

Neuigkeiten zum EMILiA-Projekt

Teilautomatisierte Bewertung von E-Mail-Postfächern

Nico Beyer, Felix Gericke, Alexander Hinze-Hüttl

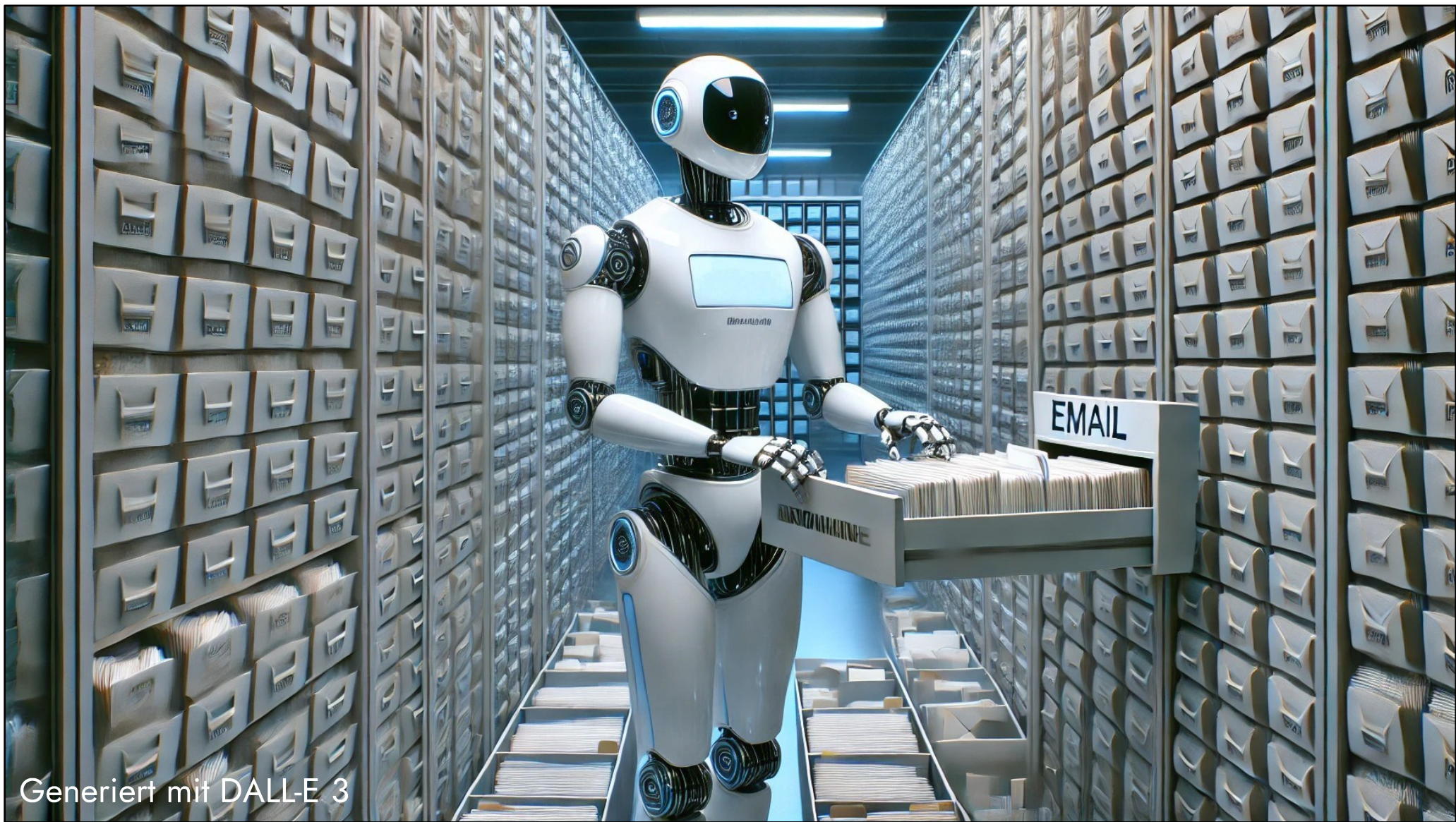


Agenda

- 1 Einleitung
- 2 Projektrahmen
- 3 Erfahrungsbericht
- 4 Lösungsansätze
- 5 Vision
- 6 Diskussionsrunde

Einleitung

1



Generiert mit DALL-E 3

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

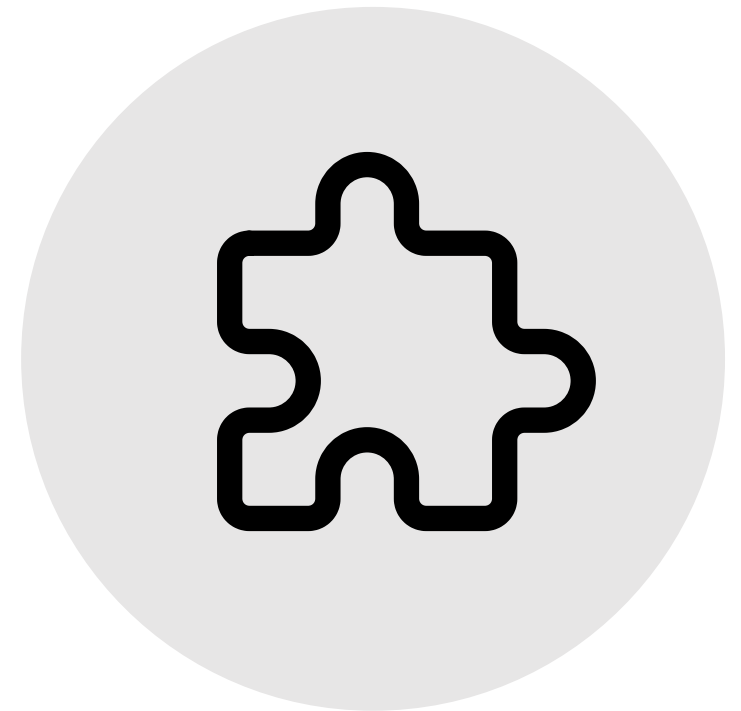


2

Projektrahmen

Werdegang

- 2015: Beginn der Konzeption im Archiv der MPG in Kooperation mit dem Fachbereich Informatik der Freien Universität Berlin
- 2017: Einstellung einer studentischen Hilfskraft und Beginn der Entwicklung eines Prototyps
- 2024: Förderprogramm „[ProValid](#)“ der Investitionsbank Berlin
- Aktuelles Kernteam: 2 Informatiker und 1 Archivar



