ETH zürich

ETH-Bibliothek

Frühjahrstagung: Records Management und Digitale Archivierung

Langzeitarchivierung von E-Mails an der ETH Zürich

Projekt E-Mail Archiv 2.0

Claudia Briellmann, Fabian Schneider

Agenda

- **Projekt E-Mail Archiv 2.0**
- **Unsere Grundlagen**
- **Bewertung und Übernahme von E-Mails**
- **Erschliessung von E-Mails**
- **Geeignete Langzeitarchivierung von E-Mails**
- **Stand und nächste Schritte**
- **Herausforderungen und unsere Empfehlungen**

Projekt E-Mail Archiv 2.0

Was ist das Projekt E-Mail Archiv 2.0?

- **Projekt E-Mail Archiv 2.0**
  - Projektleitung: Informatikdienste der ETH Zürich (ID)
  - In Zusammenarbeit mit: u.a. Hochschularchiv (HSA), Forschungsdatenmanagement und Datenerhalt (FDD)
  - Laufzeit: 2021 – 2023
- **Mailarchiv**
  - Kopien von E-Mails und Kalendereinträgen im Archiv-Store (separates Speichersystem)
    - Von Daten älter als 30 Tage (Stand 01.05.2022 neu 60 Tage)
    - Älteste E-Mail von 1998
    - Keine Langzeitarchivierungslösung
- **Projektteile, die in den Bereich der ETH-Bibliothek fallen:**
    - Bewertung und Archivierung von E-Mails (retrospektiv)
    - Ausarbeitung Workflow (prospektiv)
      - ➢ Bewertung, Übernahme, Erschliessung, Langzeitarchivierung

Rechtliche Aspekte

- Das Hochschularchiv hat folgenden Auftrag:
  «Das Archiv der ETH Zürich nimmt die Aufgaben eines öffentlichen Archivs gemäss dem BGA für die gesamte ETH Zürich sowie den ETH-Rat wahr. Es leistet damit einen Beitrag zur Rechtssicherheit sowie zur kontinuierlichen und rationellen Verwaltungsführung und schafft insbesondere Voraussetzungen für die historische und sozialwissenschaftliche Forschung. Damit dient es sowohl den Verwaltungen der ETH Zürich und des ETH-Rates als auch der Lehre und Forschung. Darüber hinaus sorgt es auch dafür, dass die Interessen der Öffentlichkeit im Rahmen der gesetzlichen Bestimmungen wahrgenommen werden können. Das Archiv wirkt zudem aktiv an der Erforschung und Vermittlung der Geschichte der Hochschule mit.»

  (ETH Zürich: Reglement für das Archiv der ETH Zürich, 01.04.2015. Online: <https://rechtssammlung.sp.ethz.ch/Dokumente/420.1.pdf> Art. 1.)

- Beinhaltet Überlieferung von E-Mails
- keine strukturierte und sichere Ablage von geschäftsrelevanten E-Mails → Archivierung aus Mailarchiv
- Schutzfrist für E-Mails (Personendaten)
- Kommunikation seit 01.05.2022: Neue **Benutzungsordnung für Informations- und Kommunikationstechnologie an der ETH Zürich (BOT)** (Online: <https://rechtssammlung.sp.ethz.ch/Dokumente/203.21.pdf>)

Grundlagen

Ausgangslage

- **Unser Grundsatz**: Vorhandene Tools nutzen
  - Mailarchiv
  - CMI AIS
  - Apache Tika
  - 3-Heights® Document Converter (2023 geplanter Wechsel auf PDF Compressor)
  - Rosetta

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

Zahlen und Fakten

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

Bewertung und Übernahme

Die archivische Bewertung

- **Früherer Ansatz: möglichst viel archivieren**
  - Bestände waren kleiner
  - Informationen waren spärlich
  - Möglichst viel Informationen sollten überliefert werden

