

EMILiA

Entwicklung einer lizenzbasierten Software
für die Archivierung von E-Mails

Nico Beyer (Freie Universität Berlin)



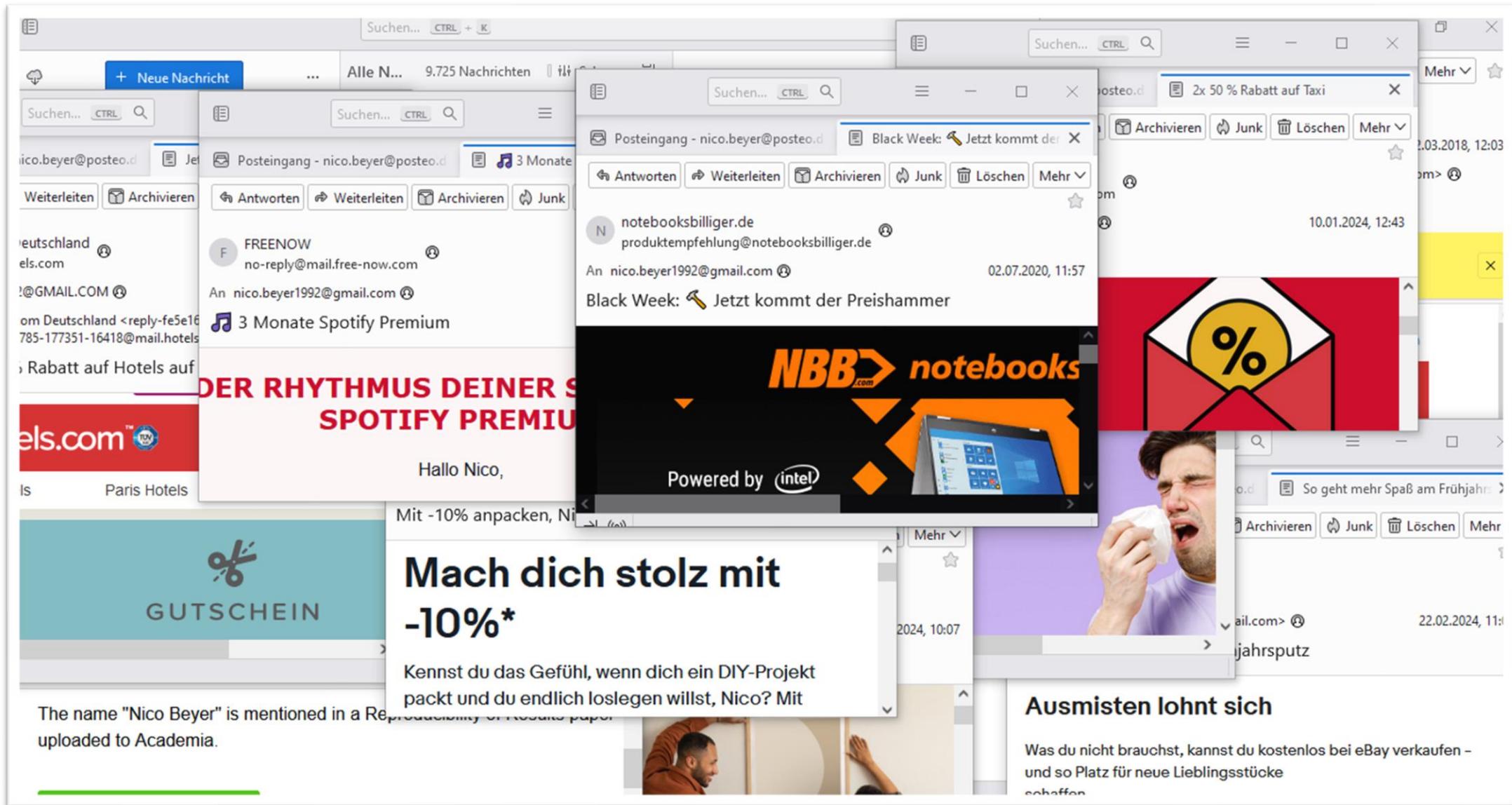
Freie Universität



Berlin



ARCHIV
DER MAX-PLANCK-GESELLSCHAFT



The name "Nico Beyer" is mentioned in a Reproducibility of Results paper uploaded to Academia.