Grundkonzept

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



Abgabe an das
Archiv



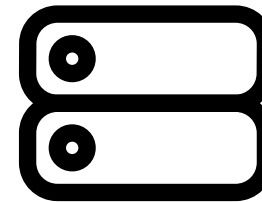
Übernahme

- 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

3

Erfahrungsbericht

Zusammensetzung von E-Mail-Konten

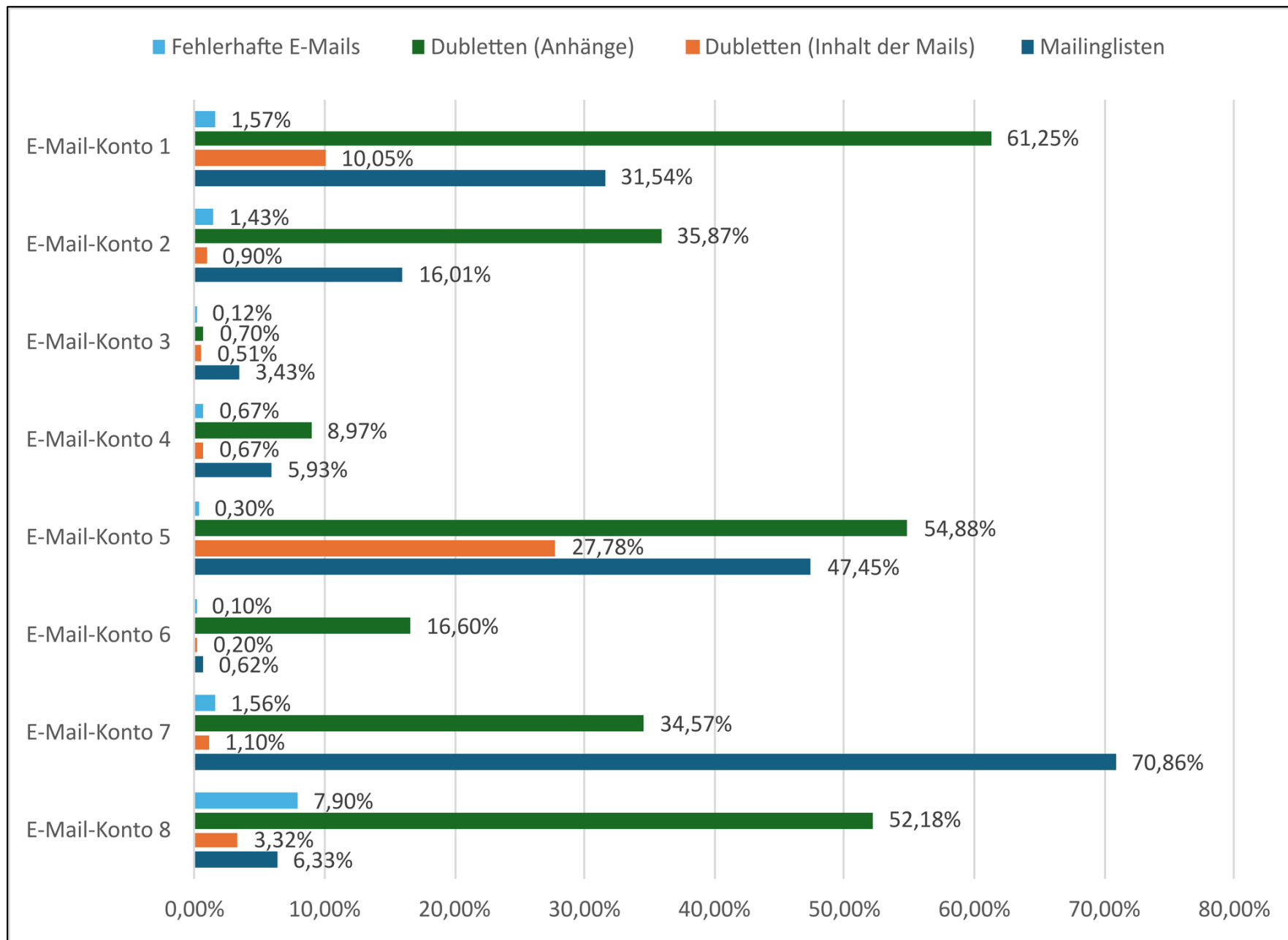
- Es werden ausschließlich E-Mail-Postfächer von Schlüsselpersonen übernommen
- Die Postfächer im Archiv der MPG umfassen durchschnittlich **50.000** Nachrichten, die bisher größte Übernahme enthielt **133.194** E-Mails und **258.180** Anhänge
- Spam, Werbemails und der teilweise inflationäre Umgang mit der elektronischen Kommunikation führen zu riesigen Mengen nicht archivwürdiger Daten
- Inmitten dieses Datenmülls können sich aber auch zahlreiche überlieferungswürdige Nachrichten befinden



Vorsortierung durch die Abgebenden

- Im Vorfeld der ersten Übernahmen wurde von den Abgebenden die Option gefordert, Postfächer innerhalb des Abgabetools noch einmal sortieren zu können
- Bei den tatsächlichen Übernahmen wurden die Filteroptionen dann jedoch von keiner einzigen Person genutzt
- Unsere bisherigen Erfahrungen zeigen, dass nur in sehr wenigen Fällen eine Vorsortierung durch die Abgebenden stattfindet



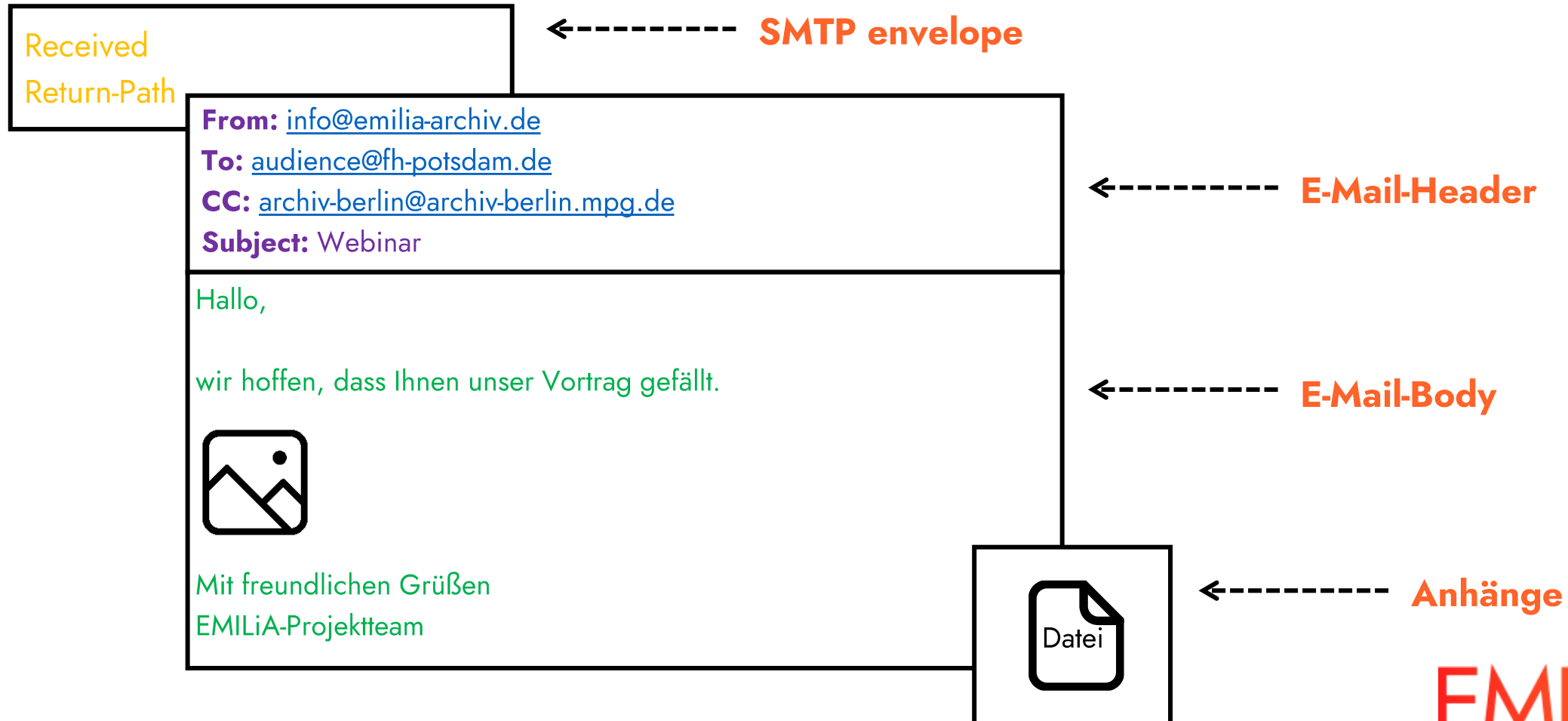


Statistische Auswertung von 8 E-Mail-Postfächern, die im Archiv der MPG verwahrt werden

4

Lösungsansätze

Aufbau einer E-Mail



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

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 (Maschinelles Lernen)

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 Prüfungsausschusses 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.