- **Vorteile von Bewertung**
  - Der Wert von Material kann definiert werden
  - Archivwürdiges Material geordnet und einfach nutzbar
  - Gutes Dokumentenmanagement = bereits bei Eröffnung eines Geschäfts definiert, was archiviert wird

- **Kritik an Bewertung**
  - Bewertungskriterien sind arbiträr und inkonsistent
  - Bewertung hängt immer an einer Person und daran, was diese wichtig findet

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

- **Mögliche Konzepte für die Bewertung von *Big Data***
  - *Big Data* zu gross für Einzelstück-Bewertung durch Personen. Gängige Modelle:
    1. Komplette Übernahme
    2. Komplette Kassation
    3. Übernahme eines Sample

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 ETH-Rat, Geschäftsführer ETH-Rat
    - **Nur E-Mails aus den Amtsperioden werden übernommen!** (Laufzeit auf Tag genau)

Andere Ansätze zur Bewertung von E-Mails

- **Übernahme von E-Mails aus RMS** (Records Management System)
  - Voraussetzung: Geschäftsrelevante E-Mails müssen korrekt abgelegt werden
  - Nachteile:
    - Entstehungskontext der E-Mails geht verloren
    - Bewertung erfolgt nicht durch Archivmitarbeitende
  - ➢ Gibt es zum jetzigen Zeitpunkt an der ETH nicht

- **Archivierung mit Hilfe von Hot-Spot-Listen:**
  - Wird vom Nationaal Archief in den Niederlanden eingesetzt
  - Es werden vorgängig Themen definiert, zu denen E-Mails archiviert werden
  - Basierend auf Trendanalysen

Was ist der beste Ansatz zur Bewertung von E-Mails?

- **Lösungen müssen auf die jeweiligen Archive angepasst werden**
  - Mehrere Ansätze können parallel genutzt werden = Absicherung
  - Vorhandene Tools zur Hilfe nehmen
  - Bei *Big Data* weniger zeitaufwändige Lösung (technische Unterstützung)

  - **Wichtig ist, dass archivwürdige E-Mails in geeigneter Weise langzeitarchiviert werden und (nach Ablauf einer Schutzfrist) auffindbar sind.**

Erschliessung

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 Schutzfrist
  - Keine inhaltliche Erschliessung → automatische Metadatenextraktion auf Dateiebene (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

- Formate
- Auswahl Metadaten und Mapping
- PDF/A-Konvertierung
- CSV-Ingest nach Rosetta
- Langzeitarchivierung in Rosetta
- Workflow
- Stand und nächste Schritte

Formate

- Exportierte Formate PST, MSG bedingt geeignet (Abhängigkeit Outlook)
- Geeignet: nicht proprietäre, offene Formate
  - Wissen über Aufbau
  - Spezifikation zugänglich und nutzbar für Formatvalidierungen
- MBOX und EML besser geeignet für Langzeitarchivierung
- PDF/A: Standard-Format für textbasierte Dateien für Langzeitarchivierung
- Risiken bei Konvertierung:
  - Informationsverlust (z.B. Angabe über Anhang oder Klassifikation)
  - Gefährdung Authentizität
  - Ressourcen
- Daher:
  - Erhaltung der Originaldateien (PST, MSG)
  - Konvertierung der MSG-Dateien nach PDF/A

Metadaten

- Metadaten auf Ebene E-Mail für Nachnutzung notwendig
- Herausforderungen bei der Evaluierung der zu extrahierenden Metadatenfelder
  - Formatabhängig **unterschiedliche Felder** für dieselben Metadaten
  - Unterschiedliche **Formatierung** der Inhalte über Felder und Formate hinweg
  - Unterschiedliche **Extraktionsrate** über Felder und Formate hinweg

Beispiel Datum (unterschiedliche Formatierung)

Message:raw-header:date	dcterms:created	dcterms:modified
Wed, 19 Jul 2006 10:07:44 +0200	2006-07-19T08:07:44Z	

