EMILiA

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

Nico Beyer & Felix Gericke

IBB Business Team

Freie Universität Berlin

ARCHIV DER MAX-PLANCK-GESELLSCHAFT

Einleitung

Jeden Tag sollen weltweit durchschnittlich ***~ 360 Mrd. E-Mails*** gesendet und empfangen werden

Riesige Datenmengen

Ein großer Teil dieser Nachrichten enthält ***Spam, Werbung*** oder nur ***kurzfristig relevante Informationen***

Viele wertlose Informationen

Es gibt jedoch ***historisch oder rechtlich relevante E-Mails***, *die* bewahrt werden sollten

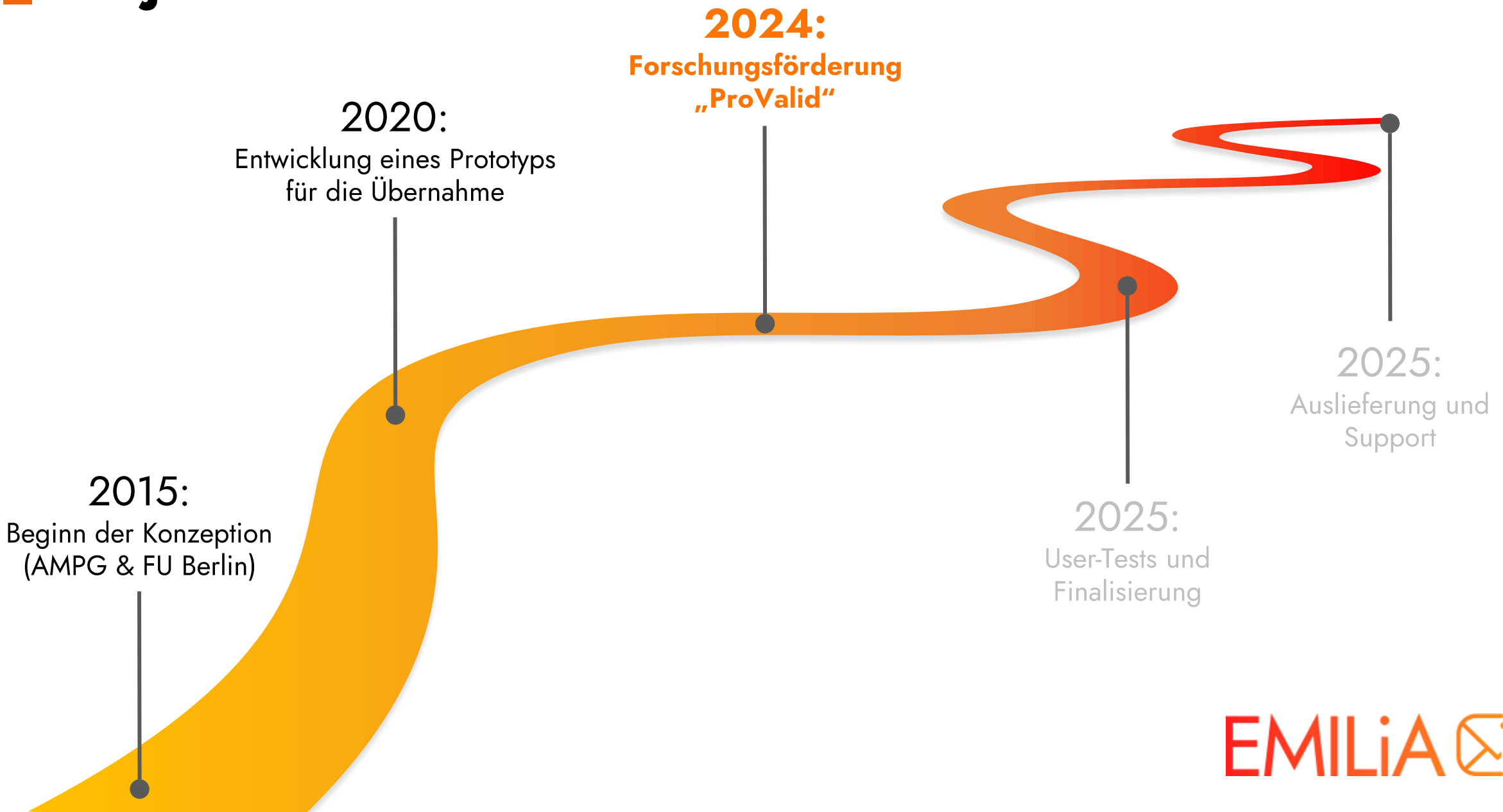
Es muss selektiert werden

Eine ***fachgerechte Auswahl, Archivierung und Auswertung*** ist nur mithilfe automatisierter Prozesse möglich

