EMILiA

Was benötigen die Forschenden?

Entwicklung einer Software für die Archivierung und Nutzbarmachung von E-Mails

Nico Beyer und Felix Gericke

IBB Business Team

Freie Universität Berlin

ARCHIV DER MAX-PLANCK-GESELLSCHAFT

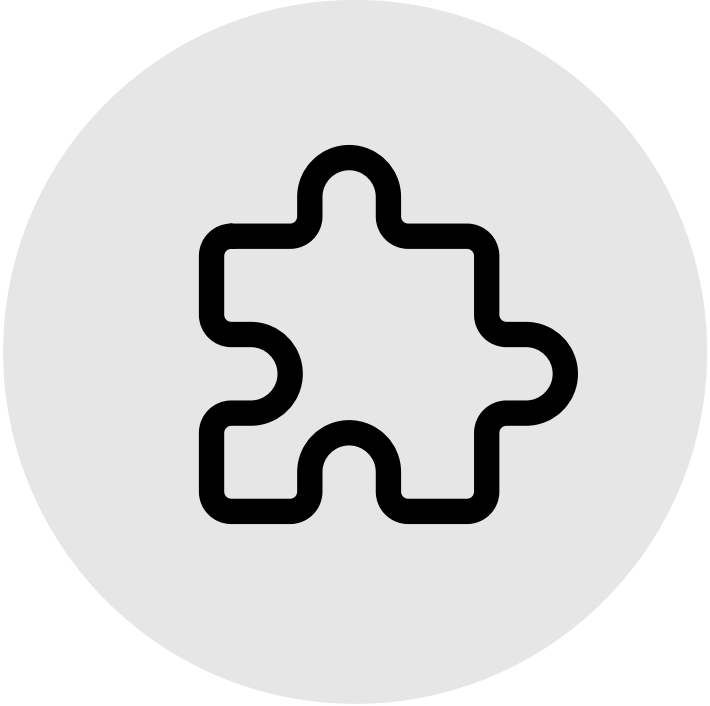
- Wie können Archive historisch relevante E-Mails überliefern und möglichst zeitnah für Forschende nutzbar machen?

EMILiA

Projektrahmen

Projektrahmen

- 2015: Beginn der Konzeption im Archiv der MPG in Kooperation mit dem Fachbereich Informatik der Freien Universität Berlin
- 2017: Beginn der Entwicklung eines Prototyps
- 2024: Förderprogramm „ProValid“ der Investitionsbank Berlin
- Aktuelles Kernteam: 2 Informatiker und 1 Archivar



EMILiA

Was wir heute vorhaben

Workshopziele

- Herausforderungen der E-Mail-Archivierung aufzeigen
- EMILiA-Lösungskonzept kennen
- Eigene Anforderungen formulieren
- Priorisierung weiterer Features
- Möglichkeiten der Partizipation kennen

EMILiA

Agenda

1 Ausgangslage
2 Herausforderungen
  2.1 Technische Aspekte
  2.2 Rechtliche Aspekte
  2.3 Inhaltliche Aspekte
3 Grundkonzept
4 Funktionsumfang
5 Live-Demonstration
6 Diskussion
7 Fazit

EMILiA

Ausgangslage

Aufbau einer E-Mail

Received
Return-Path

SMTP envelope

From: info@emilia-archiv.de
To: audience@offene-archive.de
CC: archiv-berlin@archiv-berlin.mpg.de
Subject: Workshop auf der Konferenz „Offene Archive“

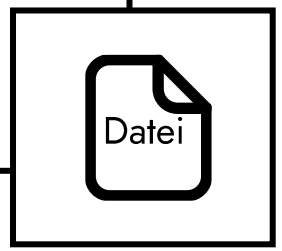
E-Mail-Header

Hallo,

wir hoffen, dass Ihnen unser Workshop gefällt.