	Message-To	Message:Raw-Header:To	Message:To-Email	Message:To-Display-Name	Message:To-Name
MSG	Name ***oder*** Adresse	Name ***und*** Adresse	***Nur*** Adresse	***Nur*** Name	
EML	Name ***und*** Adresse	[wird nicht genutzt]			

Metadaten

Auswahl der Metadaten der MSG-Dateien und der entsprechenden Felder 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)

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

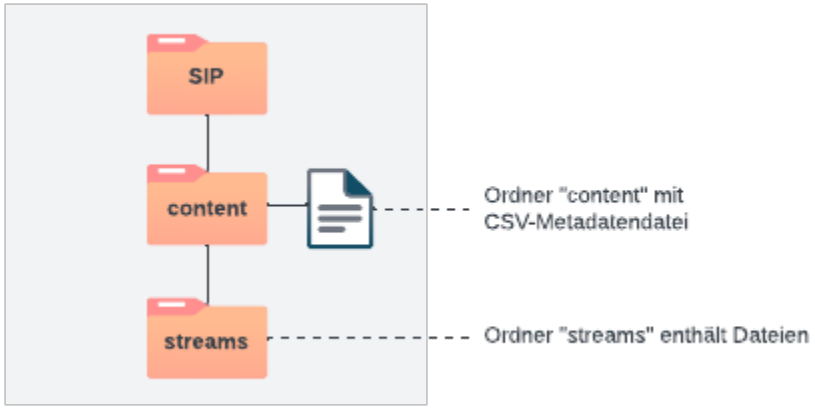
FILE - Type : meta:mapi-message-class: NOTE

PDF/A-Konvertierung

- Nutzung des 3-Heights® Document Converter
- Einstellmöglichkeiten:
  - E-Mails (MSG) mit oder ohne Anhang in PDF
  - E-Mails (MSG) mit oder ohne Header-Angaben in PDF
- Test-Set (Anteil E-Mails mit Anhang: 10%) konvertiert und Einstellmöglichkeiten getestet
- Hohe Erfolgsrate bei Konvertierung:
  - einschliesslich Anhang: 80,7%
  - ohne Berücksichtigung des Anhangs: 99,8%
- Wichtig: Prüfung der Konvertierungsergebnisse: Probleme teils mit überlangen Excel-Tabellen

CSV-Ingest nach Rosetta

Struktur für CSV-Ingest:

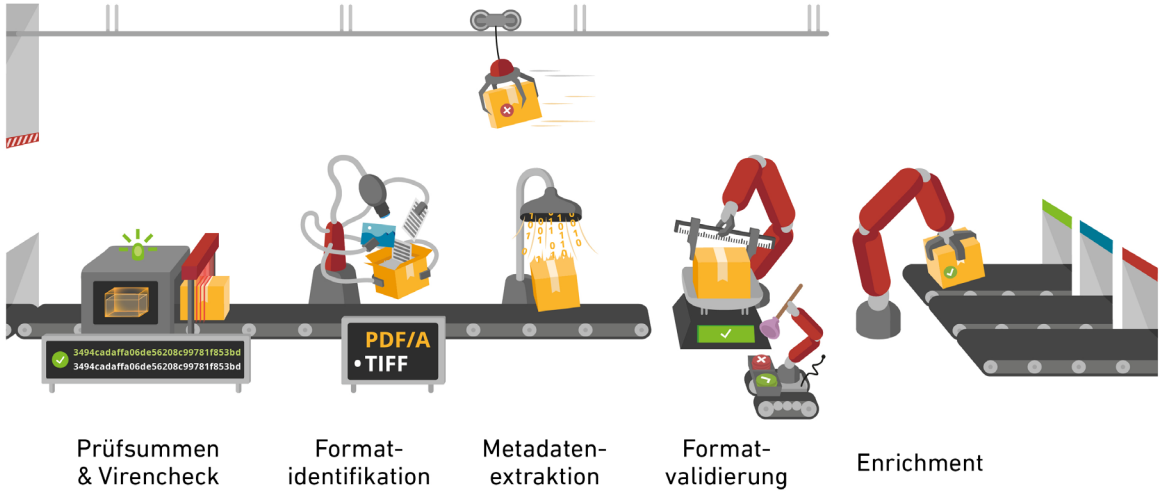


Extraktion der ausgewählten Metadatenfeldern mit dem „CsvPopulator“ (Eigenentwicklung) nutzt Apache Tika für Extraktion und erstellt CSV-Metadatendatei:

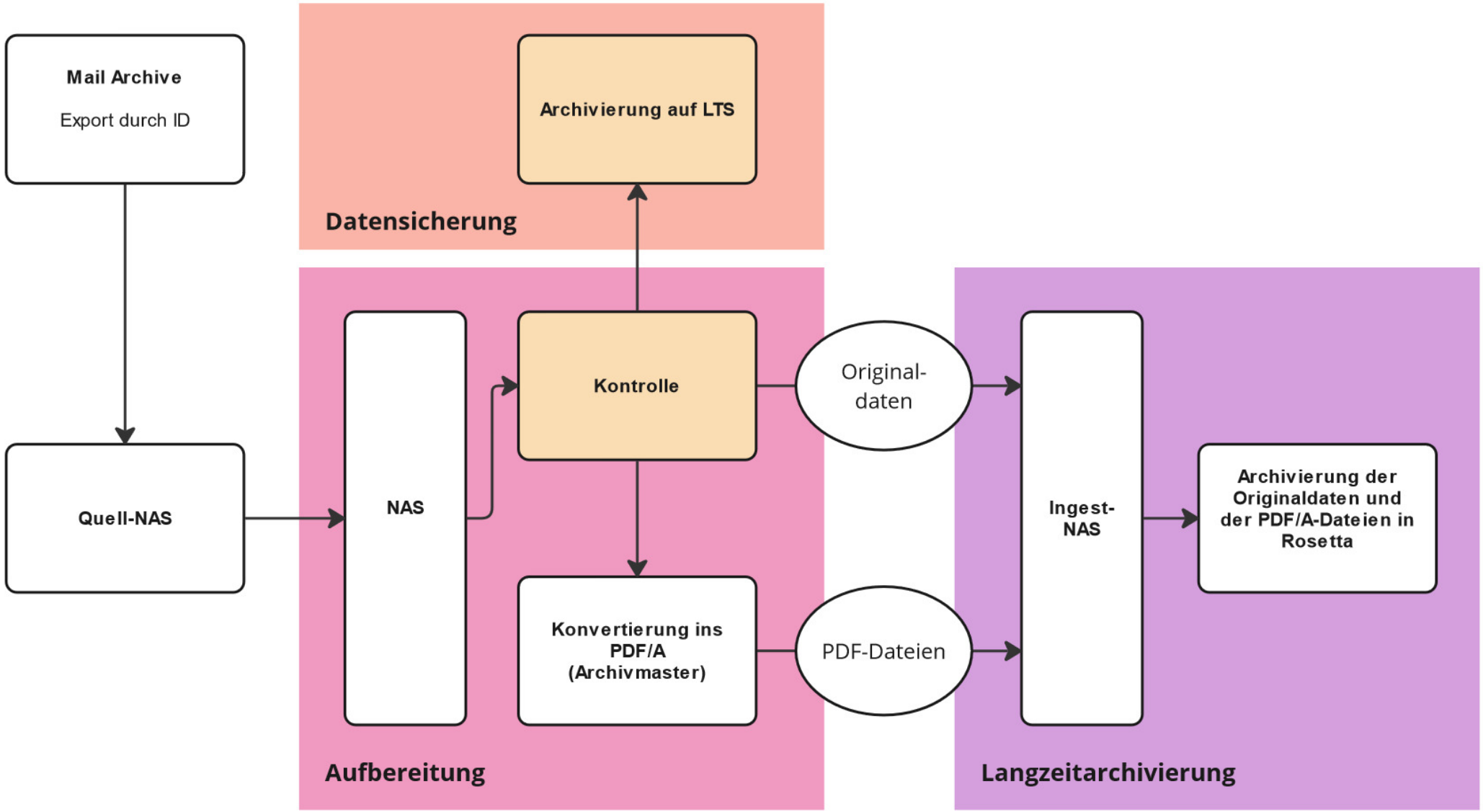
	A	L	M	N	O	P	Q	R	S	T	
1	Object Type	FILE - Creator (DC)	FILE - Creator (DC)	FILE - Contributor (DC)	FILE - Contributor (DC)	FILE - Description (DC)	FILE - Date (DC)	FILE - Type (DC)	File Original Path	File Original Name	MD5
2	SIP										
3	IE										
4	REP										
11	File	Message:From-Nan	Message:From-Email:	/O=ETHZ/OU=ETHZ-MDL/CN=RECIPIENTS/CN=		dc:title: [redacted] MIT	dcterms:created: 2008-01-30T12:29:32Z	meta:mapi-message-class: CONTACT	MSG/	10047.msg	81df07f939(
12	File	Message:From-Nan	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	MSG/	10058.msg	16a5cd40d3
13	File	Message:From-Nan	Message:From-Email:	/O=ETHZ/OU=ETHZ-MDL/CN=RECIPIENTS/CN=		dc:title: Sitzungszimmer F	dcterms:created: 2008-02-06T08:31:40Z	meta:mapi-message-class: CONTACT	MSG/	10097.msg	3c085b46dc