Automatisierung als Chance

EMILiA

Projektrahmen



Herausforderungen

E-Mail-Standard macht nur wenige klare Vorgaben

Gängige E-Mail-Container sind nicht gut für die Archivierung geeignet

Anhänge in den unterschiedlichsten Formaten

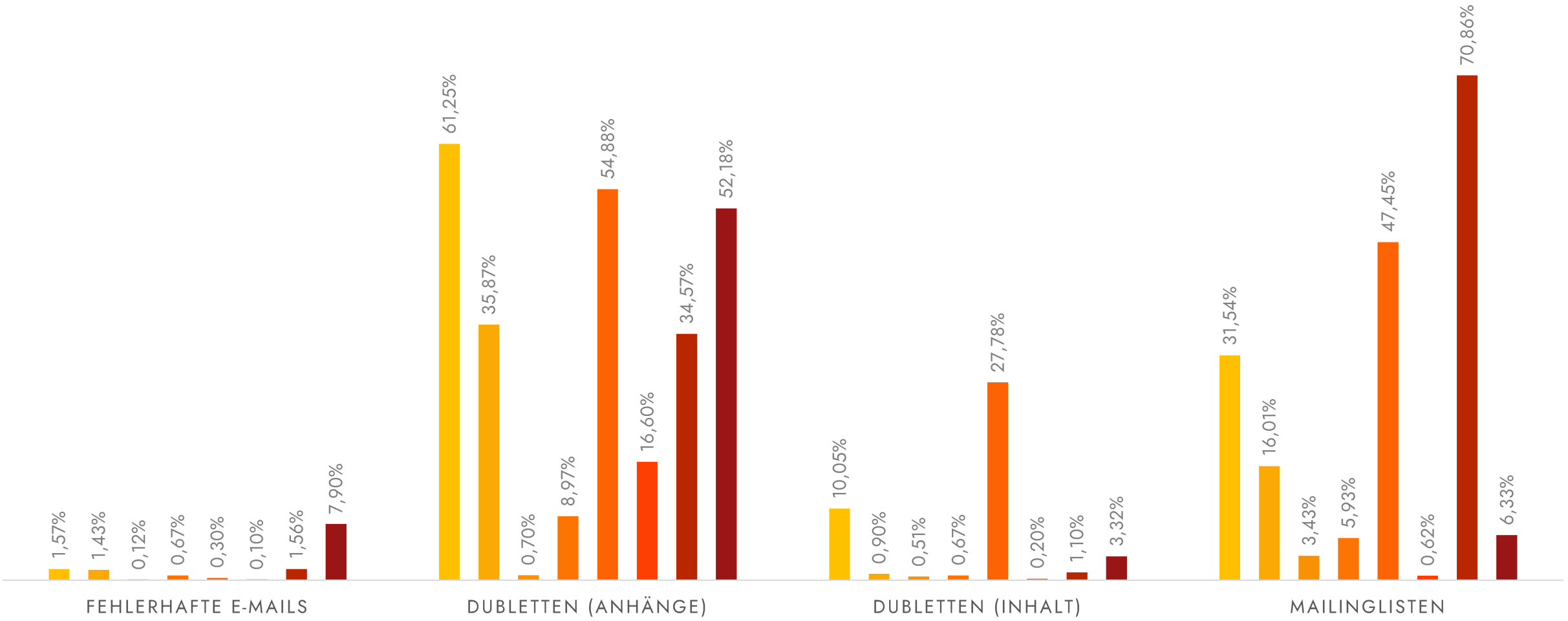
Signierte und verschlüsselte E-Mails

Personenbezogene Daten und urheberrechtlich relevante Dokumente

Archivwürdiger Kern ist nur schwer ausfindig zu machen

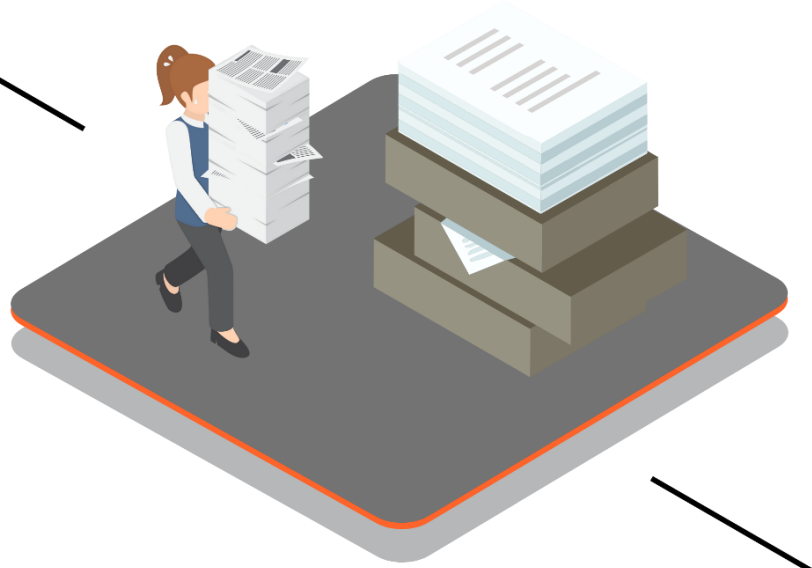
EMILiA

Inhaltliche Zusammensetzung



Die E-Mail-Accounts im Archiv der MPG umfassen im Durchschnitt ~50.000 E-Mails
Die bislang größte Übernahme enthielt 133.194 E-Mails und 258.180 Anhänge