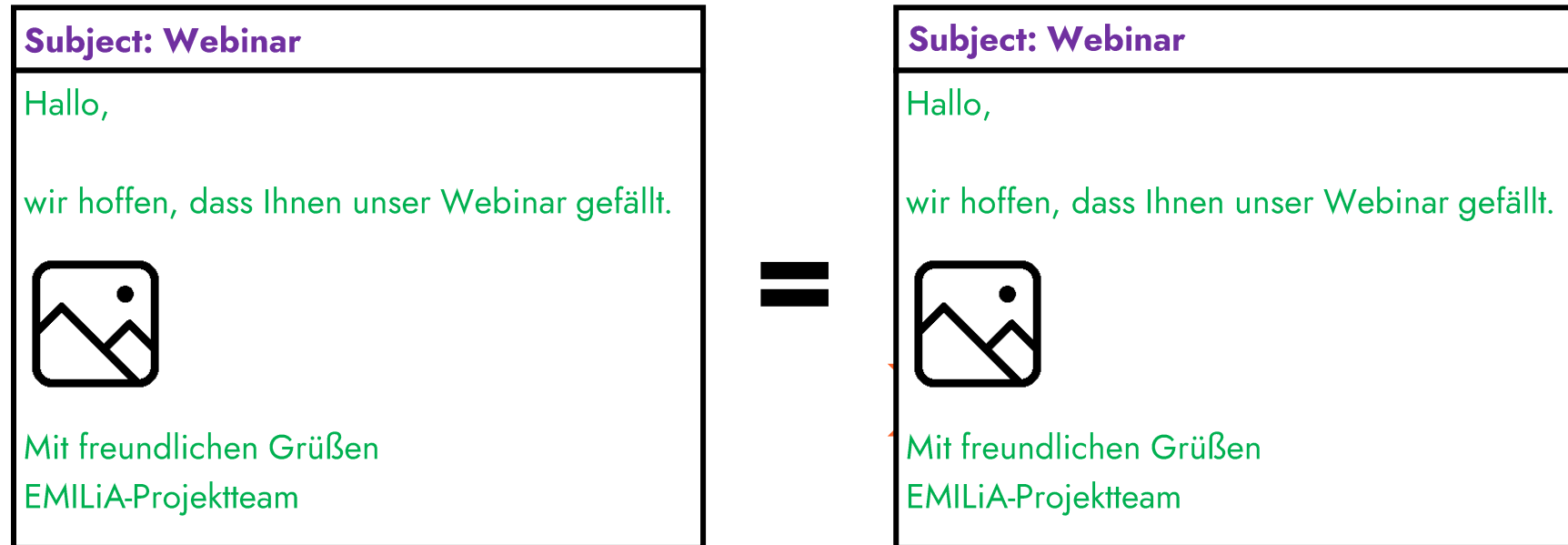
Mit freundlichen Grüßen

Judith Parker
Beraterin der Vereinten Nationen für Wiedergutmachung
Vereinigtes Königreich.



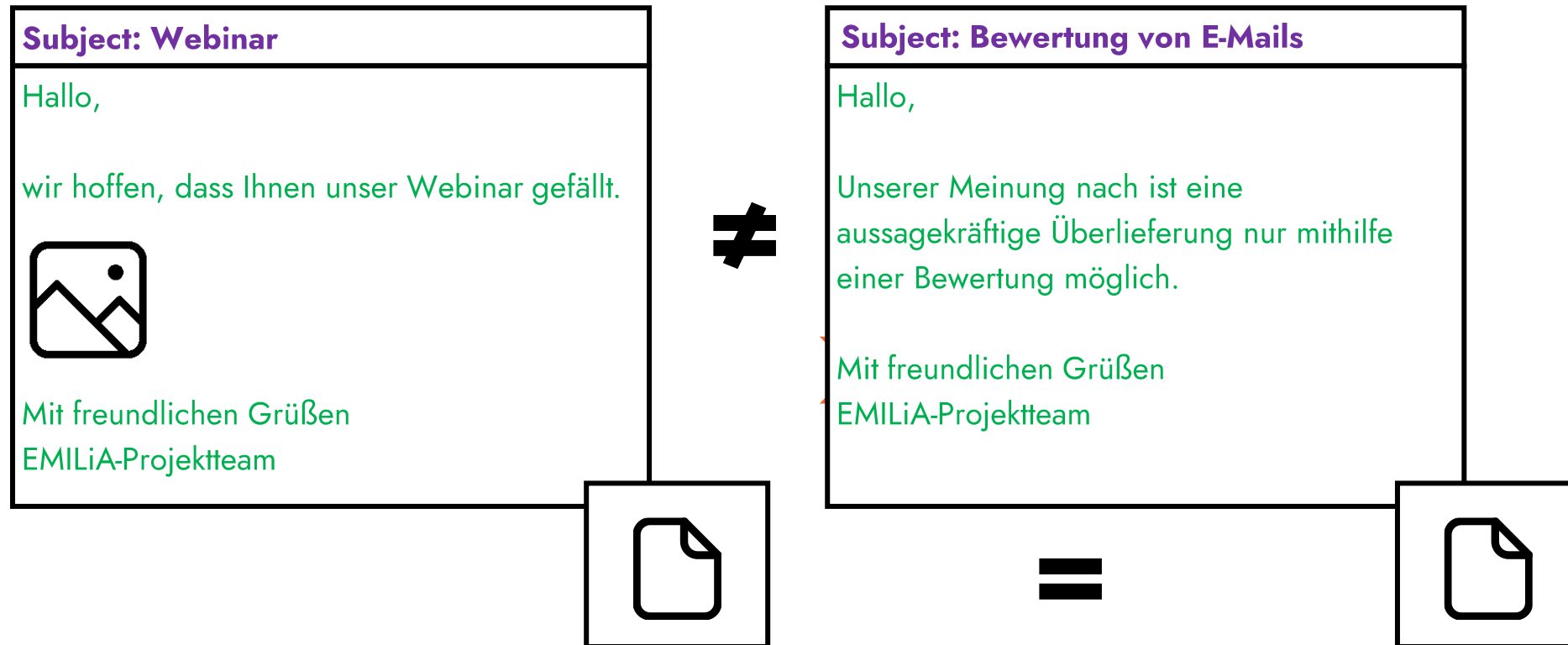
Mit der entsprechenden Datengrundlage kann ein Spamfilter trainiert werden, der unerwünschte E-Mails auf Grundlage bestimmter Schlüsselwörter identifiziert

Dubletten (Inhalt)



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

Dubletten (Anhang)



➤ Identische Anhänge werden mithilfe eines Abgleichs der Prüfsummen als Dubletten markiert

Virenbelastete E-Mails

Subject: Webinar

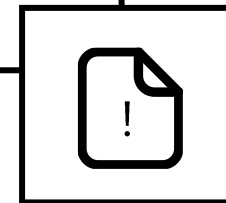
Hallo,

wir hoffen, dass Ihnen unser Webinar gefällt.