Mit freundlichen Grüßen
EMILiA-Projektteam

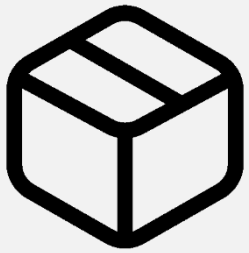
E-Mail-Body



Anhänge

EMILiA

MBOX und PST

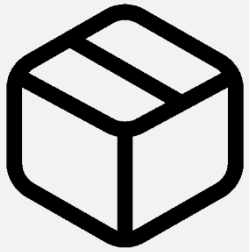


.PST

+

.MSG

- Proprietäres Format der Firma Microsoft
- Kann ohne Hilfsmittel nicht vom Menschen gelesen werden



.MBOX

+

.EML

- Speichert alle E-Mails in einer einzigen Datei
- Erschwert Weiterverarbeitung

EMILiA

Herausforderungen

Technische Aspekte

- Die E-Mail-Formate MBOX und PST gehen bei der Archivierung mit einigen Problemen einher
- Der offene E-Mail-Standard, die Besonderheiten unterschiedlicher E-Mail-Clients und unterschiedliche Zeichenkodierungen erschweren die Verarbeitung der Daten
- Umgang mit signierten und verschlüsselten Mails ist problematisch

Bei der Archivierung von E-Mail-Postfächern müssen zahlreiche technische Barrieren überwunden werden.

EMILiA

Rechtliche Aspekte

- E-Mails enthalten große Mengen personenbezogener Daten
- E-Mails und E-Mail-Anhänge können urheberrechtlich relevant sein
- Außerhalb der Anbietungspflicht können Vorbehalte gegen die Archivierung von E-Mail-Postfächern bestehen

Eine zeitnahe und rechtskonforme Nutzbarmachung von E-Mails ist nur mithilfe einer Anonymisierung denkbar.