EMILiA



- 600.257 A4-Seiten
- 7 Tage Druckzeit bei 60 S. pro Minute
- ~ 100 laufende Meter

- 5 Bäume
- 25 m ⬍
- 40 cm ⌀

EMILiA

Softwarearchitektur

Erfassung, Management, Indizierung, Limitierung, intelligente Analyse

Abgabe › Übernahme › Vorverarbeitung › Bewertung & Erschließung › Export › Recherche

EMILiA

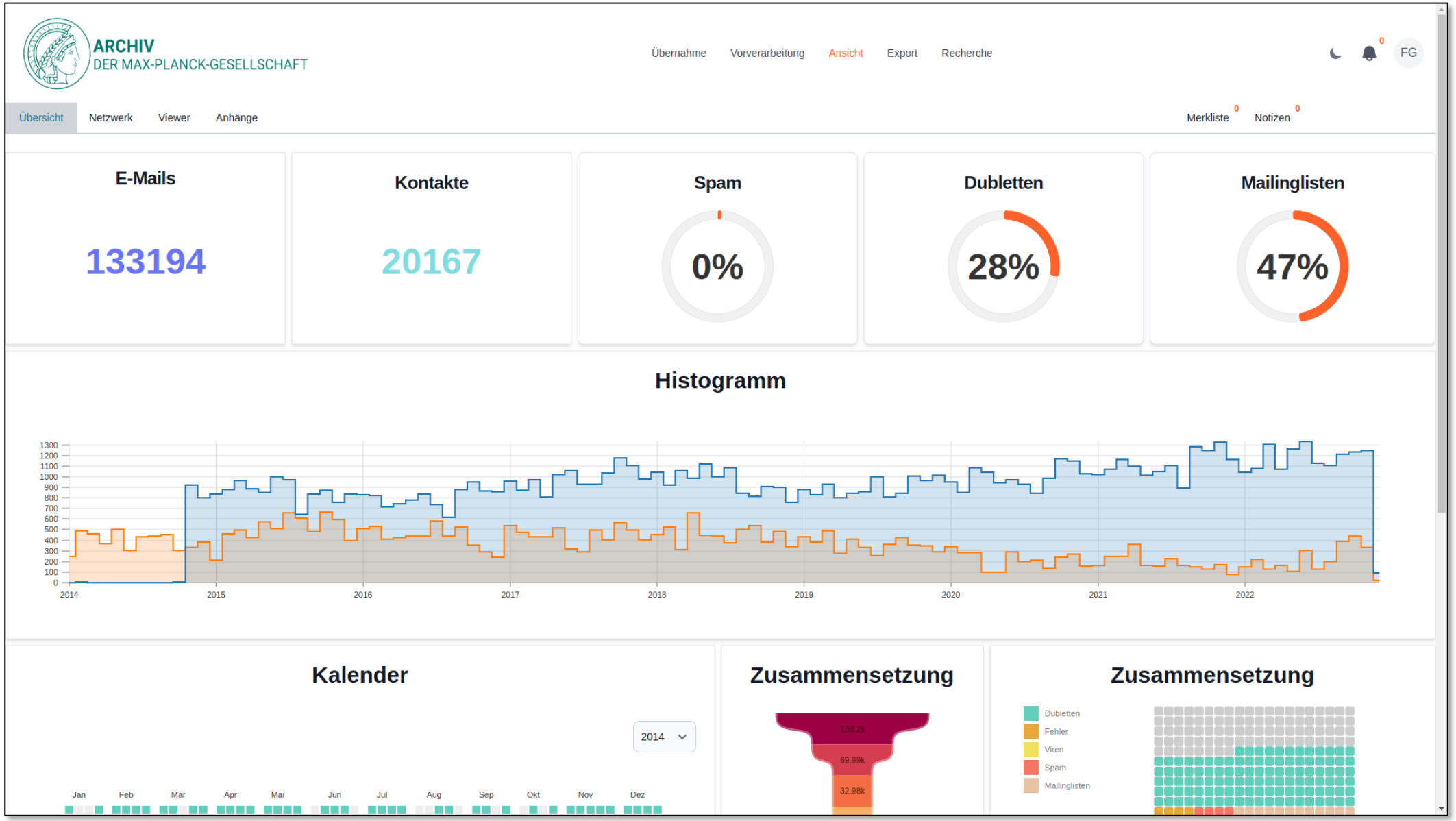
Funktionsumfang

1. **Übernahme:** Sicherer Transfer, Konvertierung, Virenprüfung, Formaterkennung, Authentizitäts- und Integritätssicherung

