EMILiA

Die Suche nach der Nadel im Heuhaufen.

Entwicklung einer teilautomatisierten Software für die Archivierung von E-Mails

Nico Beyer und Felix Gericke

IBB Business Team

Freie Universität Berlin

ARCHIV DER MAX-PLANCK-GESELLSCHAFT

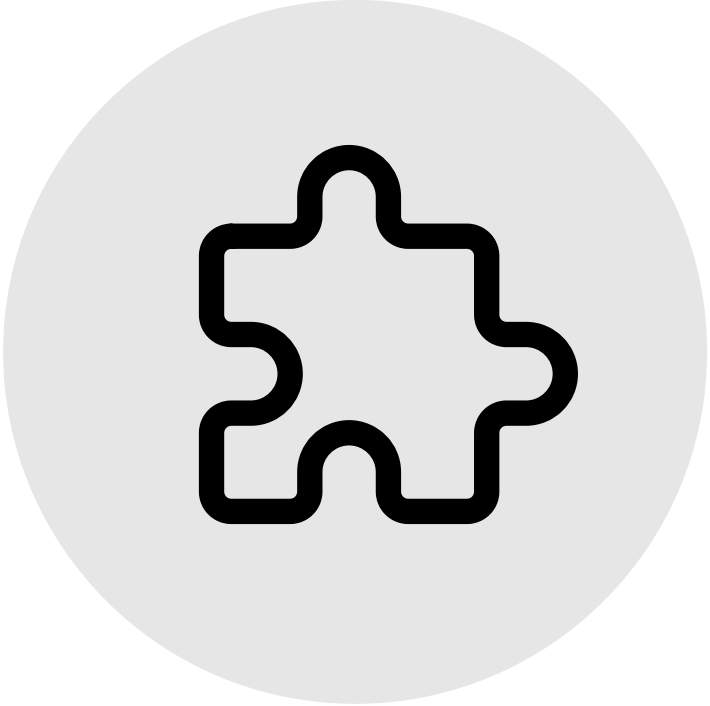
Wie können Archive historisch relevante E-Mails **übernehmen**, **bearbeiten**, für die **digitale Langzeitspeicherung vorbereiten** und möglichst **zeitnah nutzbar machen**?

EMILiA

Projektrahmen

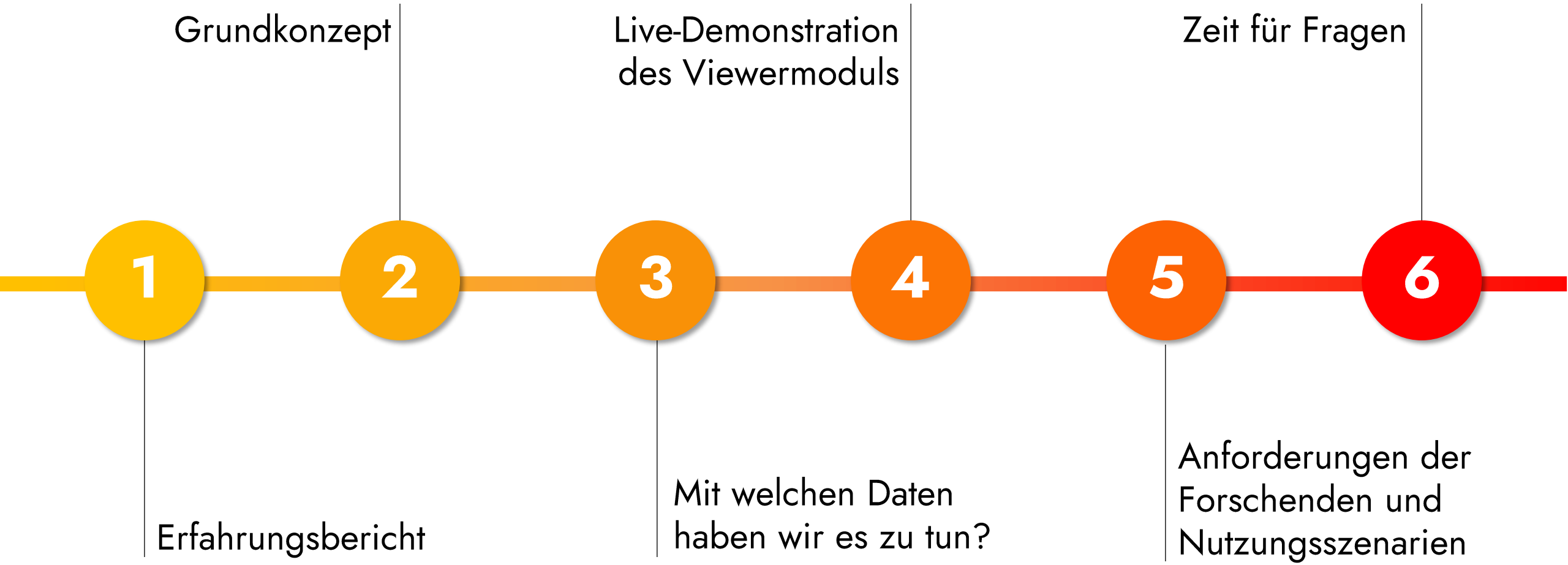
Projektrahmen

- 2015: Beginn der Konzeption im Archiv der MPG in Kooperation mit dem Fachbereich Informatik der Freien Universität Berlin
- 2020: Beginn der Entwicklung eines Prototyps
- 2024: Förderprogramm „ProValid“ der Investitionsbank Berlin
- Aktuelles Kernteam: 2 Informatiker und 1 Archivar



