bei der Übernahme von Postfächern nicht bekannt ist, wie die jeweiligen Dateien kodiert sind, kann eine Konvertierung in gängige Formate, wie beispielsweise UTF-8,[16] notwendig sein. Probleme bei der Konvertierung können zu Anzeigefehlern führen.

5 Informationsdichte von E-Mail-Postfächern

Neben technischen Barrieren sind vor allem auch die riesigen Datenmengen hervorzuheben, die bei der E-Mail-Archivierung bewältigt werden müssen. Zurückzuführen ist dies in erster Linie auf Spam- und Werbemails, Rundschreiben sowie den oft sehr inflationären Umgang mit der elektronischen Kommunikation. Nachfolgend werden die Ergebnisse einer inhaltlichen Auswertung der acht analysierten Postfächer präsentiert. Vor dem Hintergrund des Umfangs bereits übernommener E-Mail-Postfächer, erscheint eine händische Bearbeitung undenkbar. Ein möglicher Lösungsansatz könnte in der Totalüberlieferung liegen, sofern entsprechende Speicherkapazitäten zur Verfügung stehen. Als Methode für die Auswahl archivwürdiger E-Mail-Konten hat sich international der sogenannte Capstone-Ansatz durchgesetzt, bei dem ausschließlich die E-Mails von be-

14 RFC 8551. https://datatracker.ietf.org/doc/html/rfc8551. oder RFC 4880. https://datatracker.ietf.org/doc/html/rfc4880. Beide zuletzt geprüft am 17.05.2024.

15 RFC 2822. https://datatracker.ietf.org/doc/html/rfc2822. Zuletzt geprüft am 17.05.2024.

16 RFC 3629. https://datatracker.ietf.org/doc/html/rfc3629. Zuletzt geprüft am 17.05.2024.

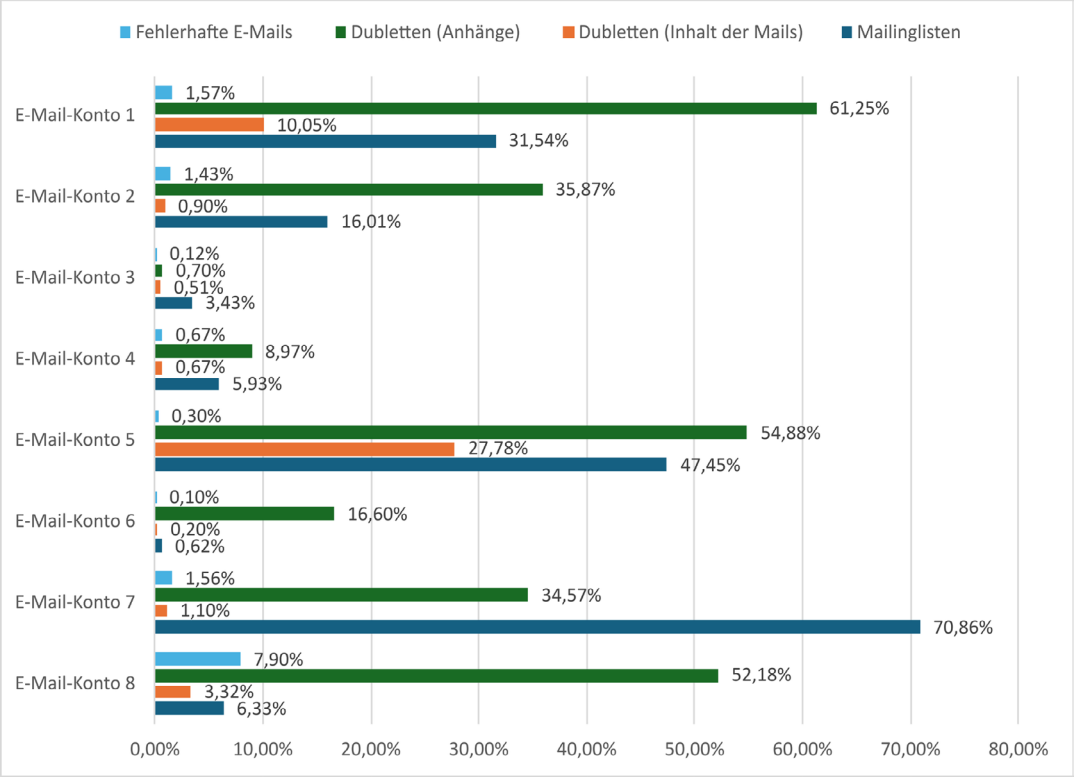


Abb. 1: Analyse der Zusammensetzung von acht E-Mail-Postfächern, die im Archiv der Max-Planck-Gesellschaft verwahrt werden (Grafische Darstellung: Nico Beyer)[17]