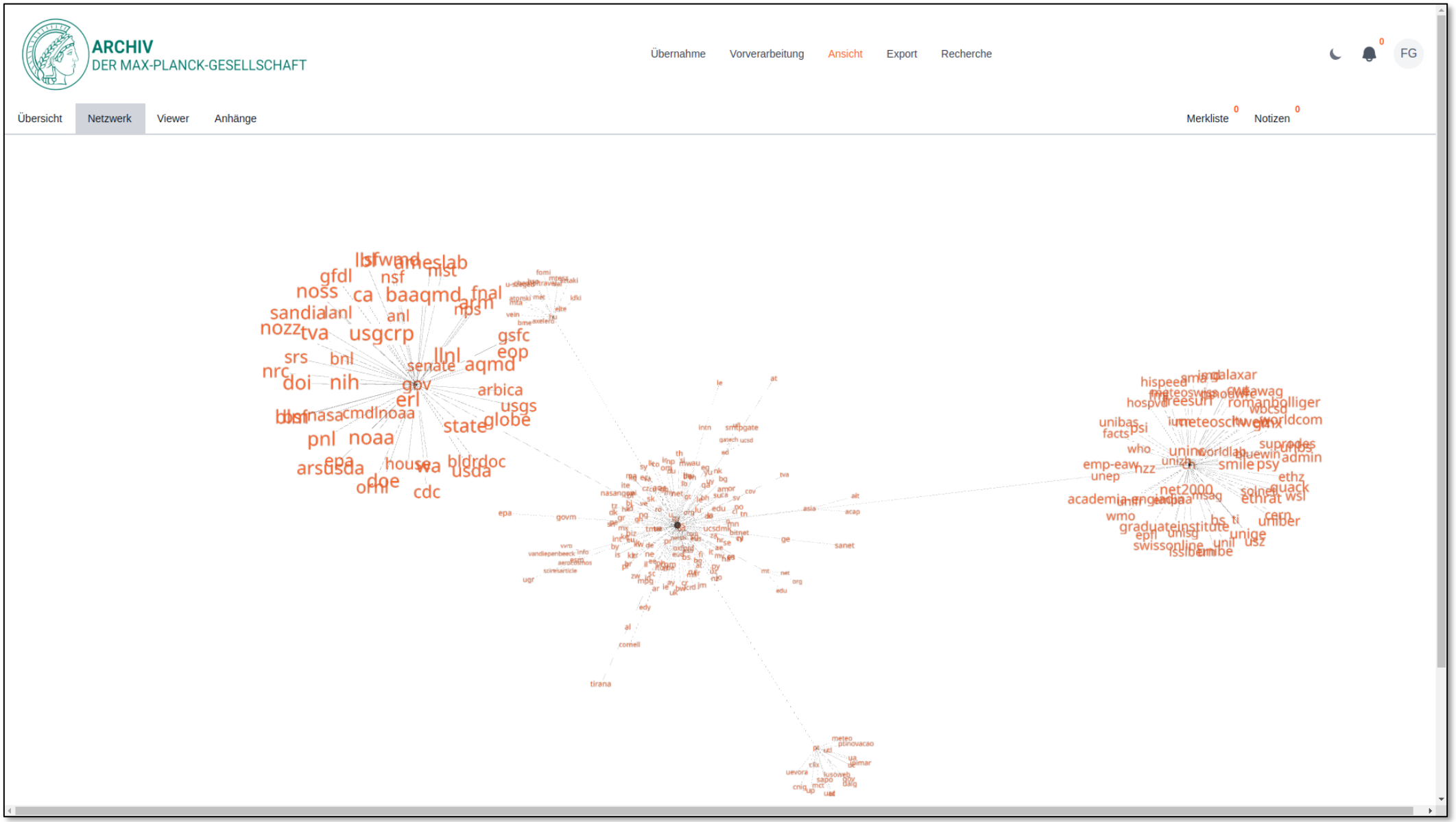
2. **Bewertung & Erschließung:** Spracherkennung, Spam- und Dublettenerkennung, Entitätenerkennung, Themenmodellierung, Identifikation personenbezogener Daten