Mit freundlichen Grüßen

EMILiA-Projektteam



Für die Erkennung potenziell gefährlicher Dateien kommt ein gängiger und bewährter Virens Scanner zum Einsatz

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 Name

Freie Universität Berlin Ort

Takustraße 9, 14195 Berlin Ort

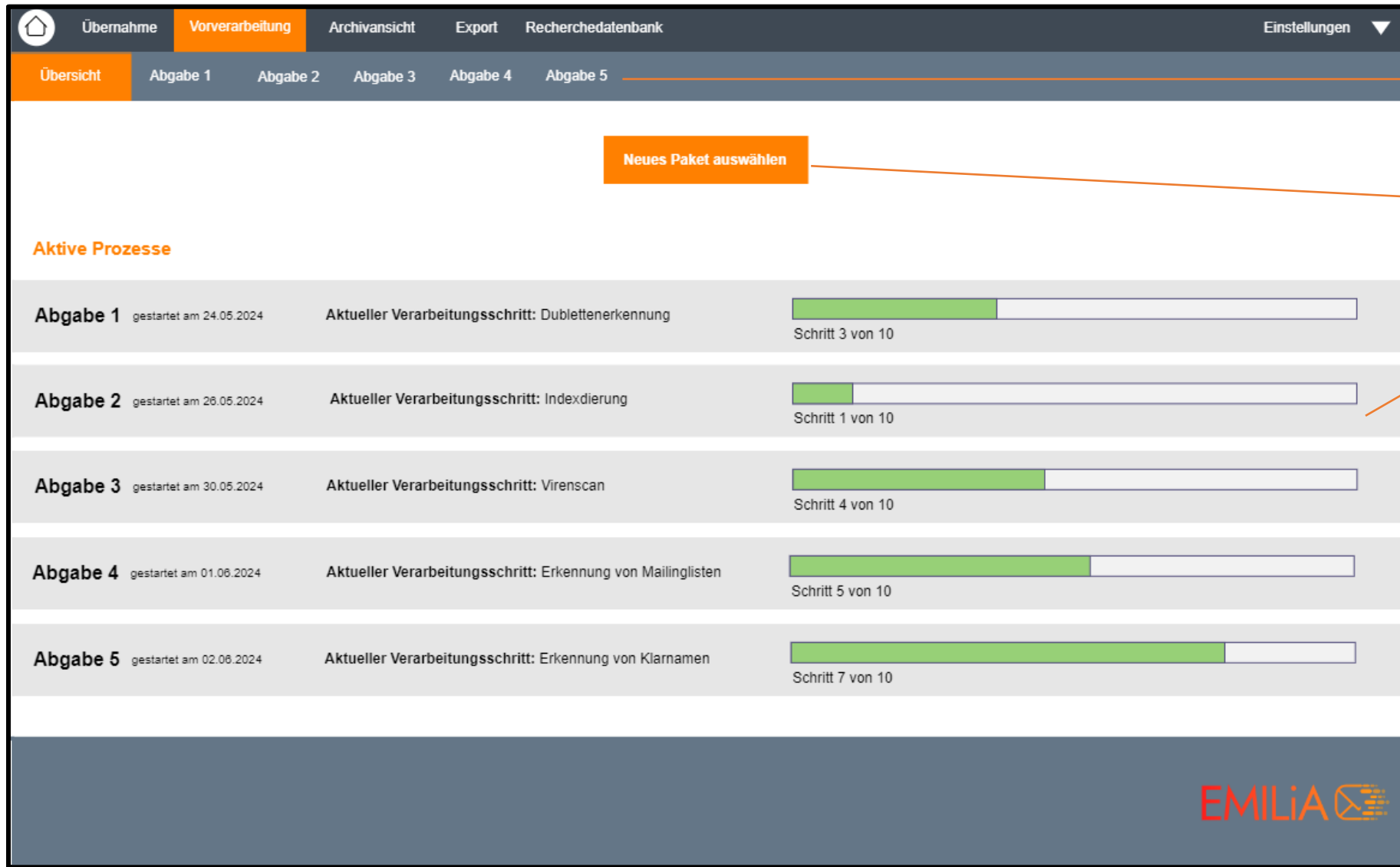


Für die Erkennung von Entitäten kommt das NLP-Framework Flair in der jeweils korrekten Sprachvariante zum Einsatz

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

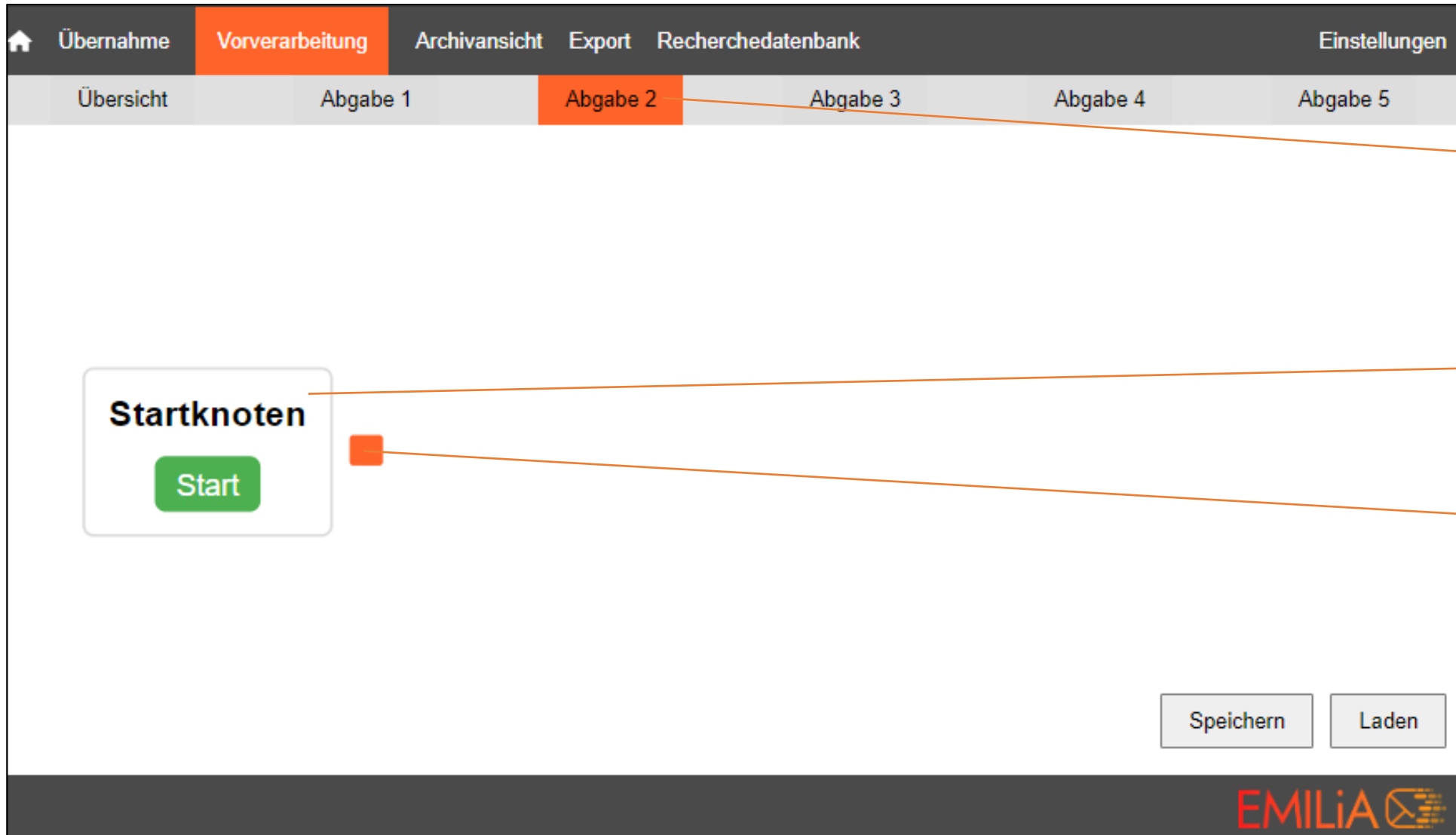


Laufende Prozesse

Paketauswahl

Übersicht über laufende Prozesse

Vorverarbeitungsübersicht (Konzept)