EMILiA

Was wir heute vorhaben


1
2
3
4
5
6
Grundkonzept
Live-Demonstration des Viewermoduls
Zeit für Fragen
Erfahrungsbericht
Mit welchen Daten haben wir es zu tun?
Anforderungen der Forschenden und Nutzungsszenarien

Erfahrungsbericht

Technische Aspekte

- Die gängigen E-Mail-Formate gehen mit zahlreichen 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

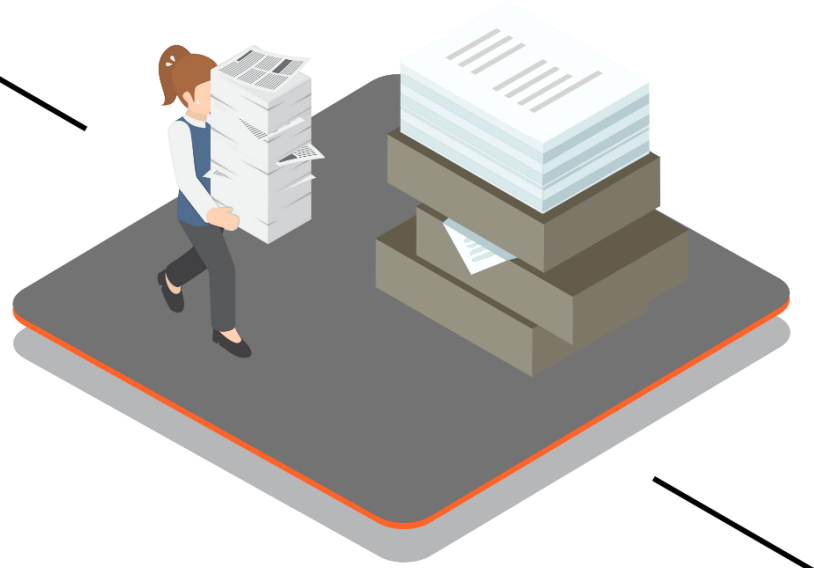
- Die Postfächer im Archiv der MPG umfassen durchschnittlich **50.000** Nachrichten, die bisher größte Übernahme enthielt **133.194** E-Mails und **258.180** Anhänge

- Spam, Werbemails und die Niedrigschwelligkeit elektronischer Kommunikation führen zu riesigen Mengen nicht archivwürdiger Daten

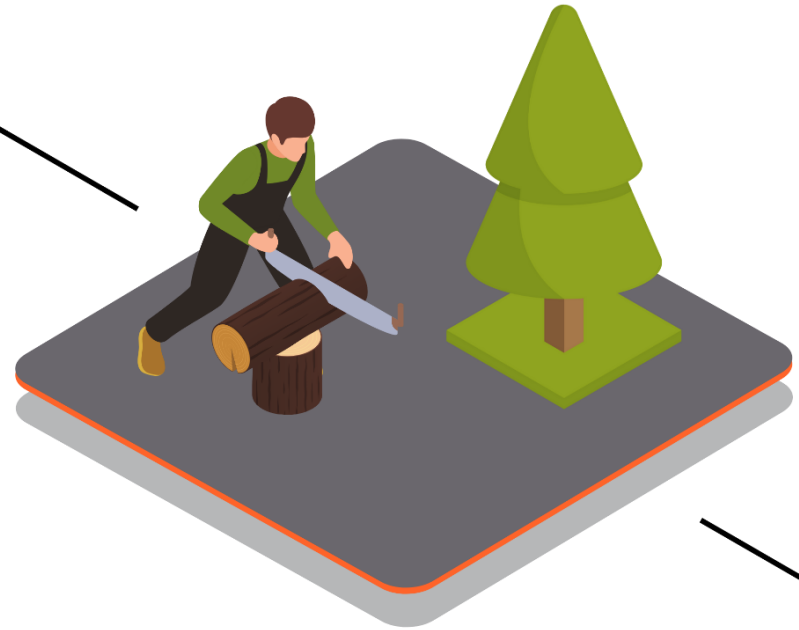
- Inmitten dieses Datenmülls können sich aber auch zahlreiche überlieferungswürdige Nachrichten befinden