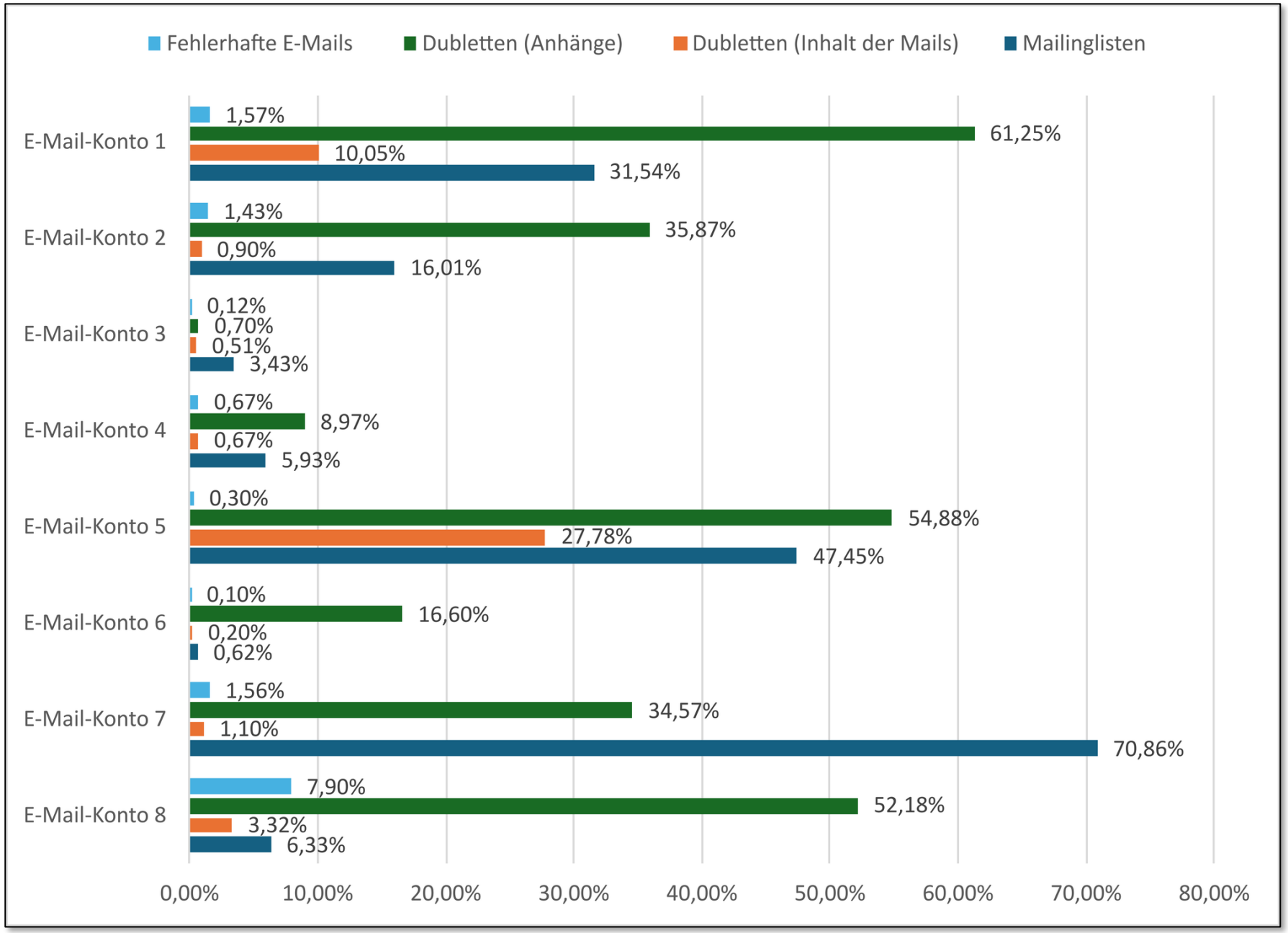
EMILiA

Inhaltliche Aspekte

- Die Postfächer im Archiv der MPG umfassen durchschnittlich **50.000** Nachrichten, die bisher größte Übernahme enthielt **133.194** E-Mails und **258.180** Anhänge
- Spam, Werbemails und Niedrigschwelligkeit elektronischer Kommunikation führen zu riesigen Mengen nicht archivwürdiger Daten
- Inmitten dieses Datenmülls können sich aber auch zahlreiche überlieferungswürdige Nachrichten befinden

Um eine aussagekräftige Überlieferung bilden zu können, ist eine umfangreiche Bewertung und Datenreduktion erforderlich.

EMILiA



Statistische Auswertung von 8 E-Mail-Postfächern, die im Archiv der MPG verwahrt werden

EMILiA

Grundkonzept

Anforderungen

EMILiA muss dazu in der Lage sein, …

… E-Mail-Postfächer inklusive aller Anhänge zu übernehmen und technisch aufzubereiten.

… Archive bei der Bewertung großer Datenmengen zu unterstützen.

… E-Mails rechtskonform und möglichst zeitnah nutzbar zu machen.

… Forschenden sinnvolle Recherche- und Auswertungsmöglichkeiten zur Verfügung zu stellen.

EMILiA

Funktionsumfang

Erfassung, Management, Indizierung, Limitierung, intelligente Analyse

Abgabe von E-Mails an das Archiv

Übernahme

- Import
- Virenprüfung
- Formaterkennung
- Authentizität
- Integrität

Bewertung & Erschließung

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

Übergabe an digitales Langzeitarchiv

Nutzung

- Recherchedatenbank
- Anonymisierung
- Darstellung
- Datenvisualisierung

EMILiA

Automatisierung als Chance

- Fortschritte im Bereich der Automatisierung können dabei helfen, Prozesse zu vereinfachen

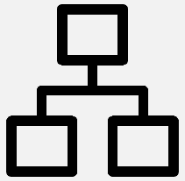
- Wichtige Entscheidungen sollen aber nach wie vor von Archivfachkräften getroffen werden

Übernahme

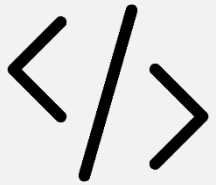
Übernahme-Workflow

Ergebnis der Übernahme

Angaben zur Abgabe (.xml)



Strukturdatei (.xml)



PREMIS-Metadaten (.xml)

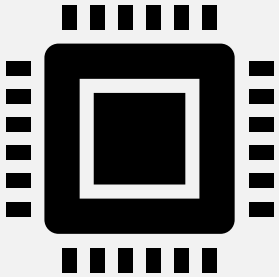
E-Mails (.txt) und Anhänge (Original)

Prüfsummen (.txt)