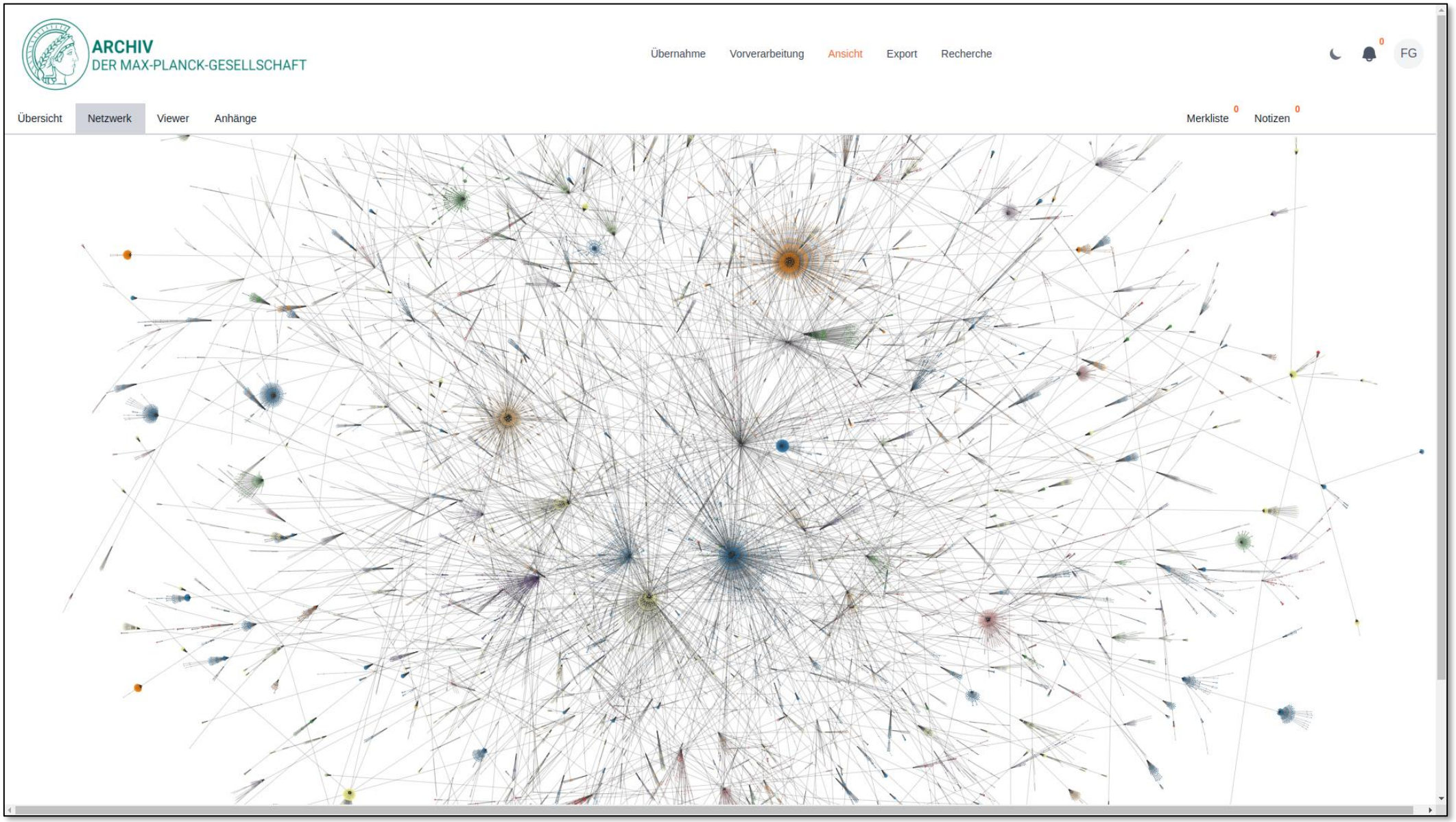
3. **Bereitstellung & Nutzung:** Nutzungsviewer, PDF-Export, Optionale Anonymisierung, Such- und Filterfunktionen, Statistiken und Netzwerke, Zusammenfassung von Threads mittels LLM