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

EMILiA

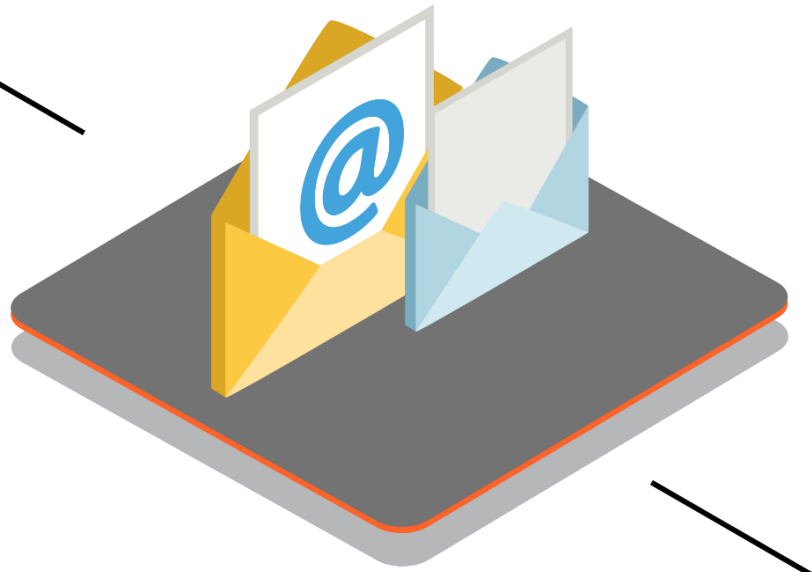


- 600.257 A4-Seiten
- 7 Tage Druckzeit bei 60 S. pro Minute
- ~ 100 laufende Meter

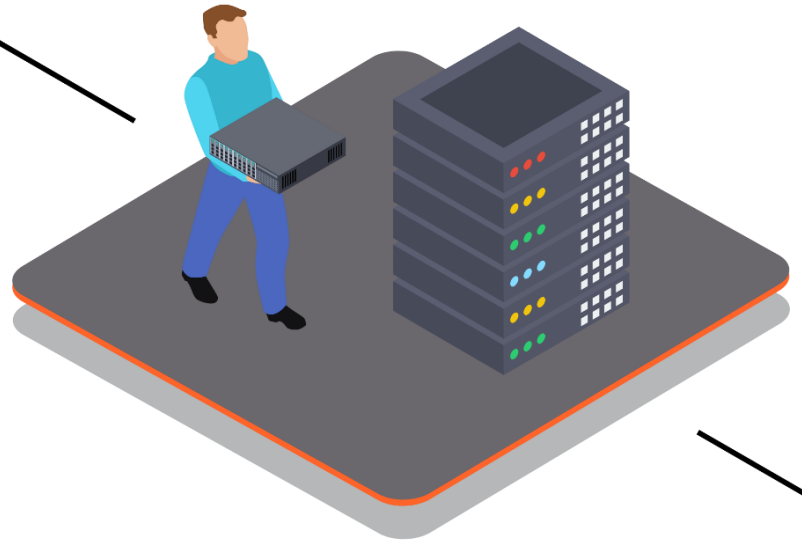


- 5 Bäume
- 25 m ⬍⬍
- 40 cm ⌀

EMILiA

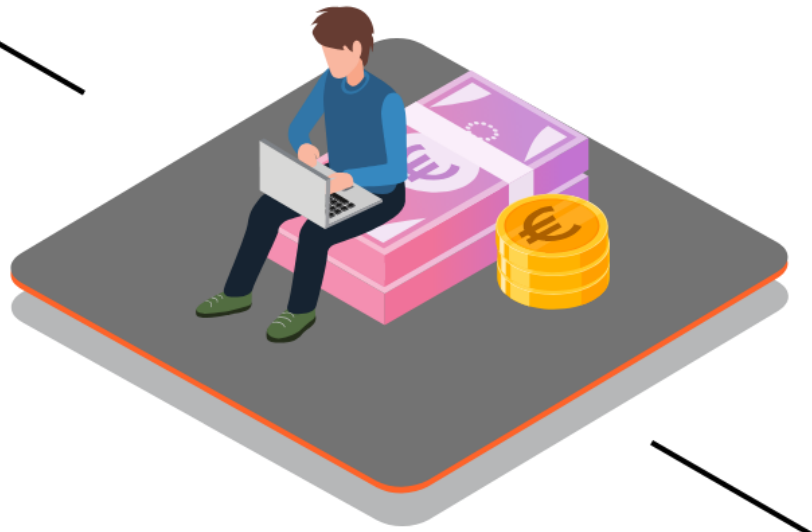


- 213.533 Seiten Dubletten
- 247.220 Seiten Mailinglisten
- 363.111 Seiten Mailinglisten und Dubletten