deutenden Schlüsselpersonen einer Organisation übernommen werden.[18] Die Untersuchung zeigte allerdings, dass die Menge tatsächlich archivwürdiger Nachrichten, unabhängig von der Position der Abgebenden, äußerst gering zu sein scheint. Aus diesem Grund ist eine umfangreiche und gezielte Datenreduktion auch innerhalb der ausgewählten E-Mail-Konten dringend anzuraten.

So zeigte sich zum Beispiel, dass die meisten analysierten Postfächer zu einem bemerkenswerten Anteil aus Dubletten, sowohl in Form identischer Dateianhänge als auch in Form identischer Inhalte, bestehen. Bei zwei Postfächern (5 und 8) handelte es sich bei etwas mehr als 50 Prozent der Anhänge um Dubletten. Bei Konto 1 waren es mehr als 60 Prozent. Zwei der untersuchten E-Mail-Postfächer (1 und 5) enthielten mit respektive 10 Prozent und 30 Prozent zudem eine nennenswerte Menge an inhaltlichen Dubletten.

Ähnlich zahlreich waren E-Mails, die sich mithilfe einer Analyse der E-Mail-Header eindeutig als Nachrichten aus Mailinglisten identifizieren ließen. Postfach 3 bestand sogar zu mehr als 70 Prozent aus diesem Nachrichtentyp. Werbe- und Spammails, die beim Versenden nicht als solche gekennzeichnet wurden, wurden bei dieser Untersuchung noch nicht einmal berücksichtigt.

Die Menge der E-Mails, die aufgrund von verschiedenen Defekten und Anomalien nicht ausgelesen werden konnten, war bei fast allen untersuchten Konten kaum erwähnenswert. Lediglich bei E-Mail-Konto 8 war mit rund 8 Prozent eine signifikante Menge der Nachrichten betroffen. Bei einem Virenscan zeigte sich, dass nur zwei der analysierten Konten eine geringe Anzahl von Dateianhängen enthalten, die eine genauere Untersuchung erforderlich machen. Bei Postfach 4 waren 30 (0,09 Prozent) und bei Postfach 7 insgesamt 19 (0,01 Prozent) der ausgewerteten E-Mails betroffen.

6 Rechtliche Aspekte

Zusätzlich zu technischen und inhaltlichen Aspekten sind vor allem auch die jeweils geltenden rechtlichen Rahmenbedingungen zu berücksichtigen. Dies ist umso wichtiger,

17 Aus datenschutzrechtlichen Gründen können die Untersuchungsergebnisse an dieser Stelle nur in anonymisierter Form präsentiert werden.

18 Siehe hierzu US National Archives and Records Administration. *White Paper on the Capstone Approach and Capstone GRS*. https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf. Zuletzt geprüft am 17.05.2024.

als es sich bei E-Mails um einen Unterlagentypus handelt, der – im Gegensatz zu vielen anderen digitalen Kommunikationswegen – zum Zeitpunkt des Absendens üblicherweise nicht für die breite Öffentlichkeit bestimmt war.[19] Eine fundierte rechtliche Grundlage kann dazu beitragen, eventuelle Bedenken von Nachlassgebern oder bestandsbildenden Institutionen auszuräumen und eine solide Vertrauensbasis zu schaffen.