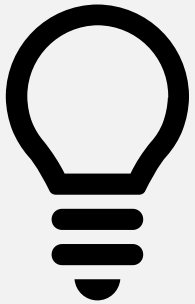
EMILiA

Bewertung

Arten der Bewertung



Teilautomatisierte Bewertung



Inhaltliche Bewertung

Mailinglisten

From: info@emilia-archiv.de
To: audience@webinar.de
Subject: Testsubjekt
X-Mailman-Version: 2.1.29
Precedence: list
List-Id: Mathematik und Informatik <ml-i-studi-mi.lists.fu-berlin.de>
List-Unsubscribe: https://lists.fu-berlin.de/options/ml-i-studi-mi,
ml-i-studi-mi-request@lists.fu-berlin.de?subject=unsubscribe
List-Archive: https://lists.fu-berlin.de/private/ml-i-studi-mi/
List-Post: ml-i-studi-mi@lists.fu-berlin.de
List-Help: ml-i-studi-mi-request@lists.fu-berlin.de?subject=help
List-Subscribe: https://lists.fu-berlin.de/listinfo/ml-i-studi-mi,
ml-i-studi-mi-request@lists.fu-berlin.de?subject=subscribe
Content-Type: text/plain; charset="utf-8"
Content-Transfer-Encoding: base64

- Mails aus Mailinglisten lassen sich über bestimmte Schlüssel-Wert-Paare im E-Mail-Header identifizieren

EMILiA

Spam (Header)

From: info@emilia-archiv.de
To: audience@webinar.de
Subject: Testsubjekt
X-Spam-Flag: YES
X-Spam-Status: Yes, score=12.5 required=5.0 tests=DCC_CHECK,DKIM_SIGNED, DKIM_VALID,DKIM_VALID_AU,DKIM_VALID_EF,FREEMAIL_FROM,FU_BOGO_UNSURE, FU_XPURGATE_SPAM,HTML_MESSAGE,MIME_HTML_ONLY,RCVD_IN_MSPIKE_H2, SPF_HELO_PASS,SPF_PASS,T_REMOTE_IMAGE
X-Spam-Checker-Version: SpamAssassin 3.4.6 on Niue.ZEDAT.FU-Berlin.DE
Content-Type: text/plain; charset="utf-8"
Content-Transfer-Encoding: base64

- Sofern die Abgebenden bereits einen Spamfilter verwendet haben, werden die Befunde im E-Mail-Header hinterlegt.

Spam (Schlüsselbegriffe)

Guten Tag,

Bis heute habe ich nichts von Ihnen bezüglich der E-Mail gehört, die ich Ihnen geschickt habe.
Bitte nehmen Sie zur Kenntnis, dass ich Sie in Bezug auf Ihre Entschädigung nicht mehr kontaktieren werde, und ich empfehle Ihnen, sich mit mir in Verbindung zu setzen, sobald Sie diese E-Mail mit Ihren Angaben erhalten haben.
Ich schreibe Ihnen erneut, um Sie davon in Kenntnis zu setzen, dass Ihre Entschädigungszahlung vom Verwaltungsrat des Überprüfungsausschusses für Entschädigungen und Zulagen der Vereinten Nationen genehmigt wurde.

Ich bitte Sie daher, Ihre Angaben zu bestätigen, damit die Finanzabteilung Ihre Auszahlung unverzüglich genehmigen kann.

Bitte füllen Sie die folgenden Felder aus.
1.Ihren vollständigen Namen
2.Ihre Anschrift
3.Ihre Telefonnummer:

Wir freuen uns, bald von Ihnen zu hören.

Mit freundlichen Grüßen

Judith Parker
Beraterin der Vereinten Nationen für Wiedergutmachung
Vereinigtes Königreich.

Mit der entsprechenden Datengrundlage kann ein Spamfilter trainiert werden, der unerwünschte E-Mails auf Grundlage bestimmter Schlüsselwörter identifiziert.

EMILiA

Dubletten (Inhalt)

Subject: Webinar

Hallo,

wir hoffen, dass Ihnen unser Vortrag gefällt.

Mit freundlichen Grüßen
EMILiA-Projektteam

=

Subject: Webinar

Hallo,

wir hoffen, dass Ihnen unser Vortrag gefällt.

Mit freundlichen Grüßen
EMILiA-Projektteam

E-Mails, deren Inhalt und Betreff identisch sind, werden mithilfe eines Abgleichs der Prüfsummen als Dubletten markiert.

EMILiA

Dubletten (Anhang)

Subject: Webinar

Hallo,

wir hoffen, dass Ihnen unser Vortrag gefällt.

Mit freundlichen Grüßen
EMILiA-Projektteam

≠

Subject: Bewertung von E-Mails

Hallo,

Unserer Meinung nach ist eine aussagekräftige Überlieferung nur mit einer Bewertung möglich.