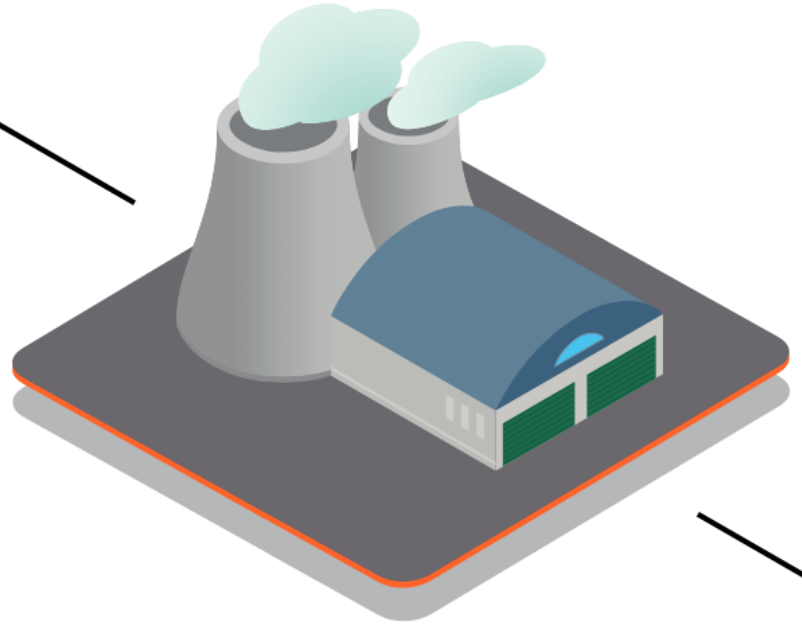


- ~21 GB
(ohne Bewertung)
- ~ 9 GB
(mit Bewertung)

EMILiA



- 37.632 € Speicherkosten pro Jahr (bei 600 Konten und 3€ pro GB)
- 21.666 € Ersparnis durch Reduktion



- **Bewertung von 600 Konten spart 324,31 KWh und 245 kg CO_2**

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

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

Mit welchen Daten haben wir es zu tun?

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.

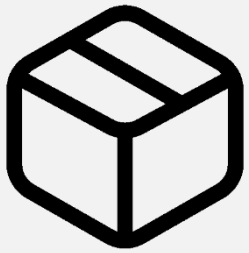
Mit freundlichen Grüßen
EMILiA-Projektteam