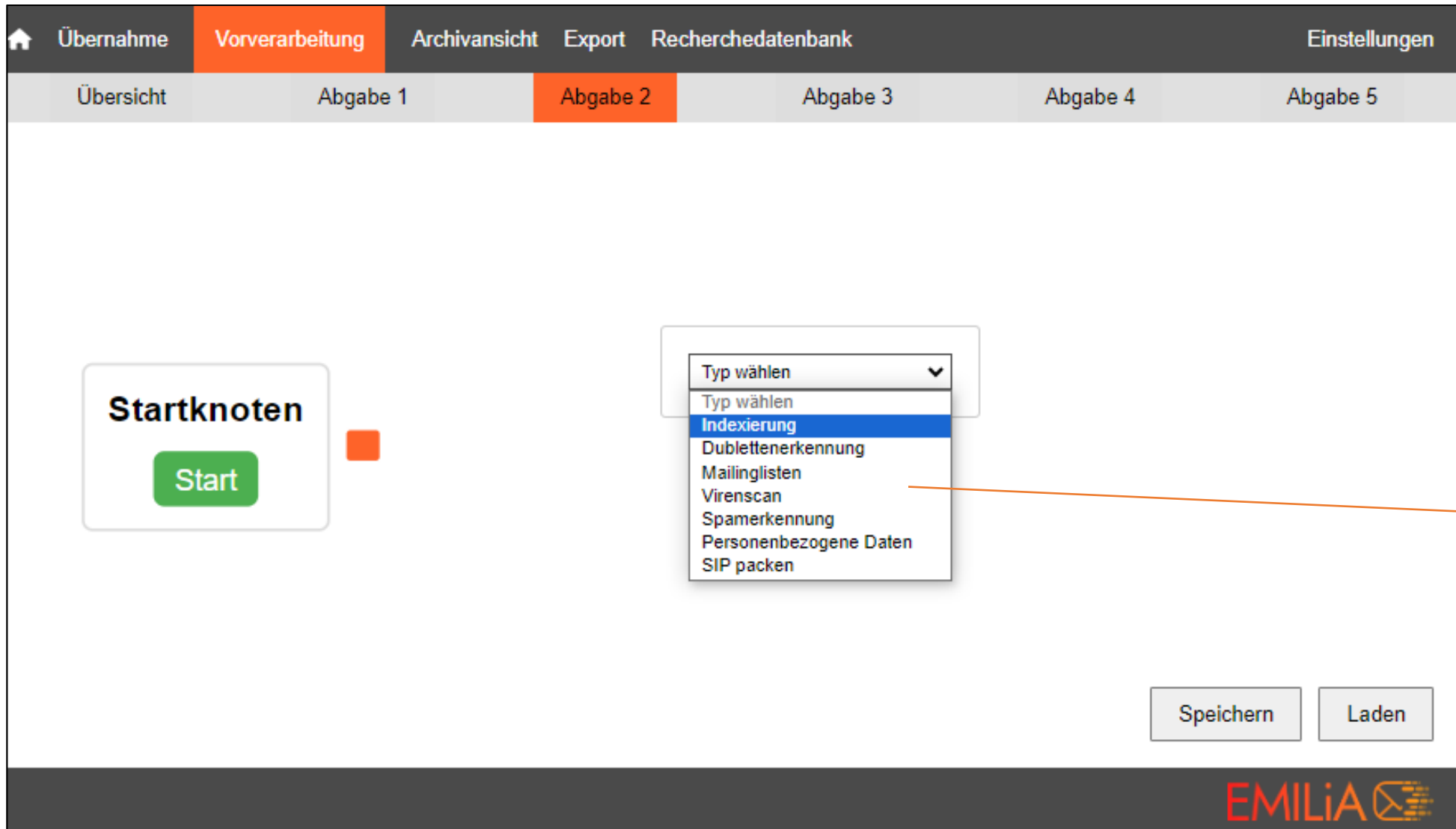
Aktiver Prozess

Startknoten

Andockpunkt

Vorverarbeitung (Prototyp)