Grundsätzlich sind bei der E-Mail-Archivierung dieselben juristischen Gegebenheiten zu beachten, die auch bei analogen Unterlagen zum Tragen kommen. In einigen Organisationen ist die Nutzung von E-Mail-Postfächern nur zu dienstlichen Zwecken zulässig. In anderen Fällen ist eine gemischte Nutzung zugelassen, woraus sich eine gänzlich andere rechtliche Situation ergibt. Eine rein dienstliche Nutzung impliziert eine Anbietungspflicht, sofern eine zuständige Gedächtnisinstitution existiert, während für die Übernahme privat oder gemischt genutzter Postfächer stets die Zustimmung der abgebenden Person notwendig ist, sofern keine anderen Regelungen existieren. Im Fall von privatem Archivgut sind vor allem die für die jeweilige Institution relevanten Schutzfristen zu berücksichtigen sowie eigentums-, datenschutz- und urheberrechtliche Aspekte zu klären. Im Nachlassbereich können hierfür Deposital- oder Schenkungsverträge verwendet werden, mit denen sich die dauerhafte Aufbewahrung der Postfächer in der jeweiligen Institution, die Zugänglichmachung im Rahmen der archivrechtlichen Schutzfristen sowie die Möglichkeit einer Schutzfristenverkürzung regeln lassen. Sofern möglich, sollte mit den abgebenden Personen darüber hinaus auch die Einräumung von urheberrechtlichen Nutzungsrechten für die verwahrende Institution vereinbart werden.

Um den Zugang zu privaten E-Mail-Postfächern zu regeln, stehen im archivischen Bereich verschiedene Optionen zur Verfügung, die von einer langfristigen Sperrung über die Notwendigkeit einer Zustimmung zur Nutzung durch die abgebende Person im Einzelfall bis hin zum Einverständnis zu einer direkten Zugänglichmachung reichen. Besonders herausfordernd ist der Umgang mit den personenbezogenen Daten Dritter, die in E-Mail-Postfächern, genau wie in traditionellen Briefwechseln, in signifikanten Mengen vorhanden sind. Für die Aufbewahrung an sich spielt dieser Umstand zunächst keine Rolle, da sich die Archivierung gemäß Artikel 14–19 und 89 DS-GVO als Löschungssurrogat auswirkt. Erst bei der Zugänglichmachung vor Ablauf der jeweils geltenden Sperrfrist müssen gegebenenfalls bestimmte Teile einer E-Mail oder eines Anhangs anonymisiert werden.

19 Vgl. Prom 2011. 5.

7 Projektrahmen

Nachfolgend wird der Projektrahmen vorgestellt, bevor schließlich der eigentliche Lösungsansatz in den Mittelpunkt der Betrachtung rücken kann. Um zur Überwindung der genannten Herausforderungen beizutragen, beschäftigt sich das EMILiA-Projektteam seit Anfang des Jahres 2024 mit der Entwicklung einer teilautomatisierten Software, die kulturelle Einrichtungen bei der Übernahme, Bewertung, Erschließung und Nutzbarmachung historisch relevanter E-Mail-Postfächer unterstützen wird. Aktuell besteht das Team aus einem Archivar und zwei Informatikern, die sich bereits im Vorfeld, unter anderem im Rahmen ihrer Qualifikationsarbeiten, mit verschiedenen Aspekten der E-Mail-Archivierung auseinandergesetzt haben. Bei EMILiA handelt es sich um ein kooperatives Forschungsprojekt der Freien Universität Berlin und dem Archiv der Max-Planck-Gesellschaft. Die Finanzierung des Entwicklungsvorhabens wird durch das Förderprogramm ProValid der Investitionsbank Berlin sichergestellt.[20]

8 Funktionsumfang

Hauptziel des EMILiA-Projekts ist es, interessierte Institutionen bei der Umsetzung der eingangs erläuterten Archivierungsstrategie zu unterstützten. Hierfür werden allen Beteiligten, also den abgebenden Personen, den Fachkräften und den Forschenden intuitive und teilautomatisierte Hilfsmittel zur Verfügung gestellt. Es wird eine Softwarelösung angestrebt, die von der Auswahl und Übernahme über die technische sowie inhaltliche Aufbereitung bis hin zur Recherche und Nutzung einige der wichtigsten archivischen Arbeitsschritte abdeckt.