Herausforderungen



Spam, Werbung, E-Mails ohne Informationswert



Händische Bewertung und Erschließung ist aufgrund von riesigen Datenmengen nicht zu bewerkstelligen



Personenbezogene Daten



Viele E-Mails müssen anonymisiert werden



Schadhafte Anhänge



E-Mails können Viren enthalten

Automatisierung als Chance

- Fortschritte im Bereich der künstlichen Intelligenz können dabei helfen, Prozesse zu vereinfachen und zu automatisieren
- Wichtige Entscheidungen werden nach wie vor von Archivfachkräften getroffen
- KI ist in erster Linie als Hilfsmittel für die Bewältigung der riesigen Datenmengen zu verstehen



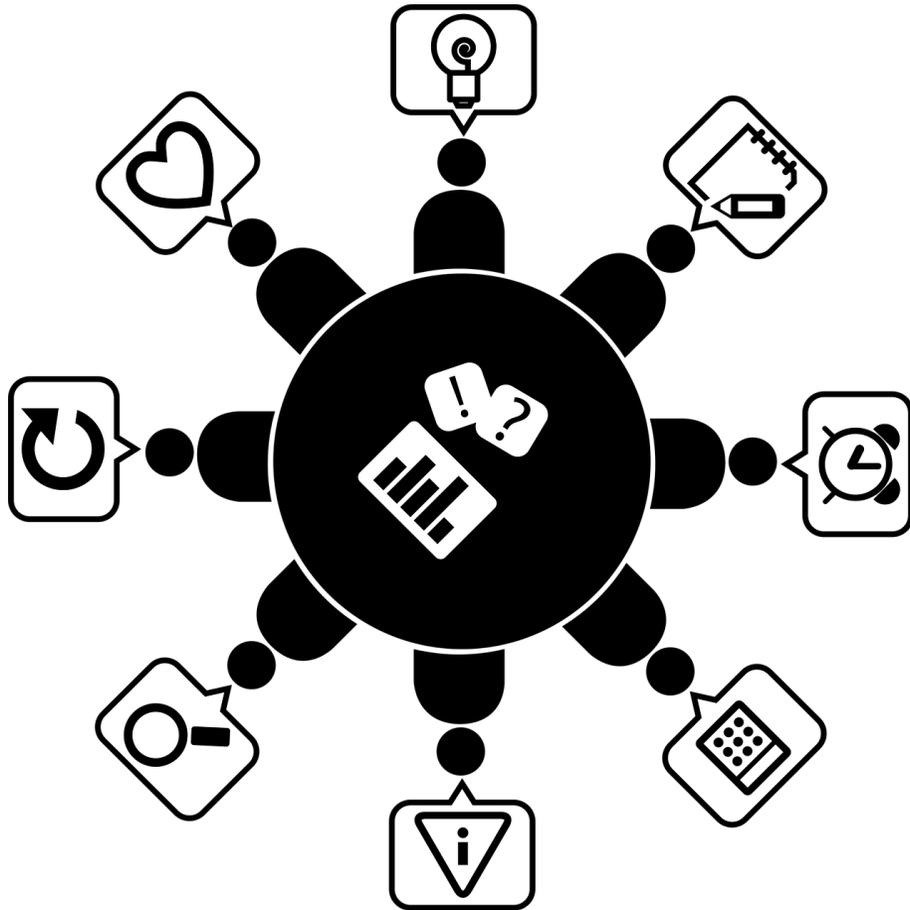
Themen

- 1 Rückblick
- 2 Aktueller Stand
- 3 Herausforderungen
- 4 Vision
- 5 Zeit für Fragen