EMILiA

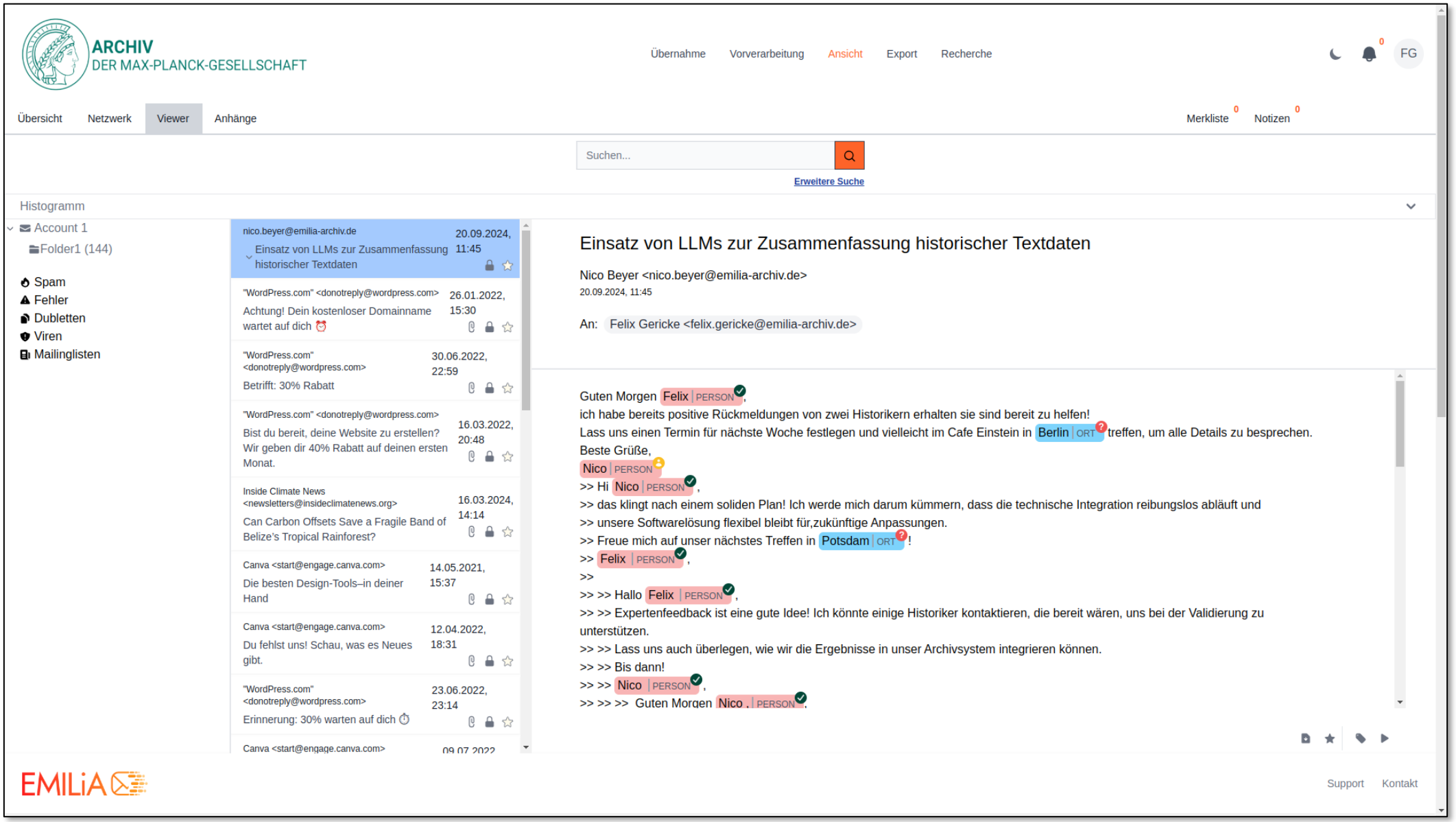

ARCHIV
DER MAX-PLANCK-GESELLSCHAFT
Übernahme
Vorverarbeitung
Ansicht
Export
Recherche
FG
Übersicht
Netzwerk
Viewer
Anhänge
Merkliste 0
Notizen 0
E-Mails
133194
Kontakte
20167
Spam
0%
Dubletten
28%
Mailinglisten
47%
Histogramm
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
2014
2015
2016
2017
2018
2019
2020
2021
2022
Kalender
2014
Jan
Feb
Mär
Apr
Mai
Jun
Jul
Aug
Sep
Okt
Nov
Dez
Zusammensetzung
133.2k
69.99k
32.98k
Zusammensetzung
Dubletten
Fehler
Viren
Spam
Mailinglisten