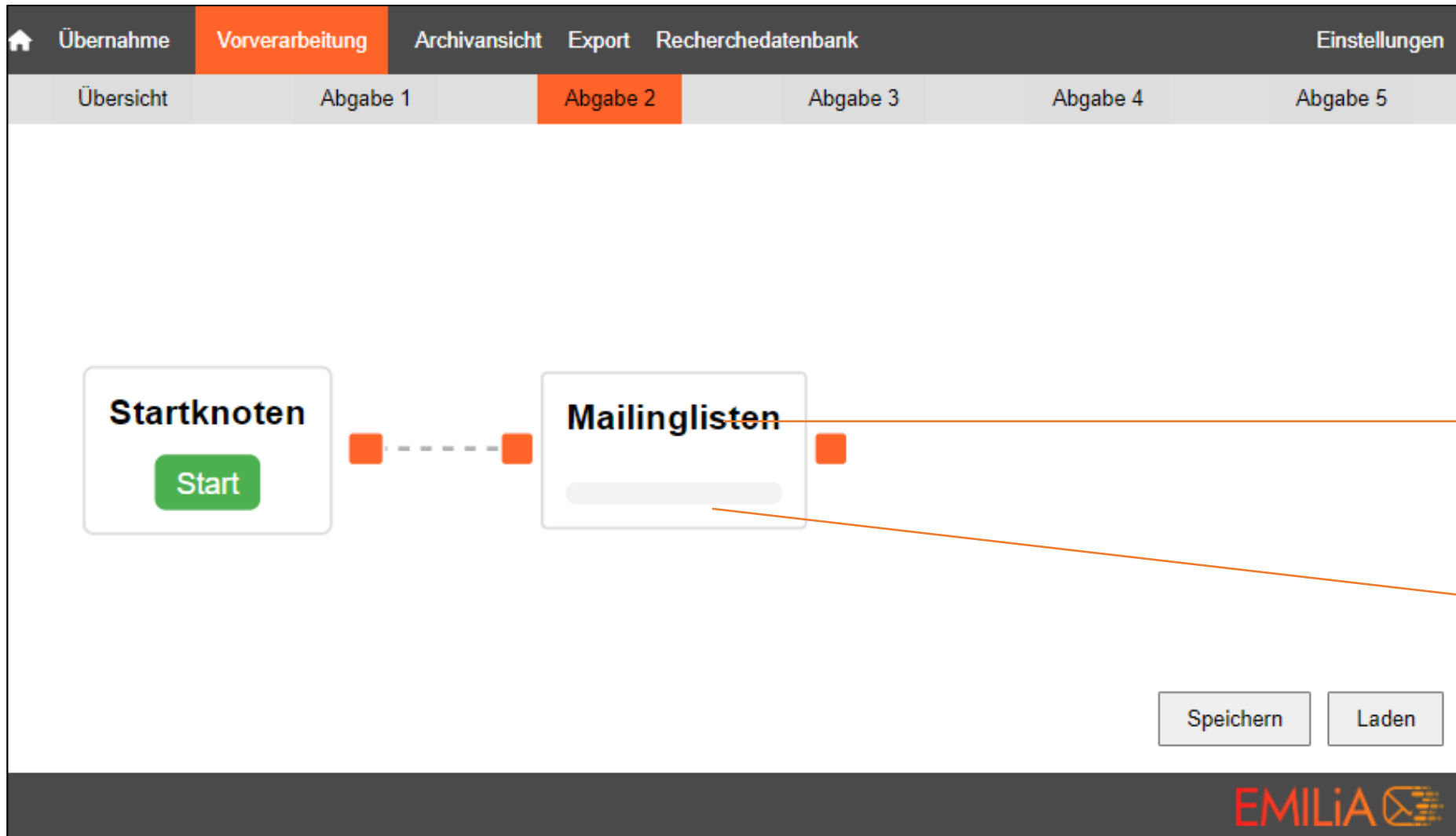




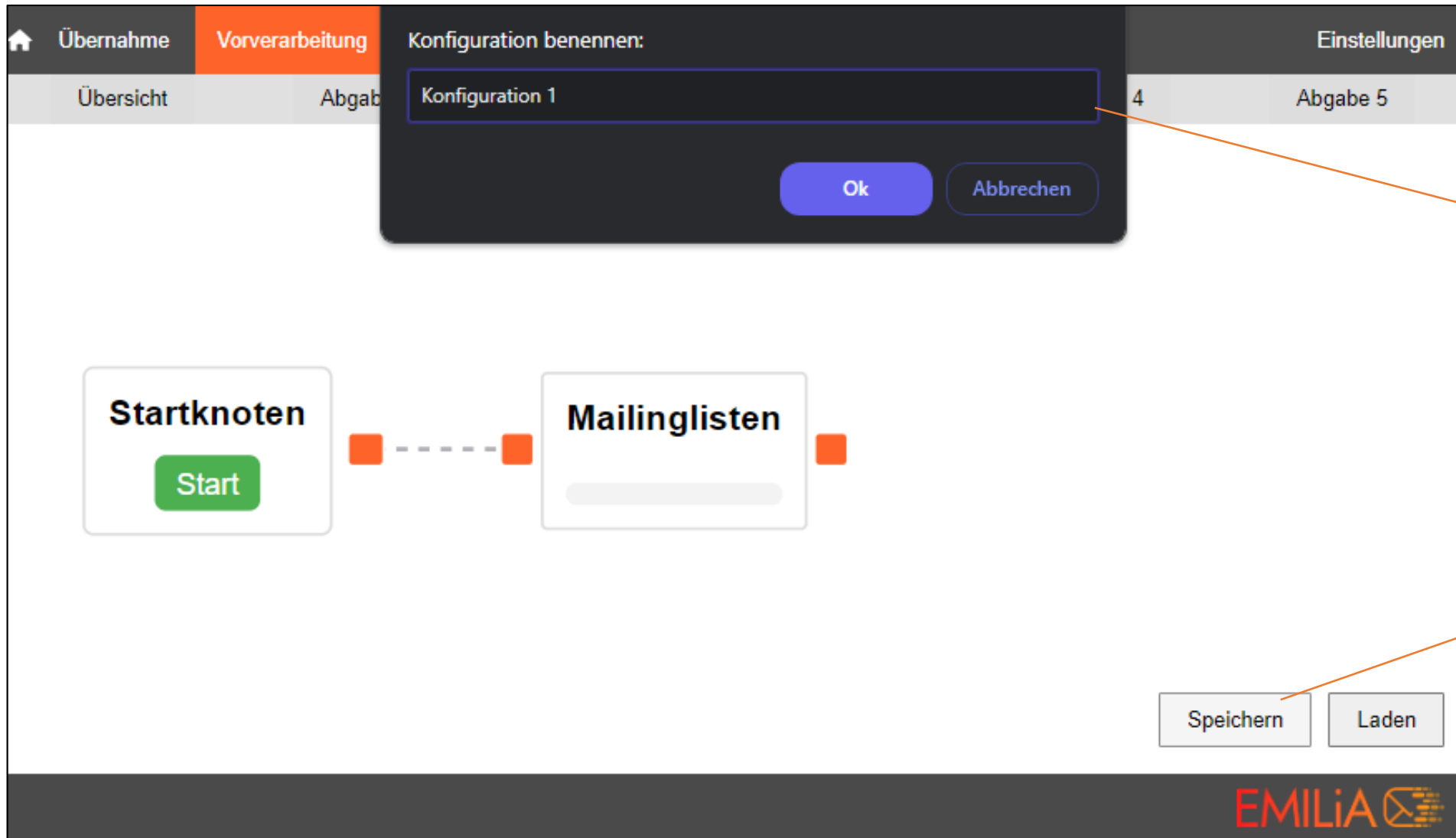
Auswahl des nächsten Knotens

Vorverarbeitung (Prototyp)