Mit EMILiA wird es möglich sein, E-Mail-Postfächer sowohl für die Abgabe an ein geeignetes digitales Langzeitarchiv als auch für die Nutzung optimal vorzubereiten. Für die Langzeitarchivierung muss ein zusätzliches standardkonformes System zur Verfügung stehen. Entsprechende Exportmöglichkeiten werden bei der Entwicklung mitgedacht. Die Software wird aus den drei Teilmodulen „Übernahme“, „Bewertung und Erschließung“ und „Nutzung“ bestehen. Alle Module werden so konzipiert, dass sie auch nach der Auslieferung noch durch weitere Funktionen und Schnittstellen ergänzt werden können. Wichtigste Grundlage des Konzepts ist das OAIS-Referenzmodell für die digitale Lang-

20 Für weitere Informationen zum Förderprogramm siehe IBB Business Team. *ProValid fördert Forschungsprojekte von Berliner Hochschulen.* https://www.ibb-business-team.de/provalid/. Zuletzt geprüft am 17.05.2024.

Abb. 2: Überblick über den Funktionsumfang von EMILiA (Grafische Darstellung: Nico Beyer)

zeitarchivierung. Nachfolgend werden die Funktionen der einzelnen Teilmodule genauer beleuchtet.

8.1 Übernahme

Das erste Modul setzt sich aus einer leicht bedienbaren Software für die Abgebenden und einem Gegenstück für die zuständigen Mitarbeitenden der jeweiligen Gedächtnisinstitution zusammen. Für die Abgabe von E-Mails können drei verschiedene Optionen verwendet werden. Die erste Option entspricht im Wesentlichen derselben Methode, die auch von klassischen E-Mail-Programmen für die Synchronisierung von Konten eingesetzt wird. Bei diesem Ansatz müssen sich die abgebenden Personen lediglich auf einem eigens für die Abgabe zur Verfügung gestellten Mailserver anmelden und ihr ursprüngliches Postfach mithilfe ihrer Anmeldedaten importieren. Die zweite Möglichkeit ist das eingangs genannte Abgabetool, mit dem sich sowohl MBOX- als auch PST-Dateien sicher über das Internet an die zuständige Institution übertragen lassen. Hierfür muss auf Seite der Institution eine Abgabe angelegt und gestartet werden. Da E-Mail-Postfächer im Regelfall riesige Datenmengen enthalten, kann diese Form der Abgabe einige Zeit in Anspruch nehmen. So ergaben Berechnungen des Projektteams, dass die vollständige Übernahme von einem E-Mail-Konto mit einer Größe von knapp 70 GB bei einer durchschnittlichen Internetgeschwindigkeit von 30 Mbit/s rund 5 Stunden in Anspruch nimmt.[21] Da dieser Prozess im Hintergrund ausgeführt werden kann, kann der verwendete Rechner während dieser Zeit aber selbstverständlich für andere Aufgaben verwendet werden.

Abb. 3: Bildschirmfoto des funktionierenden Prototyps für das Abgabetool (Bildschirmfoto: Felix Gericke)

Sollte es aufgrund mangelnder Internetgeschwindigkeit oder wegen erhöhter Sicherheitsbedenken notwendig sein, kann das Abgabetool aber auch dafür verwendet werden, die Daten auf ein physisches Speichermedium zu übertragen. Unabhängig davon, welche der drei Methoden für die Datenübertragung ausgewählt wird, entsteht am Ende des Prozesses ein BagIt-Container[22] mit einer Strukturdatei, den E-Mails im TXT- oder HTML-Format, den unveränderten Anhängen sowie PREMIS-Metadaten[23] und Prüfsummen für jede übernommene E-Mail.[24] Jeder Verarbeitungsschritt wird mithilfe des PREMIS-Standards dokumentiert, um die Authentizität und Integrität der Daten zu gewährleisten. Die Strukturdatei ist eine menschenlesbare XML-Datei und enthält die Sende- und Empfangsparteien, den Betreff und das Sendedatum einer E-Mail.[25] Zusätzlich wird die vorhan-

21 Das für die Berechnungen zugrunde gelegte Postfach umfasst etwa 104 000 E-Mails und rund 212 000 Anhänge.

22 RFC 8493. https://datatracker.ietf.org/doc/rfc8493/. Zuletzt geprüft am 17.05.2024.

23 PREMIS Editorial Committee. *PREMIS Data Dictionary for Preservation Metadata.* version 3.0. https://www.loc.gov/standards/premis/. Zuletzt geprüft am 17.05.2024.