E-Mail-Body

Anhänge

EMILiA

MBOX und PST

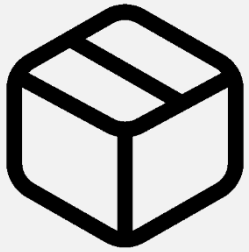


.PST

+

.MSG

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



.MBOX

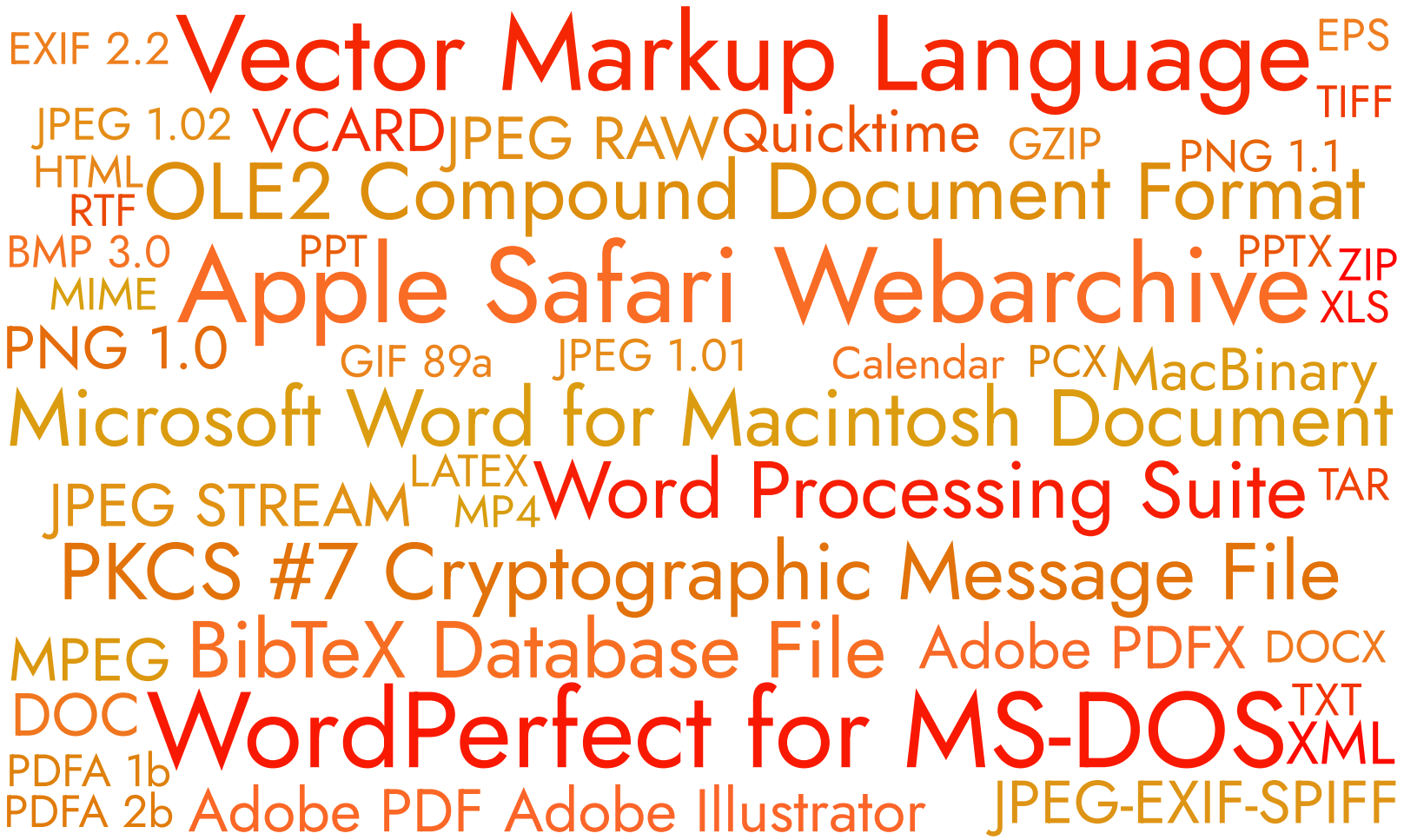
+

.EML