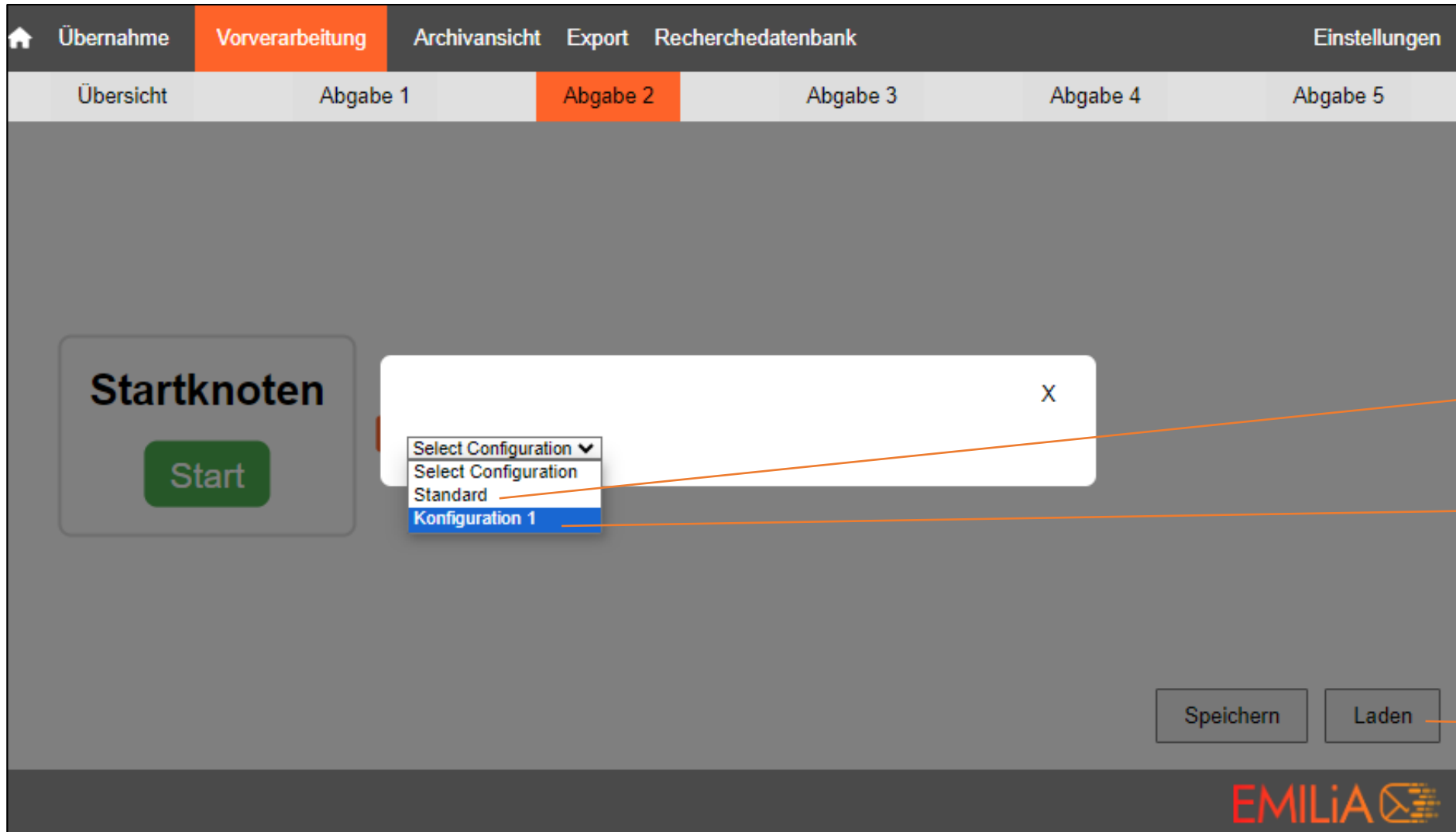
EMILiA



Vorverarbeitung (Prototyp)



Vorverarbeitung (Prototyp)



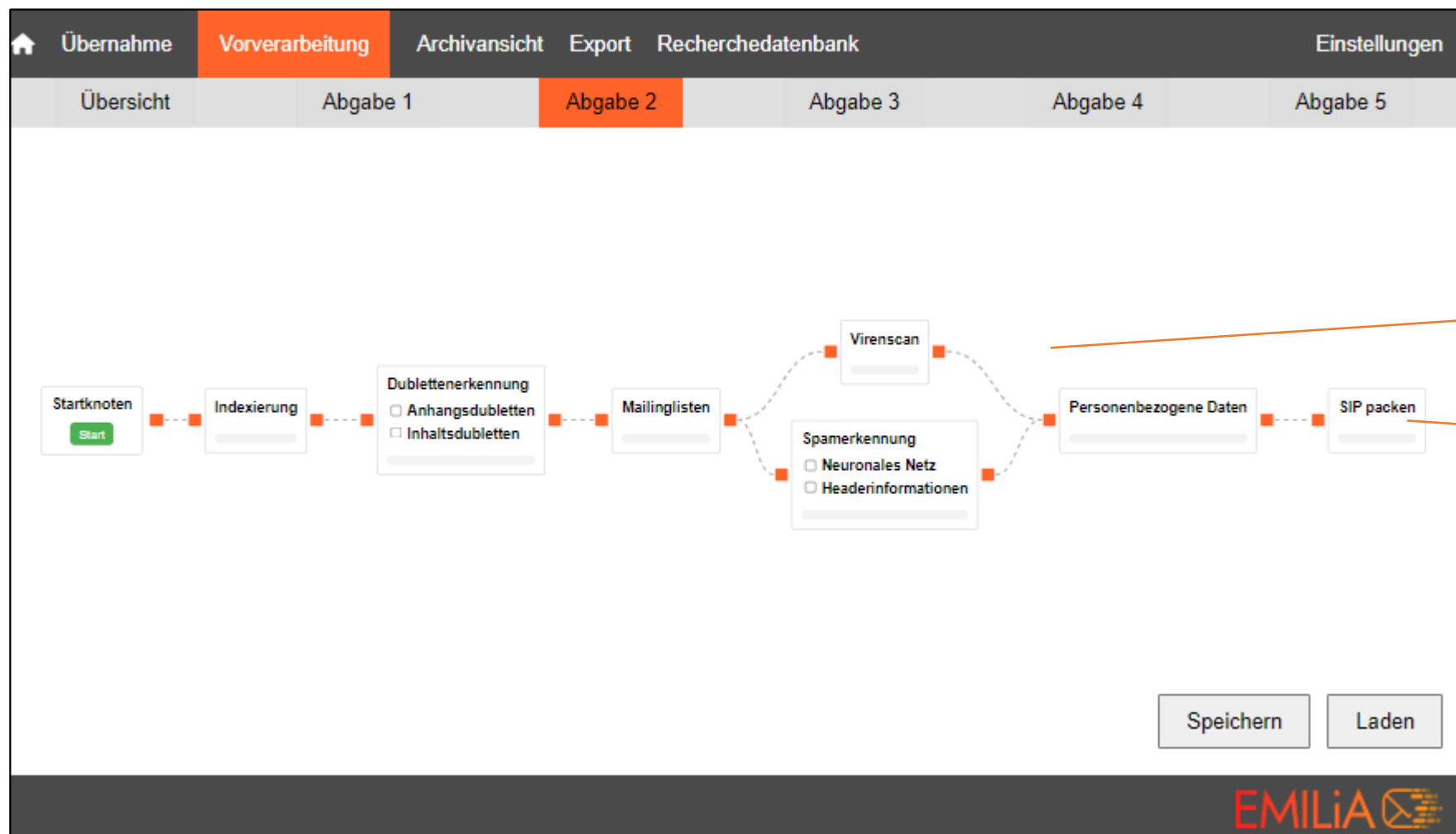
Standardkonfiguration

Eigene Konfiguration

Ladebutton

Vorverarbeitung (Prototyp)