1

Rückblick

Rückblick



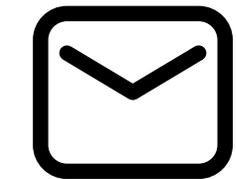
- 2015: Beginn der Konzeption im Archiv der MPG in Kooperation mit dem Fachbereich Informatik der FU Berlin
- Einstellung einer studentischen Hilfskraft im Archiv der Max-Planck-Gesellschaft
- Spezifikation und Entwicklung eines Prototyps
- Tests mit Echtdateien des Archivs der Max-Planck-Gesellschaft
- 2024: Förderprogramm „[ProValid](#)“ der Investitionsbank Berlin

2

Aktueller Stand

Grundkonzept

Erfassung, **M**anagement, **I**ndizierung, **L**imitierung, **i**ntelligente **A**nalyse



Abgabe an das
Archiv



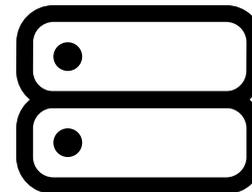
Übernahme

- Import
- Virenprüfung
- Formaterkennung
- Authentizität
- Integrität
- Erstellung von SIPs



Bewertung &
Erschließung

- Erkennung von Spam
und Dubletten
- Erkennung von
Themen, Personen
und Orten
- Identifikation
personenbezogener
Daten