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

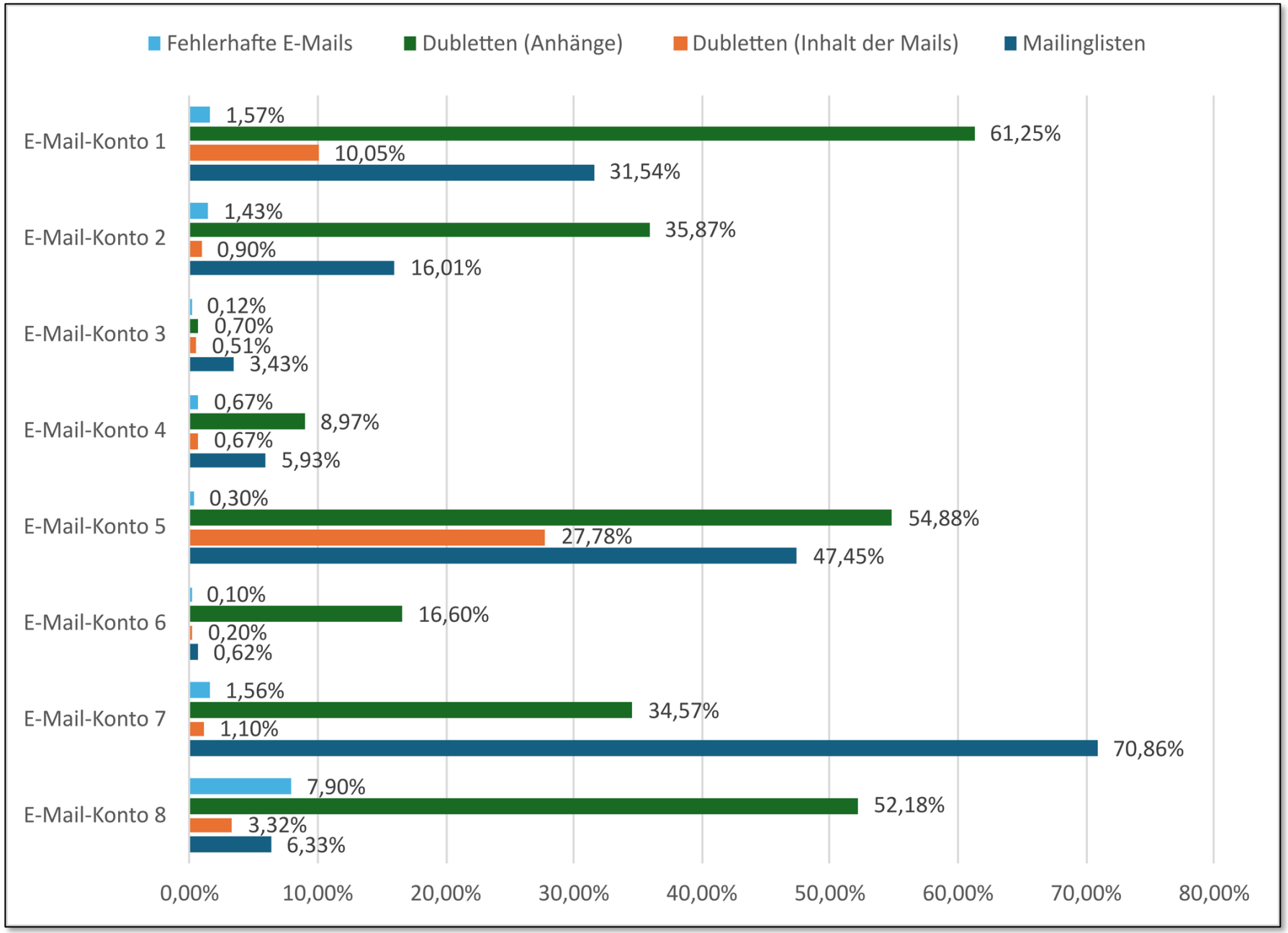
EMILiA

Anhänge



EMILiA

Live-Demonstration



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

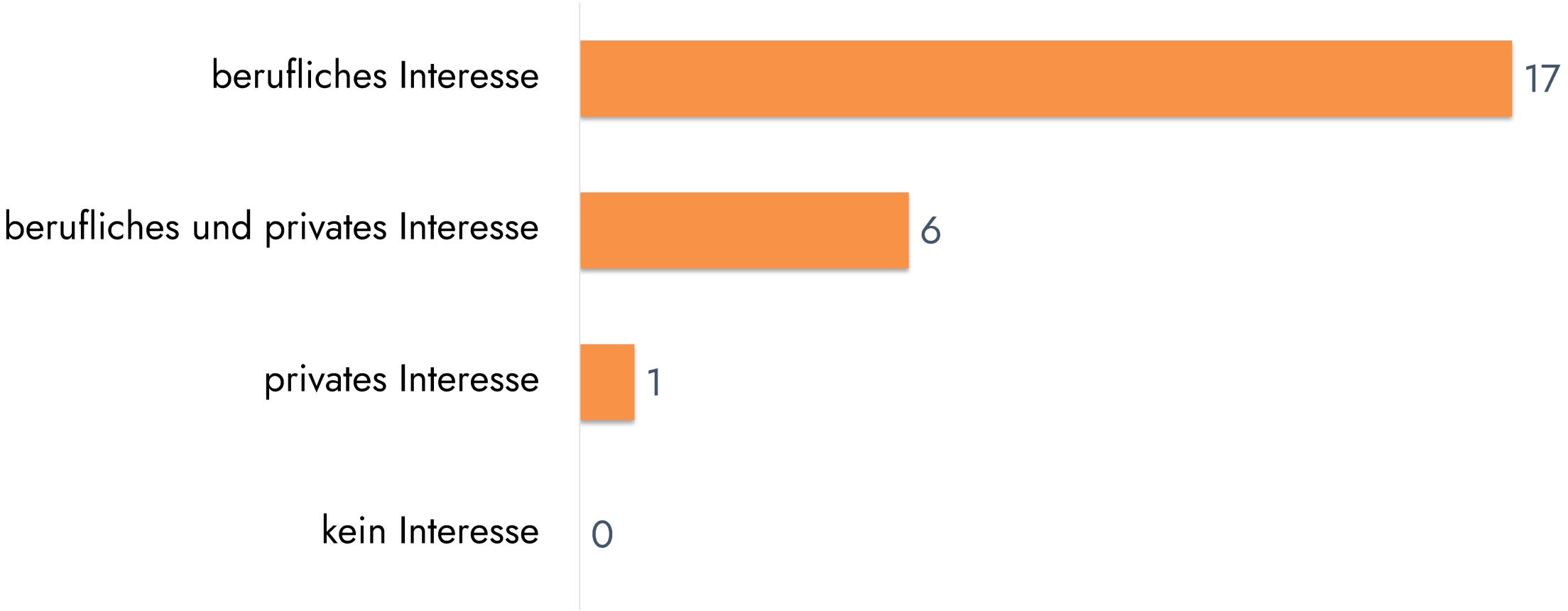
EMILiA

Anforderungen der Forschenden