EMILiA

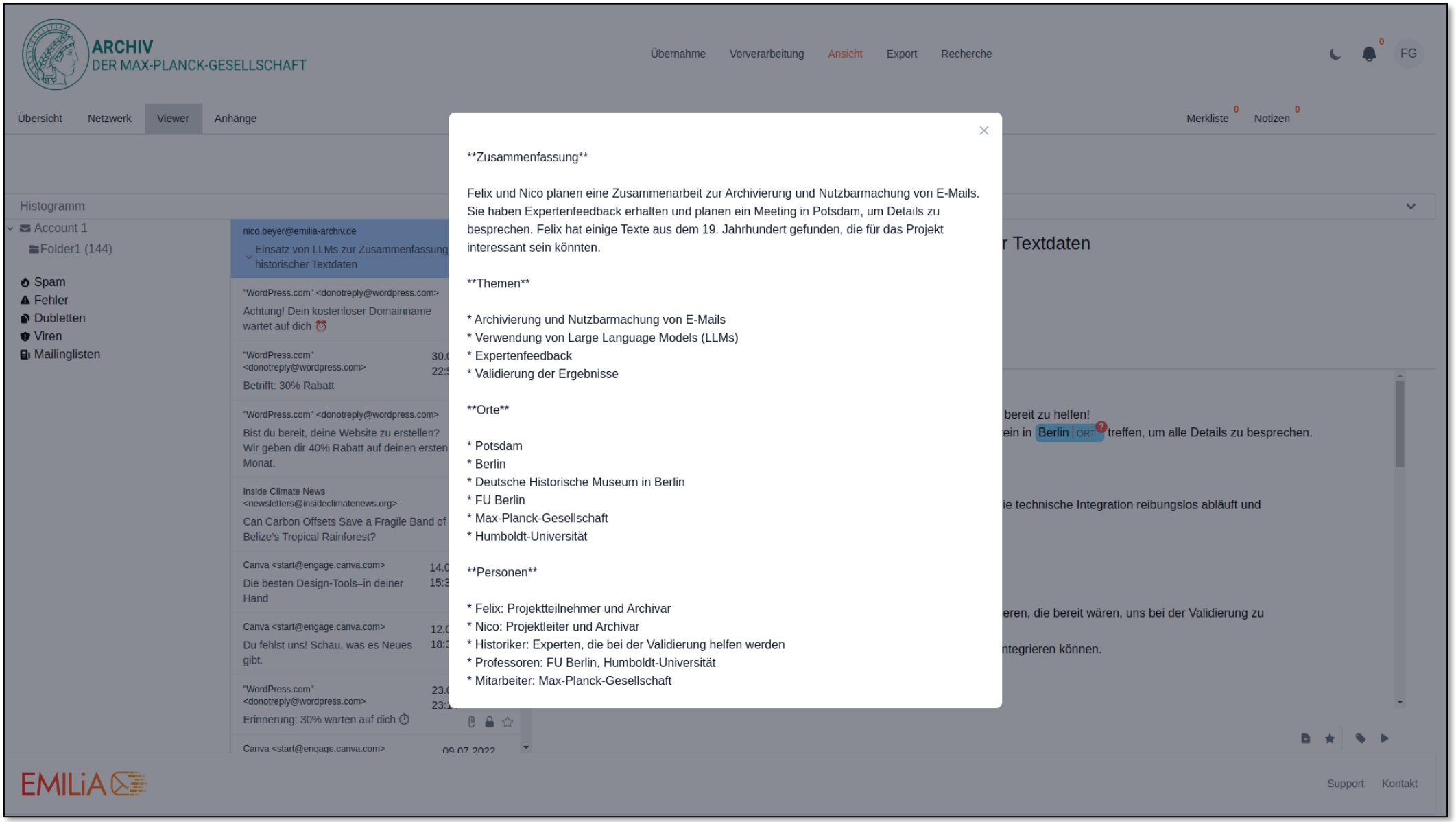

ARCHIV
DER MAX-PLANCK-GESELLSCHAFT
Übernahme
Vorverarbeitung
Ansicht
Export
Recherche
FG
Übersicht
Netzwerk
Viewer
Anhänge
Merkliste
Notizen
gfdl
nsf
nist
noss
ca
baaqmd
fnal
sandia
anl
nps
nozz
tva
usgcrp
gsfc
eop
srs
bnl
llnl
aqmd
nrc
doi
nih
gov
arbica
erl
usgs
nasa
noaa
state
globe
pnl
noaa
epa
arsusda
house
usda
doe
ornl
cdc
bldrdoc
govm
tirana
unibas
facts
bsi
who
unep
emp-eaw
hzz
unizh
smile
psy
ethz
quack
net2000
academia
wmo
graduateinstitute
epfl
unisg
swissonline
unil
usz
unige
unibe
cern
wsl
ethrat
bluewin
admin
romanbolliger
wbcsd
worldcom
suprodes
urbs


ARCHIV
DER MAX-PLANCK-GESELLSCHAFT
Übernahme
Vorverarbeitung
Ansicht
Export
Recherche
FG
Übersicht
Netzwerk
Viewer
Anhänge
Merkliste 0
Notizen 0