Mit freundlichen Grüßen
EMILiA-Projektteam

=

- Identische Anhänge werden mithilfe eines Abgleichs der Prüfsummen als Dubletten markiert.

EMILiA

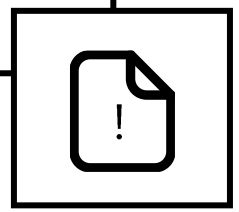
Virenbelastete E-Mails

Subject: Webinar

Hallo,

wir hoffen, dass Ihnen unser Vortrag gefällt.

Mit freundlichen Grüßen
EMILiA-Projektteam

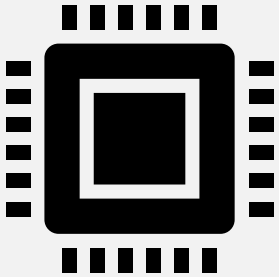


Für die Erkennung potenziell gefährlicher Dateien kommt ein gängiger Virenscanner zum Einsatz.

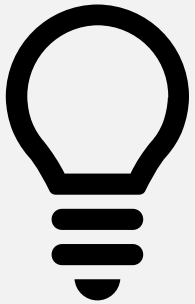
EMILiA

Erschließung

Arten der Erschließung



Teilautomatisierte Erschließung



Intellektuelle Erschließung

Rekonstruktion von Threads

From: info@emilia-archiv.de
To: audience@webinar.de
Subject: Re: Testsubjekt
Message-ID: <20230704123000.1.12345@emilia-archiv.de>
In-Reply-To: <20230757904670.12345@fu-berlin.de>
References: <202307@fu-berlin.de> <reply1.320958@archiv-berlin.mpg.de>
Content-Type: text/plain; charset="utf-8"
Content-Transfer-Encoding: base64

Mithilfe bestimmter Metadaten im E-Mail-Header können Threads wiederhergestellt werden.

Erkennung von Entitäten

Subject: Erkennung von Entitäten

Hallo,

wie funktioniert die Erkennung von Entitäten
bei EMILiA Org ?

Viele Grüße
Nico Beyer Name

- - -
Nico Beyer Name
Fachbereich Informatik Org
Freie Universität Berlin Ort
Takustraße 9, 14195 Berlin Ort

Für die Erkennung von Entitäten kommt das NLP-Framework Flair in der jeweils korrekten Sprachvariante zum Einsatz.

EMILiA

Rekonstruktion von Klarnamen

From: archivar-1992@fu-berlin.de
To: felix.gericke@fu-berlin.de
Subject: Rekonstruktion von Klarnamen

Lieber Herr Felix Gericke,

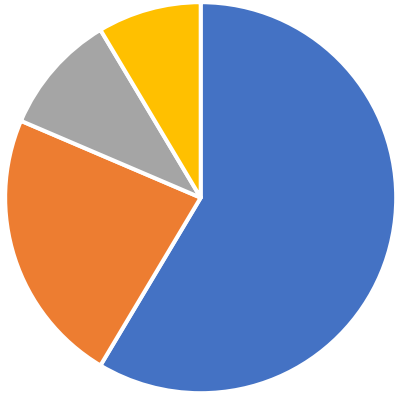
wie funktioniert die Rekonstruktion von Klarnamen bei EMILiA?

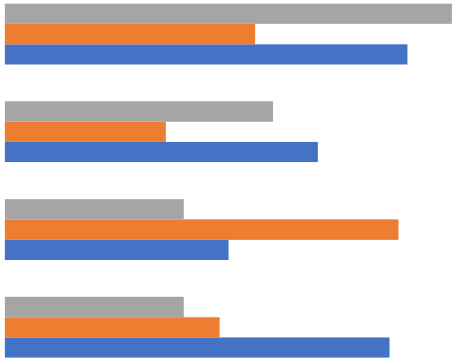
Viele Grüße
Nico Beyer

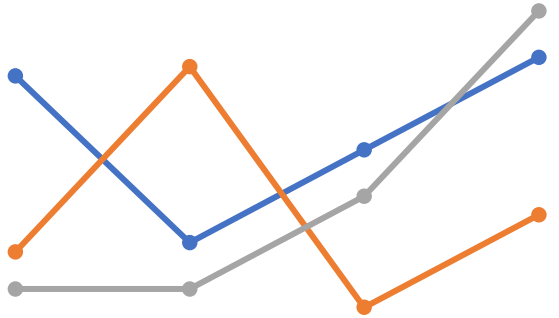
Klarnamen können sowohl mithilfe von Header-Attributen als auch aus textuellen Inhalten, Anreden, Grußformeln und Signaturen extrahiert werden.

