ETH zürich

ETH-Bibliothek

27. AUdS-Tagung 2024

Langzeitarchivierung von E-Mails an der ETH Zürich

Claudia Briellmann, Fabian Schneider

Agenda

- **Grundlagen**
- **Bewertung und Erschliessung von E-Mails**
- **Langzeitarchivierung von E-Mails**
- **Stand und nächste Schritte**
- **Herausforderungen und unsere Empfehlungen**

Grundlagen

Was ist das Projekt E-Mail Archiv 2.0?

- **Projekt E-Mail Archiv 2.0 (2021-2023)**
  - Projektleitung: Informatikdienste der ETH Zürich (ID)
  - In Zusammenarbeit mit der ETH-Bibliothek: u.a. Hochschularchiv der ETH Zürich (HSA) und dem Forschungsdatenmanagement und Datenerhalt (FDD)
- **Projektteile, die in den Bereich der ETH-Bibliothek fallen:**
  - Rechtliche Verpflichtung zur Sicherung von Verwaltungsunterlagen
  - Bewertung und Archivierung von E-Mails (retrospektiv)
  - Ausarbeitung Workflow (prospektiv)

    - ➢ Bewertung, Übernahme, Erschliessung, Langzeitarchivierung
    - ➢ Übernahme aus einem Mailarchiv, das keine Langzeitarchivierungslösung ist
    - ➢ Aufbereitung für Benutzung nach Ablauf Schutzfrist (besonders schützenswerte Personendaten)

Ausgangslage

- **Unser Grundsatz**: Vorhandene Tools nutzen
  - Mailarchiv: das Quellsystem
  - CMI AIS: vom HSA genutztes Archivsystem
  - Apache Tika: Tool zur Metadaten-Extraktion
  - Foxit: PDF-Compressor (ursprüngliche Tests mit 3-Heights® Document Converter)
  - Rosetta: Langzeitarchivsystem von ExLibris

- **Unsere Anforderungen**
  - Dauerhafte Langzeitarchivierung
  - Ressourcensparend vorgehen
  - Auffindbarkeit und Nachnutzung von E-Mails (bzw. den Inhalten) gewährleisten

Bewertung und Erschliessung

Die Bewertung von *Big Data*

- **Grundlegende Frage**
  - **«Can we keep everything?»** (Yeo, Geoffrey: Can we keep everything? The future of appraisal in a world of digital profusion, in: Brown, Caroline (Hg.): Archival Futures, London 2019, S. 45–63.)
    - Unterlagen werden bereits digital produziert und können so übernommen werden
    - Speicherplatz ist mittlerweile sehr billig
    - Metadaten geben bereits viel Auskunft über Inhalte
    - Recherchemöglichkeiten in grossen Beständen werden mit dem technischen Fortschritt immer besser (Auffindbarkeit kann besser gewährleistet werden)
    - Datenintegrität für *Big Data* zu gewährleisten ist sehr ressourcenintensiv
- **Gängige Konzepte für die Bewertung von E-Mails**
  - E-Mails sind *Big Data* und somit zu gross für Einzelstück-Bewertung durch Personen. Gängige Modelle:
    1. Capstone-Approach (National Archives and Records Administrations (NARA): White Paper on The Capstone Approach and Capstone GRS, 2015.)
    2. Übernahme aus RMS (David Bearman: Managing electronic mail, in: Archives & Manuscripts 22 (1), 1994)
    3. Hot-Spot-Listen (Nationaal Archief Niederlande: https://www.nationaalarchief.nl/archiveren/kennisbank/hotspotlijst-maken)

The Capstone-Approach

«Final disposition is based on the role or position of the end-user, not the content of each individual email record.»