Beispielkonfiguration

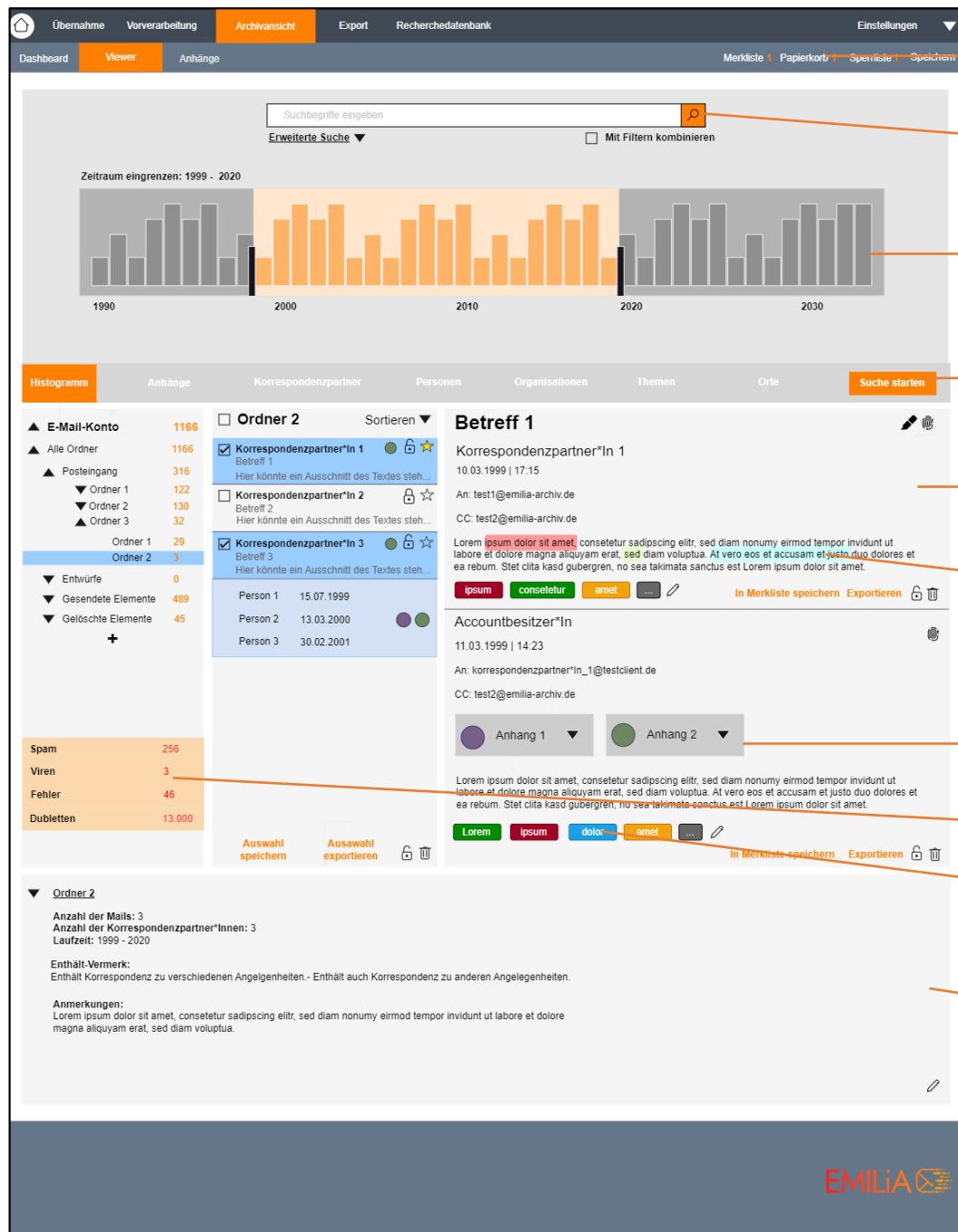
Endknoten

Vorverarbeitung (Prototyp)



Vision

5



Papierkorb

Suchschlitz

Histogramm

Filtermöglichkeiten

An E-Mail-Clients
angelegter Viewer

Annotationen

Anhänge

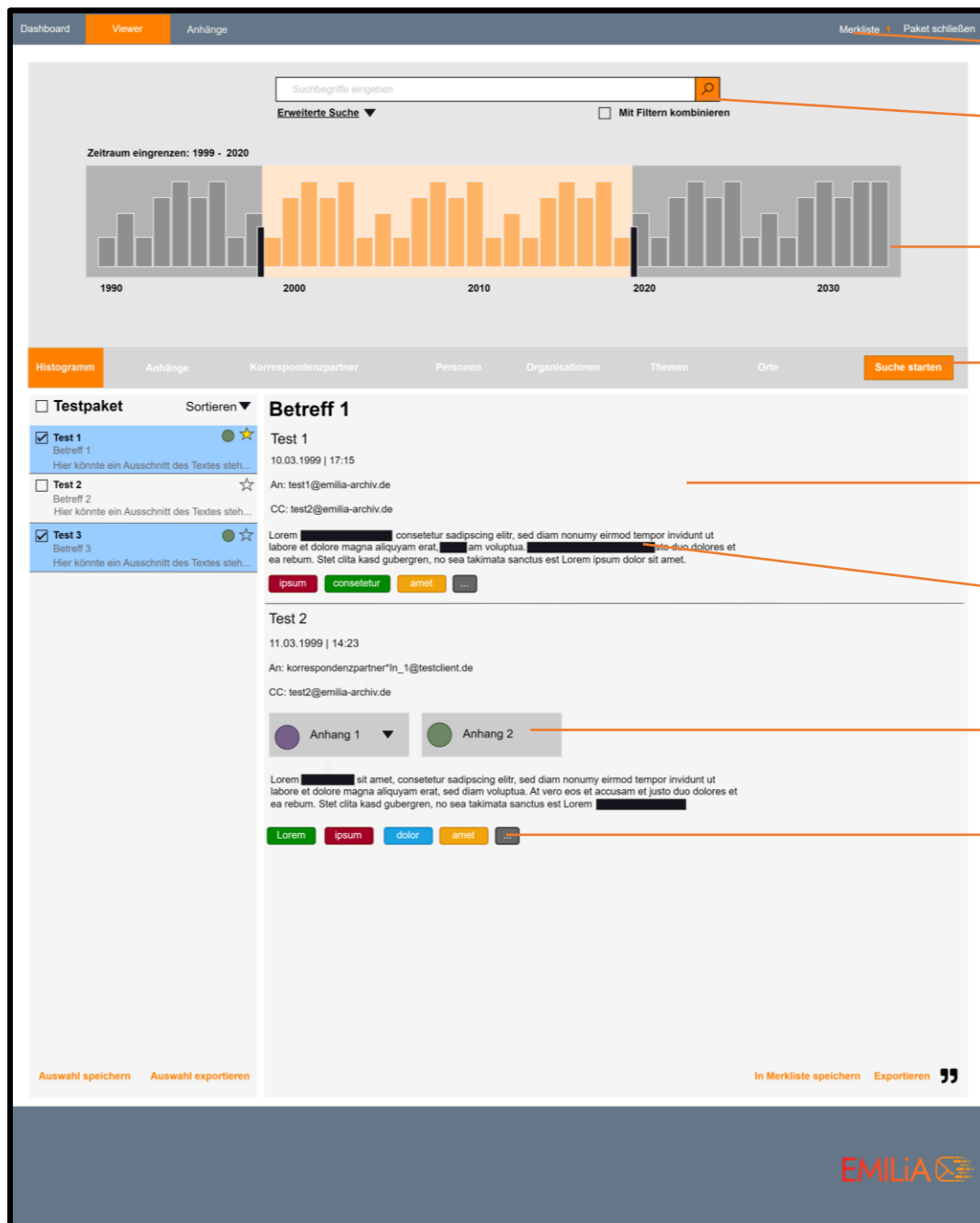
Bewertungsvorschläge

Schlagwörter

Freitextfelder für die
Erschließung

Archivansicht (Konzept)





Merkliste

Suchschlitz

Histogramm

Filtermöglichkeiten

An E-Mail-Clients
angelegter Viewer

Anonymisierung

Anhänge

Schlagwörter

Forschendenansicht (Konzept)



6

Diskussionsrunde

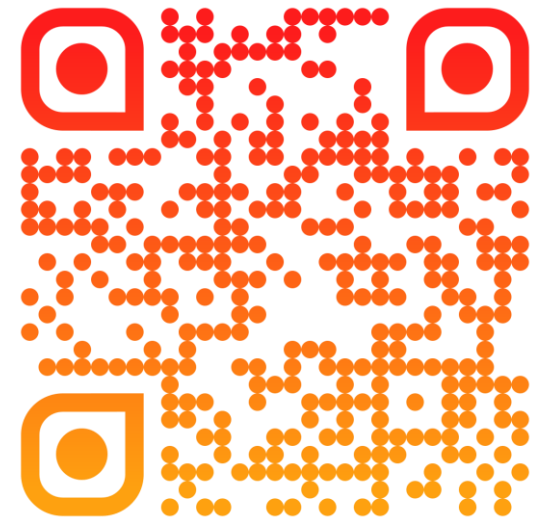
Vielen Dank für Ihre Aufmerksamkeit!

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 