EMILiA


ARCHIV
DER MAX-PLANCK-GESELLSCHAFT
Übernahme
Vorverarbeitung
Ansicht
Export
Recherche
FG
Übersicht
Netzwerk
Viewer
Anhänge
Merkliste 0
Notizen 0
Suchen...
Erweitere Suche
Histogramm
Account 1
Folder1 (144)
Spam
Fehler
Dubletten
Viren
Mailinglisten
nico.beyer@emilia-archiv.de
20.09.2024, 11:45
Einsatz von LLMs zur Zusammenfassung historischer Textdaten
"WordPress.com" <donotreply@wordpress.com>
26.01.2022, 15:30
Achtung! Dein kostenloser Domainname wartet auf dich ⏰
"WordPress.com" <donotreply@wordpress.com>
30.06.2022, 22:59
Betrifft: 30% Rabatt
"WordPress.com" <donotreply@wordpress.com>
16.03.2022, 20:48
Bist du bereit, deine Website zu erstellen? Wir geben dir 40% Rabatt auf deinen ersten Monat.
Inside Climate News <newsletters@insideclimatenews.org>
16.03.2024, 14:14
Can Carbon Offsets Save a Fragile Band of Belize's Tropical Rainforest?
Canva <start@engage.canva.com>
14.05.2021, 15:37
Die besten Design-Tools–in deiner Hand
Canva <start@engage.canva.com>
12.04.2022, 18:31
Du fehlst uns! Schau, was es Neues gibt.
"WordPress.com" <donotreply@wordpress.com>
23.06.2022, 23:14
Erinnerung: 30% warten auf dich ⏱
Canva <start@engage.canva.com>
09.07.2022
Einsatz von LLMs zur Zusammenfassung historischer Textdaten
Nico Beyer <nico.beyer@emilia-archiv.de>
20.09.2024, 11:45
An: Felix Gericke <felix.gericke@emilia-archiv.de>
Guten Morgen Felix | PERSON,
ich habe bereits positive Rückmeldungen von zwei Historikern erhalten sie sind bereit zu helfen!
Lass uns einen Termin für nächste Woche festlegen und vielleicht im Cafe Einstein in Berlin | ORT treffen, um alle Details zu besprechen.
Beste Grüße,
Nico | PERSON
>> Hi Nico | PERSON,
>> das klingt nach einem soliden Plan! Ich werde mich darum kümmern, dass die technische Integration reibungslos abläuft und
>> unsere Softwarelösung flexibel bleibt für,zukünftige Anpassungen.
>> Freue mich auf unser nächstes Treffen in Potsdam | ORT!
>> Felix | PERSON,
>>
>> >> Hallo Felix | PERSON,
>> >> Expertenfeedback ist eine gute Idee! Ich könnte einige Historiker kontaktieren, die bereit wären, uns bei der Validierung zu unterstützen.
>> >> Lass uns auch überlegen, wie wir die Ergebnisse in unser Archivsystem integrieren können.
>> >> Bis dann!
>> >> Nico | PERSON,
>> >> >> Guten Morgen Nico , | PERSON
EMILiA
Support
Kontakt

EMILiA


ARCHIV
DER MAX-PLANCK-GESELLSCHAFT
Übernahme
Vorverarbeitung
Ansicht
Export
Recherche
FG
Übersicht
Netzwerk
Viewer
Anhänge
Merkliste
Notizen
Histogramm
Account 1
Folder1 (144)
Spam
Fehler
Dubletten
Viren
Mailinglisten
nico.beyer@emilia-archiv.de
Einsatz von LLMs zur Zusammenfassung historischer Textdaten
"WordPress.com" <donotreply@wordpress.com>
Achtung! Dein kostenloser Domainname wartet auf dich ⏰
"WordPress.com" <donotreply@wordpress.com>
Betrifft: 30% Rabatt
"WordPress.com" <donotreply@wordpress.com>
Bist du bereit, deine Website zu erstellen? Wir geben dir 40% Rabatt auf deinen ersten Monat.
Inside Climate News <newsletters@insideclimatenews.org>
Can Carbon Offsets Save a Fragile Band of Belize's Tropical Rainforest?
Canva <start@engage.canva.com>
Die besten Design-Tools–in deiner Hand
Canva <start@engage.canva.com>
Du fehlst uns! Schau, was es Neues gibt.
"WordPress.com" <donotreply@wordpress.com>
Erinnerung: 30% warten auf dich ⏱
Canva <start@engage.canva.com>
09.07.2022
Zusammenfassung
Felix und Nico planen eine Zusammenarbeit zur Archivierung und Nutzbarmachung von E-Mails. Sie haben Expertenfeedback erhalten und planen ein Meeting in Potsdam, um Details zu besprechen. Felix hat einige Texte aus dem 19. Jahrhundert gefunden, die für das Projekt interessant sein könnten.
Themen
* Archivierung und Nutzbarmachung von E-Mails
* Verwendung von Large Language Models (LLMs)
* Expertenfeedback
* Validierung der Ergebnisse
Orte
* Potsdam
* Berlin
* Deutsche Historische Museum in Berlin
* FU Berlin
* Max-Planck-Gesellschaft
* Humboldt-Universität
Personen
* Felix: Projektteilnehmer und Archivar
* Nico: Projektleiter und Archivar
* Historiker: Experten, die bei der Validierung helfen werden
* Professoren: FU Berlin, Humboldt-Universität
* Mitarbeiter: Max-Planck-Gesellschaft
r Textdaten
bereit zu helfen!
ein in Berlin ORT treffen, um alle Details zu besprechen.
ie technische Integration reibungslos abläuft und
eren, die bereit wären, uns bei der Validierung zu
ntegrieren können.
EMILiA
Support
Kontakt

EMILiA

Vielen Dank für Ihre Aufmerksamkeit!

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