Langzeitarchivierung in Rosetta

- Rosetta als „Digital Asset Management and Preservation System“
- Erhebt und speichert PREMIS konform Metadaten u.a. zu:
  - Fixity (Prüfsummen CRC32, MD5, SHA1, SHA256)
  - Formatidentifikation mittels DROID (Pronom-PUID)
  - Extraktion technischer Metadaten durch JHOVE oder andere Plugins
  - Formatvalidierung mittels JHOVE (nur unterstützte Formate)



Workflow


Mail Archive
Export durch ID
Quell-NAS
Archivierung auf LTS
Datensicherung
NAS
Kontrolle
Konvertierung ins PDF/A (Archivmaster)
Aufbereitung
Original-daten
PDF-Dateien
Ingest-NAS
Archivierung der Originaldaten und der PDF/A-Dateien in Rosetta
Langzeitarchivierung

Stand und nächste Schritte

Und wo stehen wir jetzt?

- Übernahme und Datensicherung abgeschlossen
- Optimierung der PDF/A-Konvertierung durch Einsatz des PDF Compressor
- Ressourcenbedingt: Abkehr von ursprünglichem Plan, MSG und PDF/A-Dateien beide offen zu langzeitarchivieren
- Stattdessen: MSG-Dateien in gezippter Form und PDF/A-Dateien offen
- Herausforderungen:
  - Speicherung im Langzeitarchivsystem: Originaldaten und PDF/A-Dateien zusammen oder separat
  - «Matching»: aus MSG-Dateien extrahierte Metadaten mit entsprechenden PDF/A-Dateien in CSV-Datei zusammenführen

Herausforderungen und Empfehlungen

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

Vielen Dank für die Aufmerksamkeit

Fragen?

Referenzen

- National Archives and Records Administrations (NARA): White Paper on The Capstone Approach and Capstone GRS, 2015. Online: <https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf>.
- National Archives and Records Administrations (NARA): Guidance on a New Approach to Managing Email Records, 2013-02, NARA Bulletin, 2013. Online: <https://www.archives.gov/records-mgmt/bulletins/2013/2013-02.html>.
- PREMIS Data Dictionary for Preservation Metadata, Version 3, 2015: https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf
- 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.

Claudia Briellmann
Fabian Schneider

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

ETH-Bibliothek
Rämistrasse 101
8092 Zürich

www.library.ethz.ch