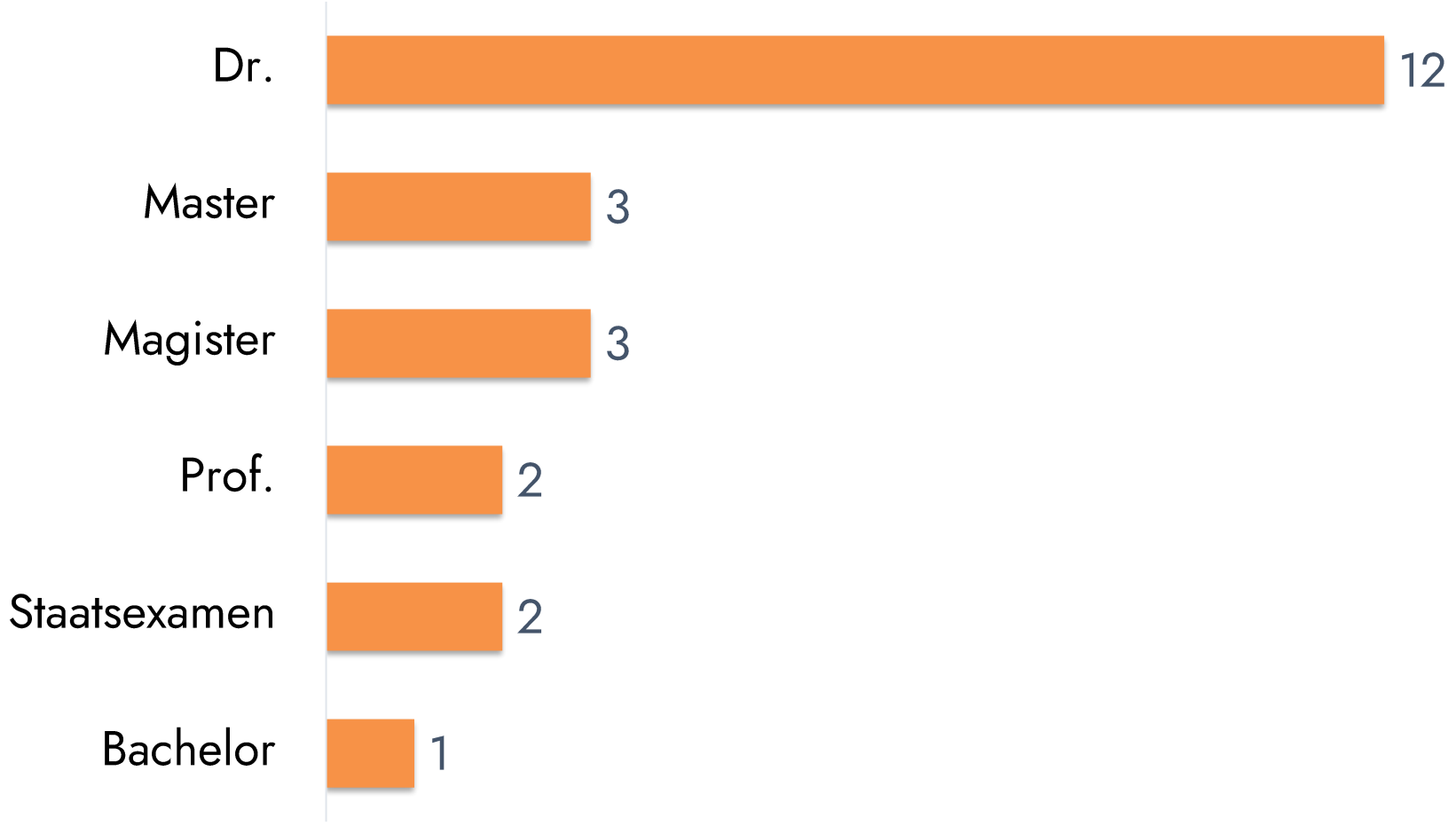
Auswertung einer Umfrage zur Nutzung von E-Mails (vorläufiges Ergebnis)

Allgemeine Angaben

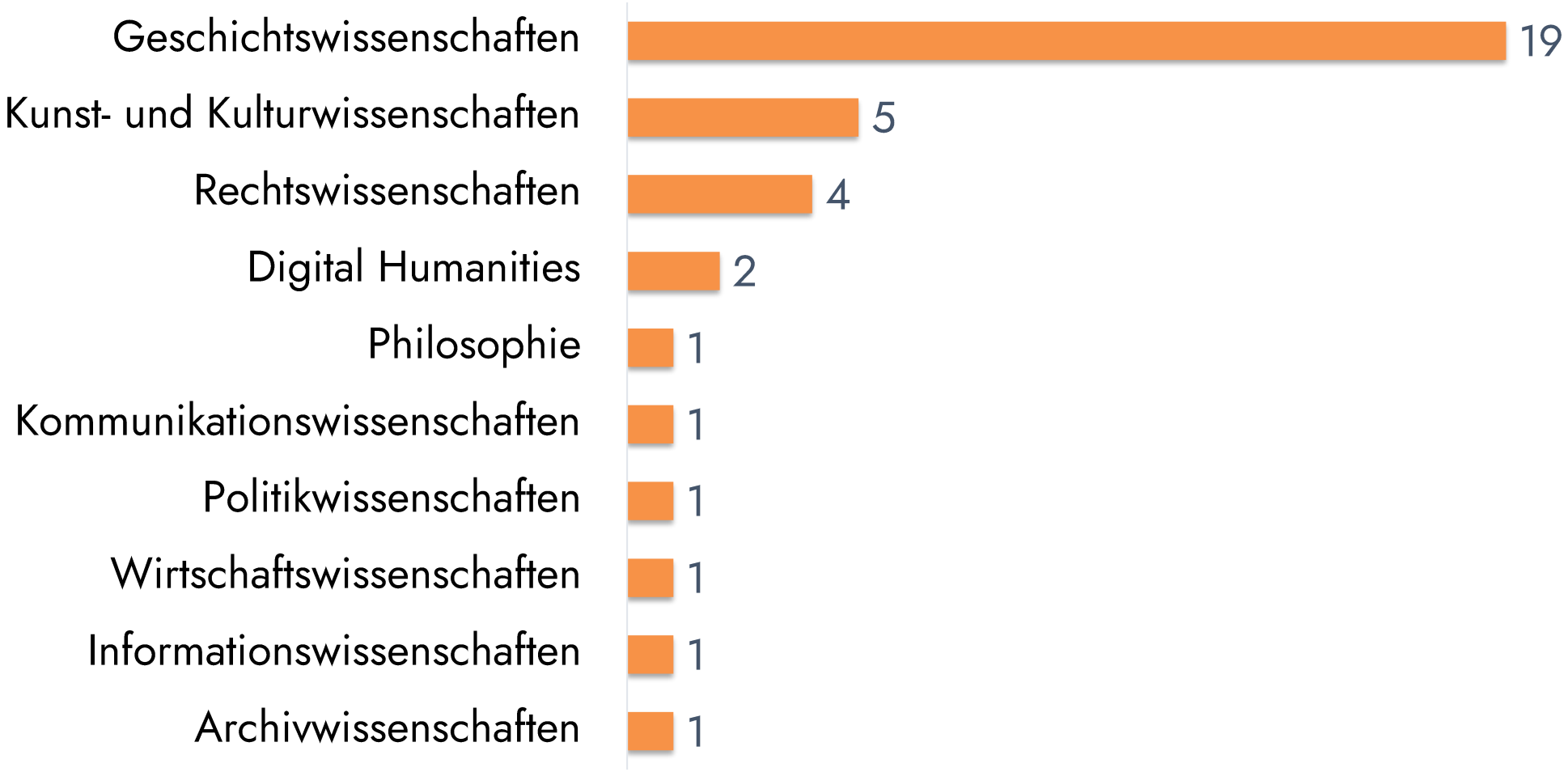
Auf welcher Ebene interessieren Sie sich für Archivgut?