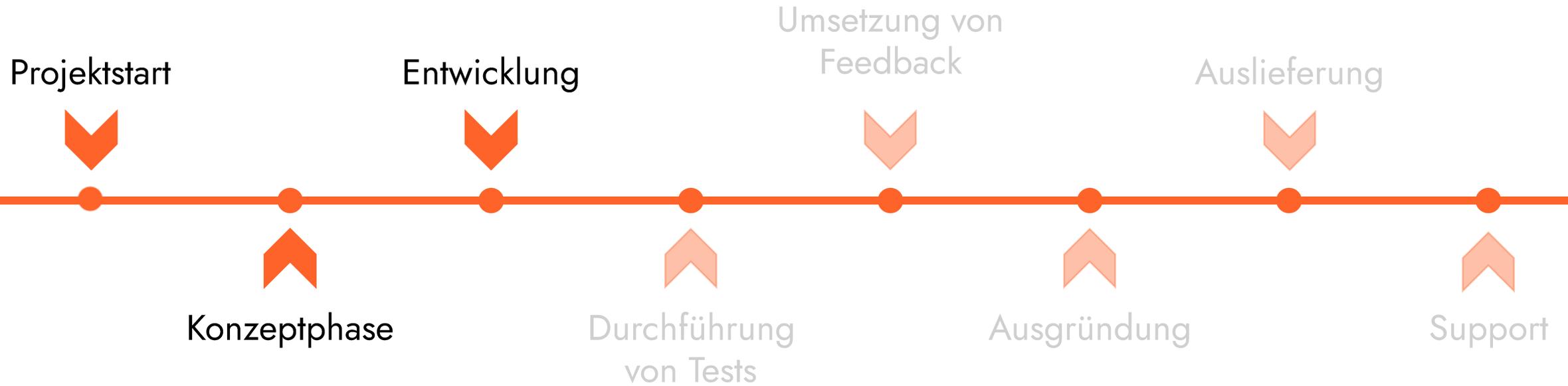
Übergabe an
digitales
Langzeitarchiv



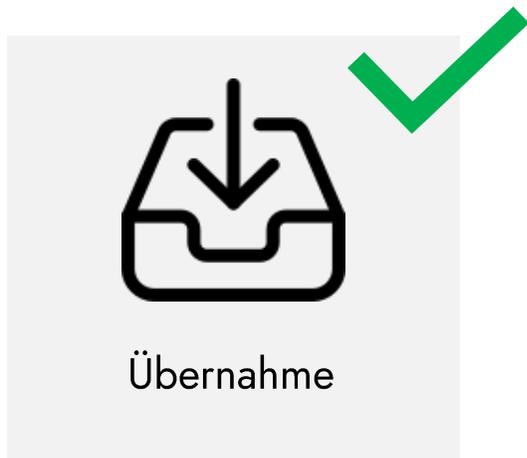
Nutzung

- Recherchedatenbank
- Anonymisierung
- Erstellung von DIPs
- Darstellung

Meilensteine



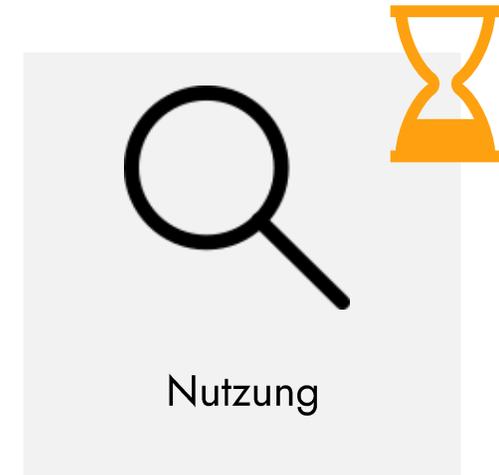
Entwicklungsstand



- Übernahme und Erstellung von SIPs funktioniert bereits



- Erkennung von Klarnamen, Sprachen, Themen und Threads wurde getestet aber noch nicht implementiert
- Dubletten- und Spamererkennung müssen noch evaluiert werden



- Recherchedatenbank und Viewer müssen noch entwickelt werden

Optionen für die Übernahme

1

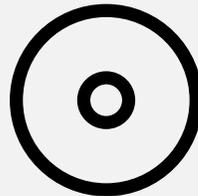


Abgabe über
Mailserver



Deponent*In meldet sich auf
Mailserver an und importiert
den Account

2

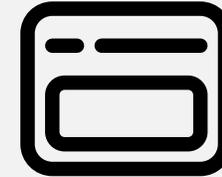


Physische
Abgabe



Deponent*In überträgt E-Mails
mithilfe eines AbgabETOOLS auf
physisches Speichermedium

3



AbgabETOOL



E-Mails werden mithilfe eines
AbgabETOOLS übertragen

Archiv: Deponent*In anlegen

MainWindow

Users User Add X Contributions Contribution Add Authentication

username test

Role CONTRIBUTOR

Name Test Contributor

Contact test@emilia-archiv.de

Institution TEst

batchMode

Add

Meldung

Meldung

User added successful

OK

Name

Rolle

Vor- und Nachname

E-Mail-Adresse

Institution

Stapelverarbeitung

Prototyp

EMILIA 

Archiv: Übernahme vorbereiten

MainWindow

Users User Add Contributions Contribution Add X Authentication

archivist felix

contributor test

startDate 01.03.2024

endDate 06.03.2024

blockingPeriod 10

filtering

prefix TEST

authenticationMode CERTIFICATE

Add

Meldung

Meldung

contribution added successful

OK

Archivar*In

Deponent*In

Startdatum

Enddatum

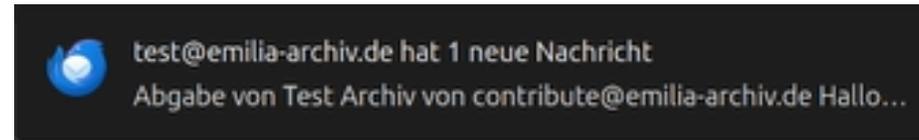
Sperrfrist

Filtern erlauben

Dateiprefix

Authentifizierungsmethode

Deponent*In: Benachrichtigung



Konfigurationsdatei mit Spezifikationen für
Abgabe wird automatisch versendet

Deponent*In: Abgabe

The screenshot displays the ContributionTool interface. The main window, titled 'MainWindow', contains the following sections:

- Allgemeine Informationen zur Abgabe:** Datum der Abgabe: 01.03.2024 - 06.03.2024
- Informationen über ihren Archivar:** Name: Felix G, Kontakt: felix@archiv-berlin.mpg.de
- Informationen über Sie:** Name: Test Contributor, Kontakt: test@emilia-archiv.de
- Fortfahren** (button)

Overlaid on the main window are three other windows:

- Upload:** A dialog box with the text: '...in E-Mail-Konto-Datei im Format MBox oder PST aus oder legen Sie per Drag...er'.
- File (1/1):** A progress dialog showing 'abgabe.mbox' with a progress bar at 100% (3,32 MB / 3,32 MB), a speed of 2,28 MB/sec, and an ETA of 0 s.
- Meldung:** A message dialog with the text: 'Successful contributed' and an 'OK' button.

A text box in the bottom left corner reads: 'Prototyp Willkommen beim ContributionTool. Es wurde keine Konfigurationsdatei gefunden. Bitte wählen Sie eine Konfigurationsdatei aus. Möchten Sie das Programm beenden Klicken Sie hier'.

1. Abgabewerkzeug öffnen
2. Konfigurationsdatei auswählen
3. Daten bestätigen
4. MBOX oder PST auswählen
5. Kommentar abgeben (optional)

Archiv: Übernahme initiieren

MainWindow

id	archivist	contribu...	blocking...	startDate	endDate	submissi...	prefix
31	felix	test	10	2024-03-01	2024-03-06	2024-03-06	TEST

Messages

- archive/event/user/STATUS_CHANGED
- archive/event/notification/SEND_USER
- archive/event/job/CREATED
- archive/event/contribution/STATUS_CHANGE
- archive/event/job/STATUS_CHANGED