EMILiA

Statistische Auswertung



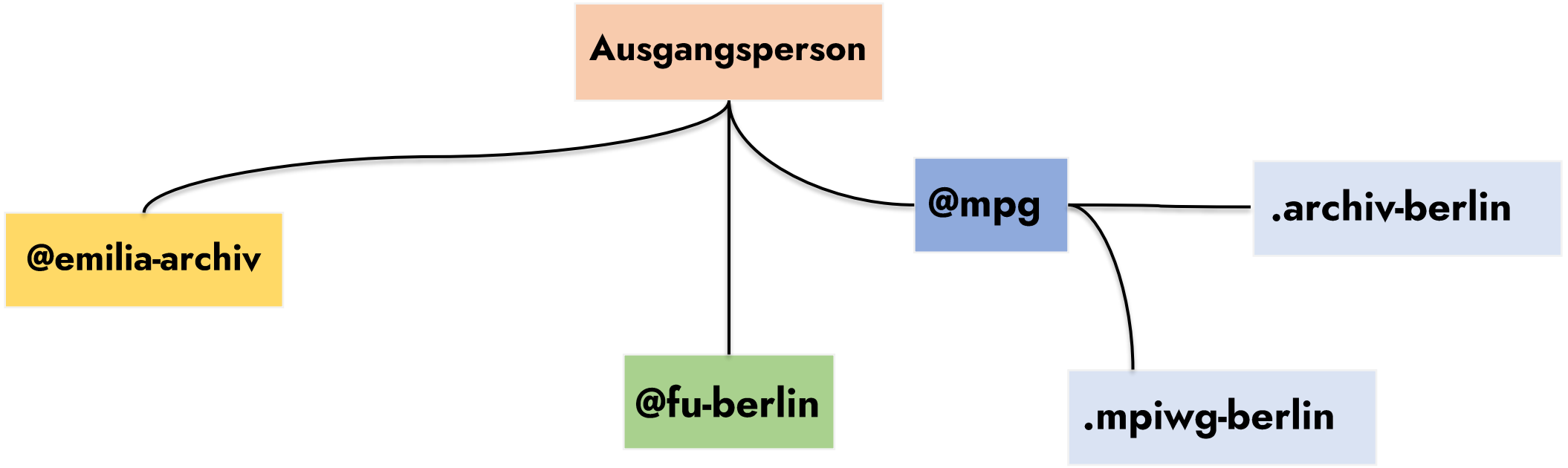




- Im Zuge der technischen Aufbereitung eines E-Mail-Accounts wird eine Vielzahl von statistischen Daten erhoben.

EMILiA

Netzwerkanalyse



Auf Basis der Domains kann das Korrespondenznetzwerk einer Person rekonstruiert werden.

Live-Demonstration

Diskussion

Bleiben Sie auf dem Laufenden

Wir arbeiten aktiv an der Implementierung weiterer Funktionen. Um auf dem Laufenden zu bleiben, können Sie unseren Newsletter abonnieren und unsere Webinare besuchen.

Vielen Dank für Ihre Aufmerksamkeit!

Fragen und Vorschläge sind jederzeit willkommen.

E-Mail: info@emilia-archiv.de
Telefon: +49 30 841 337 15

Archiv der Max-Planck-Gesellschaft
EMILiA-Projekt
Boltzmannstraße 14
14195 Berlin-Dahlem

www.emilia-archiv.de

EMILiA

Umfrage zur Nutzung von E-Mails Quellen

https://emilia-archiv.de/survey/e1645009-308a-4582-bde9-ae38d46a70e0/

Lizenz

- Diese Präsentation kann gemäß der Creative Commons Lizenz CC-BY-SA 4.0 verwendet werden.

- Die in dieser Präsentation enthaltenen Logos und Softwarekonzepte sind von dieser Lizenz ausgenommen und dürfen ohne ausdrückliche Genehmigung des Rechteinhabers nicht weiterverwendet oder verändert werden.

EMILiA