24 Ob E-Mails im TXT- oder HTML-Format gespeichert werden, hängt von der Beschaffenheit ihres Inhalts ab.

25 XML-Schema der EMILiA-Strukturdatei. https://schema.emilia-archiv.de/structure/1.0.0/structure.xsd. Zuletzt geprüft am 17.05.2024.

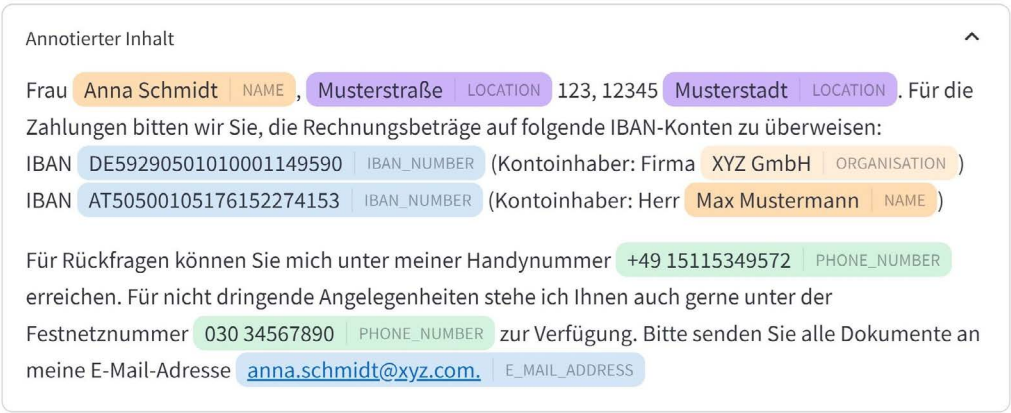


Abb. 4: Bildschirmfoto des funktionierenden Prototyps für die automatisierte Erkennung und Schwärzung personenbezogener Daten (Bildschirmfoto: Nico Beyer)

dene Ordnerstruktur mit Informationen über die Anhänge des E-Mail-Kontos hinterlegt.

8.2 Bewertung und Erschließung

Das zweite Teilmodul umfasst verschiedene Funktionen für die teilautomatisierte Datenreduktion, inhaltliche Bewertung und Erschließung der übernommenen Postfächer. Theoretisch können E-Mail-Postfächer direkt nach der Übernahme an ein digitales Langzeitarchiv übergeben oder genutzt werden. Um von allen Funktionen von EMILiA profitieren zu können, ist eine Vorverarbeitung notwendig. Die meisten Schritte dieses Prozesses sind optional und können je nach Bedarf ein- oder ausgeschaltet oder nach Priorität sortiert werden. Neben einer obligatorischen Indexierung der Daten, können im Rahmen der Vorverarbeitung eine Spam- und Dublettenerkennung, eine Formatkonvertierung sowie eine Extraktion der in den E-Mails enthaltenen Personen, Körperschaften, Themen und Orte durchgeführt werden. Im Rahmen seiner Qualifikationsarbeit war es einem Mitglied des Projektteams außerdem möglich, automatisiert Klarnamen aus E-Mail-Adressen zu extrahieren sowie E-Mail-Verläufe zu rekonstruieren. Klarnamen können nicht nur aus vorhandenen Headerinformationen, sondern auch aus textuellen Inhalten abgeleitet werden. Eine mögliche Identifizierung erfolgt durch eine Auswertung von Attributen wie Anreden, Grußformeln und Signaturen. Die gesammelten Informationen werden separat in einer Adressen-Namen-Datenbank gespeichert. Perspektivisch wäre es auch denkbar, bildliche Anhänge einer OCR-Texterkennung zu unterziehen, wodurch sich gegebenenfalls auch Dokumentenscans oder ähnliche Bilddateien mithilfe einer Volltextsuche auffindbar machen ließen. Die wohl wichtigste Funktion für die archivische Aufbereitung und Nutzbarmachung von E-Mail-Konten dürfte aber wohl die automatisierte Erkennung von personenbezogenen Daten sein. Mithilfe dieses bereits erfolgreich getesteten Moduls wird es möglich sein, sensible Daten in E-Mail-Headern und -Inhalten ausfindig zu machen und für eine optionale Schwärzung vorzuschlagen.[26] Nur durch diesen Schritt ist es möglich, E-Mails bereits vor Ablauf der Schutzfristen rechtskonform zugänglich zu machen.