Bestätigung

comment

comment

Abbrechen OK

Jobstatus

id = 31, status = FINISHED

Prototyp

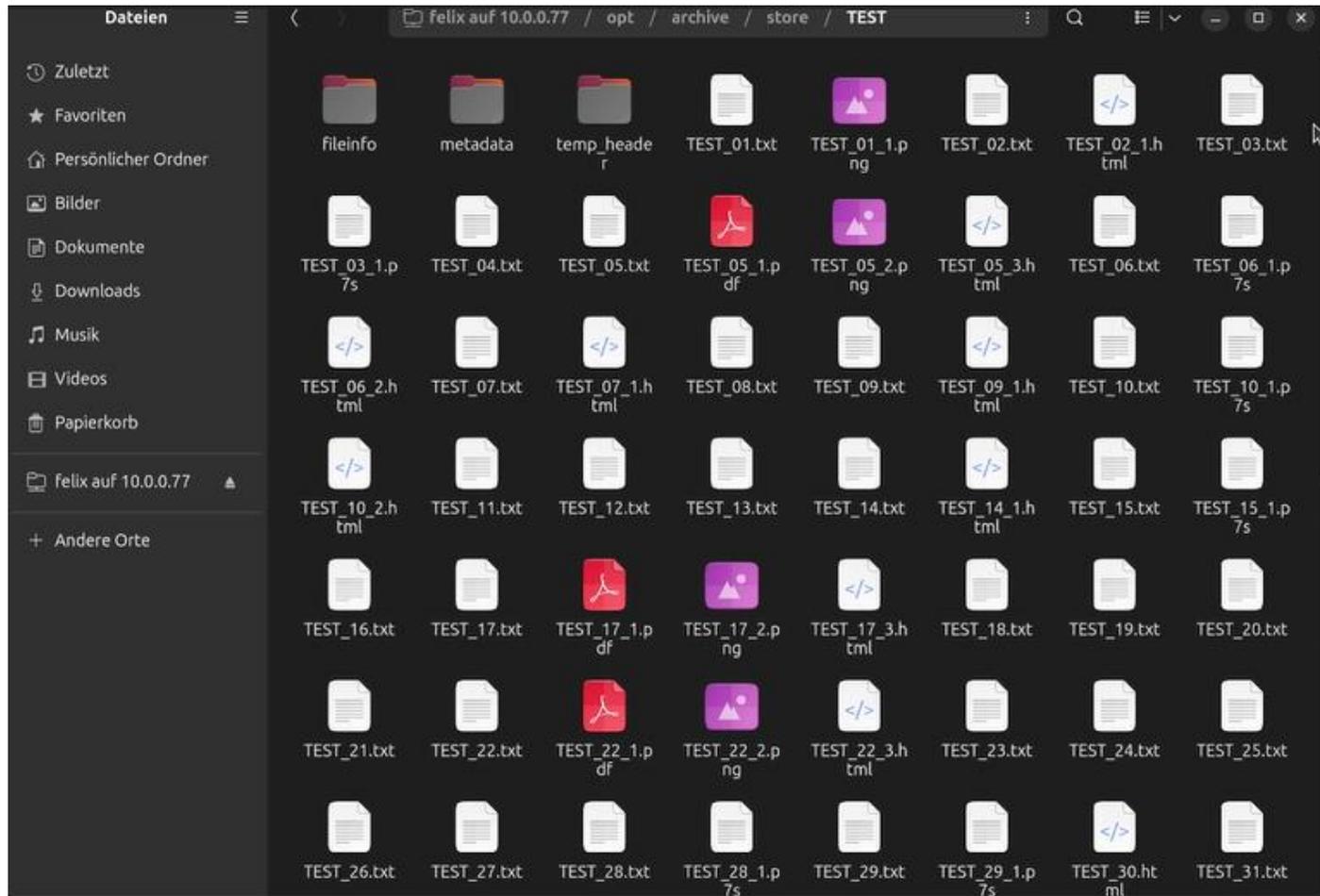
1. Abgabewerkzeug öffnen
2. Kommentar abgeben (optional)
3. Übernahme starten
4. Übertragung abwarten

➔ Dauer der Übernahme hängt von der Leitung der abgebenden Person und des Archivs ab

Übernahme-Workflow



Ergebnis der Übernahme

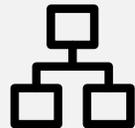


Container mit Strukturdatei, Prüfsummen, PREMIS-Metadaten, E-Mails und Anhängen

Inhalt der SIPs



Angaben zum SIP
(.xml)



Strukturdatei (.xml)



PREMIS-Metadaten
(.xml)



E-Mails (.txt) und
Anhänge (Original)



Prüfsummen
(.txt)

3

Herausforderungen

Herausforderungen

1

E-Mail Standard ist sehr offen, weshalb die eingehenden Metadaten sehr unterschiedlich aussehen können

2

Unterschiedliche E-Mail-Clients, proprietäre Dateiformate und Alter der Daten können zu Defekten und Anomalien führen

3

Umgang mit verschlüsselten E-Mails ist noch unklar

4

Unterschiedliche Zeichenkodierungen können dazu führen, dass einige E-Mails nicht korrekt dargestellt werden

5

Entscheidung für ein archivtaugliches Containerformat steht noch aus

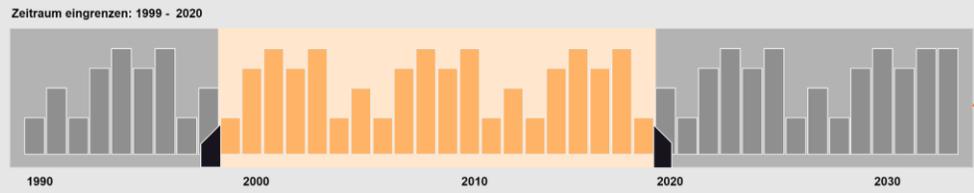
6

Test an Echtdaten zeigt aktuell, dass die Informationsdichte sehr gering ist

Vision

4

Suchbegriffe eingeben 
Erweiterte Suche ▼



- ▲ E-Mail-Konto 1 1166
- ▲ Alle Ordner 1166
- ▲ Posteingang 316
 - ▼ Ordner 1 122
 - ▼ Ordner 2 130
 - ▲ Ordner 3 32
- Ordner 1 29
- Ordner 2 3
- ▼ Entwürfe 0
- ▼ Gesendete Elemente 489
- ▼ Gelöschte Elemente 45