(National Archives and Records Administrations (NARA):
White Paper on The Capstone Approach and Capstone GRS, 2015, S. 7)

- Konzept von NARA (National Archives and Records Administration)
  - 2013: Guidance on a New Approach to Managing Email Records.
  - 2015: White Paper on The Capstone Approach and Capstone GRS.
    - Angedacht als Ergänzungslösung

- **Vorteile des Capstone-Approach**
  - Es müssen nicht unzählige E-Mails einzeln bewertet werden.
  - In einer gut strukturierten Institution wie der ETH Zürich lassen sich Personen, die für das Abbilden des Verwaltungshandelns relevant sind, gut definieren.
    - Unsere Capstones: Mitglieder Schulleitung, Präsident:innen und Geschäftsführer:innen ETH-Rat
    - **Nur E-Mails aus den Amtsperioden werden übernommen!** (Laufzeit auf den Tag genau)

Erschliessung

- **Gliederung**
  - Serienbildung nach Person und Account (wenn eine Person mehrere Accounts hat, gibt es auch mehrere Serien)
  - Innerhalb der Serien werden Jahresdossiers angelegt
    - ➢ Dies ermöglicht die kontinuierliche Freigabe von Jahresdossiers nach Ablauf der Schutzfrist
  - Keine inhaltliche Erschliessung → automatische Metadatenextraktion auf Dateiebene (und Anzeige in Rosetta)

- **Schutzfristen**
  - Nach Bundesgesetz über die Archivierung (BGA): 50 Jahre für besonders schützenswerte Personendaten
    - Nach Bundesgesetz über den Datenschutz (CH), Art. 3: Daten über:
      1. die religiösen, weltanschaulichen, politischen oder gewerkschaftlichen Ansichten oder Tätigkeiten,
      2. die Gesundheit, die Intimsphäre oder die Rassenzugehörigkeit,
      3. Massnahmen der sozialen Hilfe,
      4. administrative oder strafrechtliche Verfolgungen und Sanktionen
  - **Ohne inhaltliche Einzelbewertung muss davon ausgegangen werden, dass die E-Mails besonders schützenswerte Personendaten enthalten.**

Langzeitarchivierung

Grundlegende Fragen

- Dateiformate und Metadaten
- CSV-Ingest nach Rosetta
- Workflow (vereinfacht)
- Stand und nächste Schritte

Formate

- Empfohlene Dateiformate für Langzeitarchivierung: MBOX und EML
- Exportierte Formate PST, MSG nur bedingt geeignet (Abhängigkeit Outlook), aber:
  - enthalten für uns wichtige Metadaten (Angabe zu Attachement)
  - erlauben (vorerst) native Nutzung der Dateien
- Unsere Strategie:
  - Erhaltung des Postfachs im Originaldateiformat (PST)
  - Konvertierung der MSG-Dateien nach PDF/A (inkl. Metadaten-Extraktion)
  - Archivierung der PST und PDF/A-Dateien
- Zur PDF-Konvertierung:
  - Konvertierung mit oder ohne Anhang möglich
  - Erfolgsrate mit 3-Heights® Document Converter: mit Anhang 80%, ohne Anhang 99.8%

Metadaten

Auswahl der Metadaten und Mapping mit den entsprechenden Feldern in Rosetta:

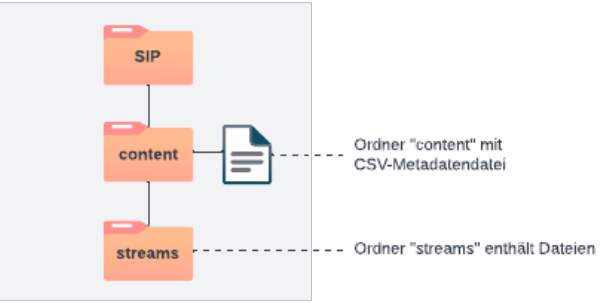
Ebene	TIKA	DC-Feld Rosetta
File	Message:From-Name	FILE - Creator (DC)
File	Message:From-Email	FILE - Creator (DC)
File	Message-To	FILE - Contributor (DC)
File	Message-Recipient-Address	FILE - Contributor (DC)
File	Message-Cc	FILE - Contributor (DC)
File	Message-Bcc	FILE - Contributor (DC)
File	dc:title	FILE - Description (DC)
File	dcterms:created	FILE - Date (DC)
File	Message:Raw-Header:X-MS-Has-Attach	FILE - Description (DC)
File	meta:mapi-message-class	FILE - Type (DC)

Metadaten in Rosetta:

FILE - Creator : Message:From-Name: [redacted]
FILE - Creator : Message:From-Email: [redacted]
FILE - Description : dc:title: WG: [redacted]
FILE - Description : Message:Raw-Header:X-MS-Has-Attach: yes
FILE - Contributor : Message-To: [redacted]
FILE - Contributor : Message-Recipient-Address: [redacted]
FILE - Contributor : Message-Cc: [redacted]
FILE - Date : dcterms:created: 2005-05-30T15:03:55Z
FILE - Type : meta:mapi-message-class: NOTE

CSV-Ingest nach Rosetta

Generelle Struktur für CSV-Ingest



CSV-Metadatendatei

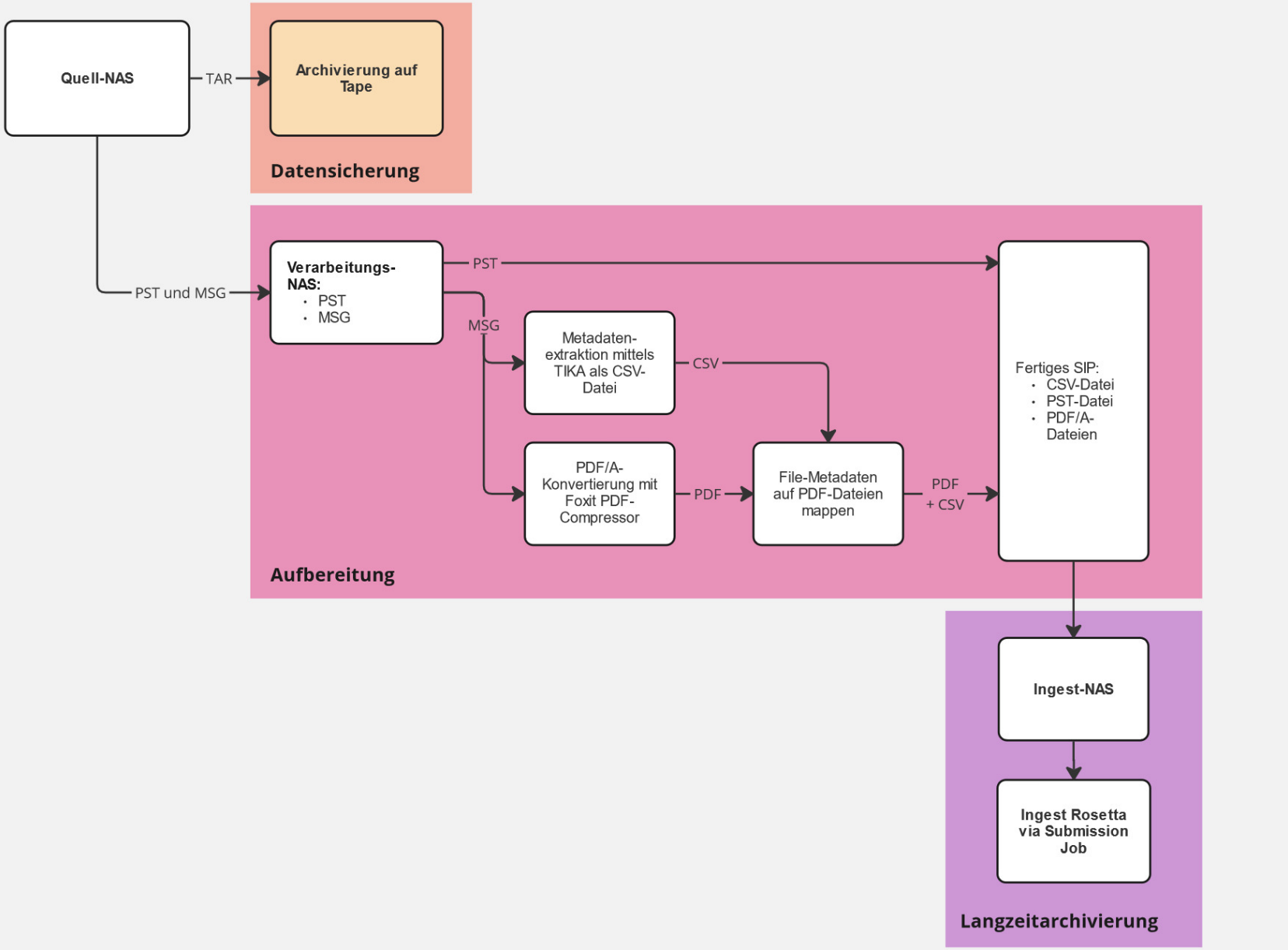
IE-Ebene (Postfach):