Sobald die Vorverarbeitung abgeschlossen ist, können die übernommenen Postfächer entweder direkt an ein digitales Langzeitarchiv übergeben oder für die weitere Bearbeitung geöffnet werden. Wichtig ist es, darauf hinzuweisen, dass endgültige Bewertungsentscheidungen nach wie vor von den zuständigen Mitarbeitenden getroffen werden. EMILiA markiert lediglich Nachrichten, die höchstwahrscheinlich kassiert werden können. Neben verschiedenen Such- und Filteroptionen bietet die grafische Oberfläche auch eine Reihe von verschiedenen Statistiken sowie die Möglichkeit, E-Mails unter der Angabe von Gründen manuell für die Benutzung zu sperren oder gänzlich zu löschen. Weiterhin können bereits vorhandene Ordnerstrukturen angepasst oder neue Ordner erstellt werden. Ebenso wird ein Export ausgewählter Inhalte möglich sein.

8.3 Nutzung

Mit den Funktionen des dritten Moduls können die Daten schließlich aus dem digitalen Langzeitarchiv exportiert und für die Nutzung vorbereitet werden. Um für Anfragen

26 Der Prototyp kann auf der Website des Projekts getestet werden. Obwohl wir keinerlei Daten speichern oder weiterleiten, empfehlen wir aus rechtlichen Gründen, keine sensiblen Daten für die Tests zu verwenden. https://demo.emilia-archiv.de/. Zuletzt geprüft am 13.05.2024.

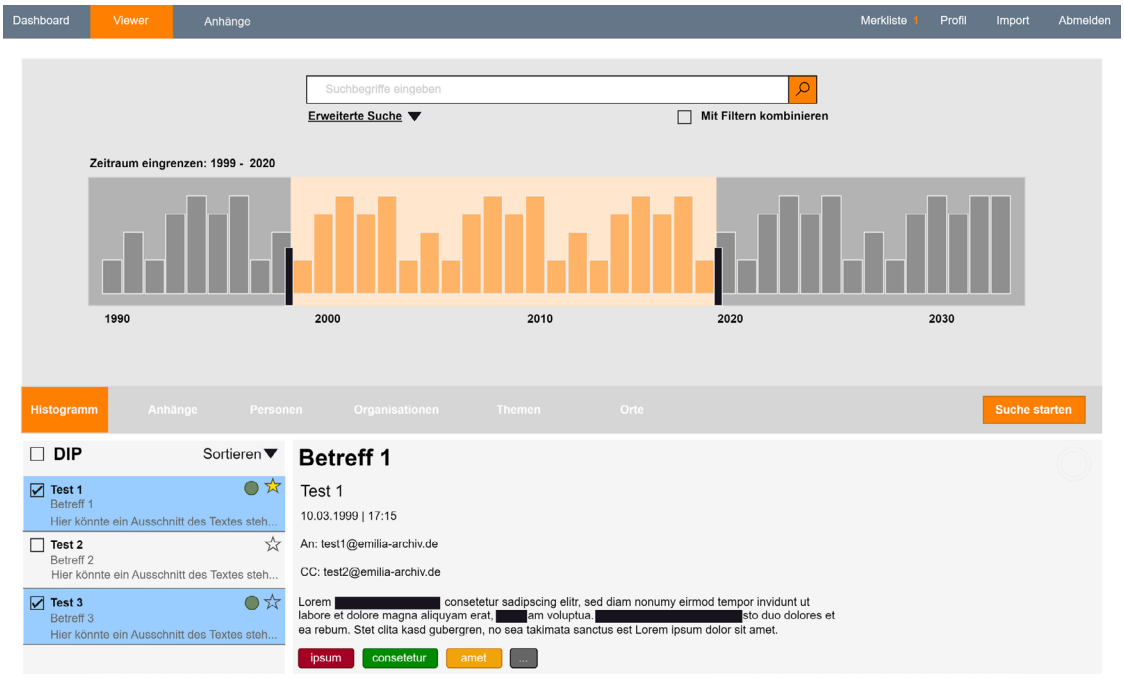


Abb. 5: Konzept für den Viewer (Grafische Darstellung: Nico Beyer)