- Ordner 2 Sortieren ▼
- Korrespondenzpartner*In 1
 - Betreff 1
 - Hier könnte ein Ausschnitt des Textes steh...
 - Korrespondenzpartner*In 2
 - Betreff 2
 - Hier könnte ein Ausschnitt des Textes steh...
 - Korrespondenzpartner*In 3
 - Betreff 3
 - Hier könnte ein Ausschnitt des Textes steh...

Betreff 1
Korrespondenzpartner*In 1
10.03.1999 | 17:15
An: test1@emilia-archiv.de
CC: test2@emilia-archiv.de
Lorem [redacted] consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, [redacted] am voluptua. [redacted] sto duo dolores et ea rebum. Stet citta kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Accountbesitzer*In
11.03.1999 | 14:23
An: korrespondenzpartner*In_1@testclient.de
CC: test2@emilia-archiv.de

 Anhang 1 ▼  Anhang 2 ▼

Lorem [redacted] sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet citta kasd gubergren, no sea takimata sanctus est Lorem [redacted]

[In Merkliste speichern](#) [Exportieren](#)

Ordner 2
Anzahl der Mails: 3
Anzahl der Korrespondenzpartner*Innen: 3
Laufzeit: 1999 - 2020
Enthält-Vermerk:
Enthält Korrespondenz zu verschiedenen Angelegenheiten.- Enthält auch Korrespondenz zu anderen Angelegenheiten.
Anmerkungen:
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Betreff 1
Kurzregest:
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet citta kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.
Enthält-Vermerk:
Enthält Korrespondenz zu verschiedenen Angelegenheiten.
Schlagworte:
>Lorem ipsum dolor amet
magna consetetur

Suchschlitz

Histogramm

An E-Mail-Clients angelegter Viewer

Anhänge

Schwärzung

Merkliste und Exportfunktion

Erschließungsdaten und von der KI erkannte Entitäten

5

Zeit für Fragen

Vielen Dank für Ihre Aufmerksamkeit!

Fragen und Vorschläge sind jederzeit willkommen.

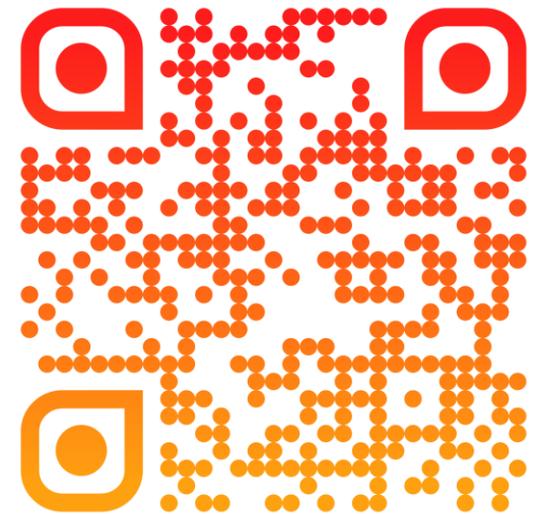
Nico Beyer

Freie Universität Berlin
AG Zuverlässige Systeme
EMILiA-Projekt

E-Mail: nico.beyer@emilia-archiv.de

Telefon: +49 30 841 337 15

Boltzmannstraße 14
14195 Berlin-Dahlem



www.emilia-archiv.de