	A	B	C	D	E	F	
1	Object Type	Title (DC)	Date (DC)	Relation (DC)	Identifier (DC)	Publisher (DC)	Rights (DC)
2	SIP	ER-3/1.2 E-Mails von [redacted]					
3	IE	ER-3/1.2 E-Mails von [redacted]	2005	ER-3/1.2	385353	ETH-Bibliothek, Hochschularchiv der ETH Zürich	Gemäss Bundesgesetz über die Archivierung
4	REP						
5	File						
6	File						
7	File						

File-Ebene (einzelne E-Mails):

FILE - Creator (DC)	FILE - Creator (DC)	FILE - Contributor (DC)	FILE - Contributor (DC)	FILE - Description (DC)	FILE - Date (DC)	FILE - Type (DC)
Message:From-Nam	Message:From-Email:	/O=ETHZ/OU=ETHZ-MDL/CN=RECIPIENTS/CN=		dc:title: [redacted] MIT	dcterms:created: 2008-01-30T12:29:32Z	meta:mapi-message-class: CONTACT
Message:From-Nam	Message:From-Email:	/O=ETHZ/OU=ETHZ-MDL/CN=RECIPIENTS/CN=		dc:title: [redacted] -	dcterms:created: 2008-02-15T13:49:56Z	meta:mapi-message-class: CONTACT
Message:From-Nam	Message:From-Email:	/O=ETHZ/OU=ETHZ-MDL/CN=RECIPIENTS/CN=		dc:title: Sitzungszimmer	dcterms:created: 2008-02-06T08:31:40Z	meta:mapi-message-class: CONTACT

Workflow (vereinfacht)



Und wo stehen wir jetzt?

- Übernahme und Datensicherung Altbestand abgeschlossen
- Neues Quellsystem bei ID erfordert möglicherweise Anpassungen bei Übernahme
- Optimierung der PDF/A-Konvertierung durch Einsatz von Foxit PDF-Compressor (Workflow in Arbeit)
- Script entwickeln für das Mapping der Metadaten aus den MSG-Dateien mit den entsprechenden PDF/A-Dateien in der CSV-Datei
- Umsetzung Workflow Hochschularchiv und Forschungsdatenmanagement und Datenerhalt

Herausforderungen und Empfehlungen

Zahlen und Fakten

Was	Anzahl
Postfächer	42
Jahrgänge	230
E-Mails	2’247’634
Speicherbedarf	1.1 TB

Das haben wir gelernt!

Herausforderung	Empfehlung
Auswahl Metadaten für Archivierung	Verschiedene Metadaten-Felder analysieren und für Nachnutzung geeignete auswählen
PDF-Konvertierung von Anhängen nicht immer optimal	Qualitätskontrolle und Originaldaten immer mitarchivieren!
Viele kleine Dateien: Zeit- und Ressourcenintensiv bei Kopiervorgängen, Verarbeitung und Speicherung	Nutzung von Archivformaten (zip, tar) für Kopiervorgänge, Verarbeitung timen
Viren in E-Mails	Auf Virenwarnung vorbereiten, betroffene E-Mails/Postfächer dokumentieren
Verschlüsselte E-Mails	Policy anpassen (geschäftsrelevante E-Mails unverschlüsselt ablegen)
Fehlendes RMS	Bei vorhandenem RMS ggf. weniger Bedarf ganze Postfächer zu archivieren

ETH zürich

ETH-Bibliothek

Vielen Dank für die Aufmerksamkeit

Fragen?

ETH zürich

Claudia Briellmann
Fabian Schneider

claudia.briellmann@library.ethz.ch
fabian.schneider@library.ethz.ch

ETH-Bibliothek
Rämistrasse 101
8092 Zürich

www.library.ethz.ch