relevante E-Mails zu finden, wird den Verantwortlichen eine Recherchedatenbank zur Verfügung stehen, mit deren Hilfe über mehrere E-Mail-Postfächer hinweg recherchiert werden kann. Da diese Datenbank potenziell auch sensible Informationen enthalten kann, ist vorerst ausschließlich eine Nutzung durch das Personal der verwahrenden Institution angedacht. Sobald mithilfe der Datenbank alle relevanten Nachrichten ausgewählt sind, wird es möglich sein, personenbezogene Daten zu anonymisieren und gegebenenfalls problematische Anhänge auszuschließen. Für den Fall, dass bei der Vorverarbeitung nicht alle Inhalte erkannt wurden, kann zudem auch manuell anonymisiert werden. Im Anschluss können die Daten an Forschende weitergegeben werden. Diese können die E-Mails dann in einer vereinfachten Version der grafischen Oberfläche durchsuchen, filtern, betrachten.

9 Fazit und Ausblick

Inzwischen ist die Konzeptphase weitgehend abgeschlossen. Derzeit ist das Projektteam mit der praktischen Umsetzung der einzelnen Programmfunktionen sowie dem Austausch mit der Fachcommunity befasst. Die Übernahme von E-Mail-Konten funktioniert bereits reibungslos. Zahlreiche Funktionen des zweiten Teilmoduls können nach erfolgreicher Testphase implementiert werden. Verschiedene Ansätze für die Filterung von Spam, Dubletten und schadhaften Anhängen werden derzeit mithilfe von Testdaten evaluiert. Darüber hinaus existiert ein Prototyp für die Darstellung und Filterung von übernommenen Daten, der stetig um weitere Funktionen ergänzt wird.

Nach dem Projekt ist die Gründung eines Start-up-Unternehmens geplant. Diese Vorgehensweise ermöglicht es, die Software auch nach der Projektlaufzeit und der Auslieferung an Kunden aktiv zu pflegen und stetig bedarfsgerecht zu erweitern. Darüber hinaus können Schulungen angeboten und nutzende Einrichtungen bei technischen Problemen unterstützt werden.

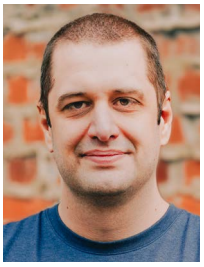
Das Projektteam bietet aktuell in regelmäßigen Abständen Webinare an, bei denen sich Interessierte über Fortschritte informieren und aktiv in die Entwicklung einbringen können. Weitere Informationen finden sich auf der Website des Projekts.[27] Auch außerhalb von offiziellen Veranstaltungen sind Vorschläge und Kooperationsangebote jederzeit willkommen.

27 https://www.emilia-archiv.de/. Zuletzt geprüft am 13.05.2024.

Autoren

Nico Beyer
Freie Universität Berlin
Fachbereich Mathematik und Informatik
EMILiA-Projekt
Takustraße 9
14195 Berlin
nico.beyer@emilia-archiv.de
https://orcid.org/0009-0003-8984-3572

Alexander Hinze-Hüttl
Archiv der Max-Planck-Gesellschaft
EMILiA-Projekt
Boltzmannstraße 14
14195 Berlin
alexander.hinze-huettl@emilia-archiv.de
https://orcid.org/0009-0009-7040-1110



Felix Gericke
Archiv der Max-Planck-Gesellschaft
EMILiA-Projekt
Boltzmannstraße 14
14195 Berlin
felix.gericke@emilia-archiv.de
https://orcid.org/0009-0009-0916-2202