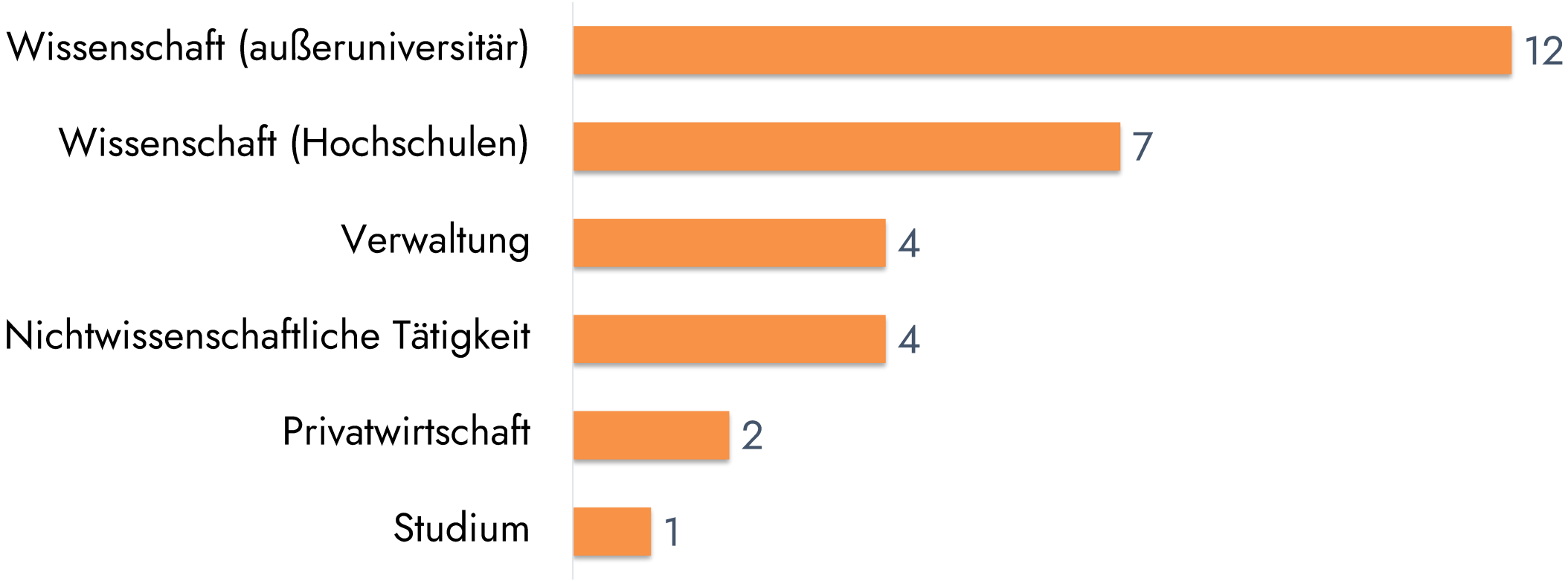
Bitte geben Sie Ihren akademischen Grad an.



Welchen Fachgebieten gehören Sie an?

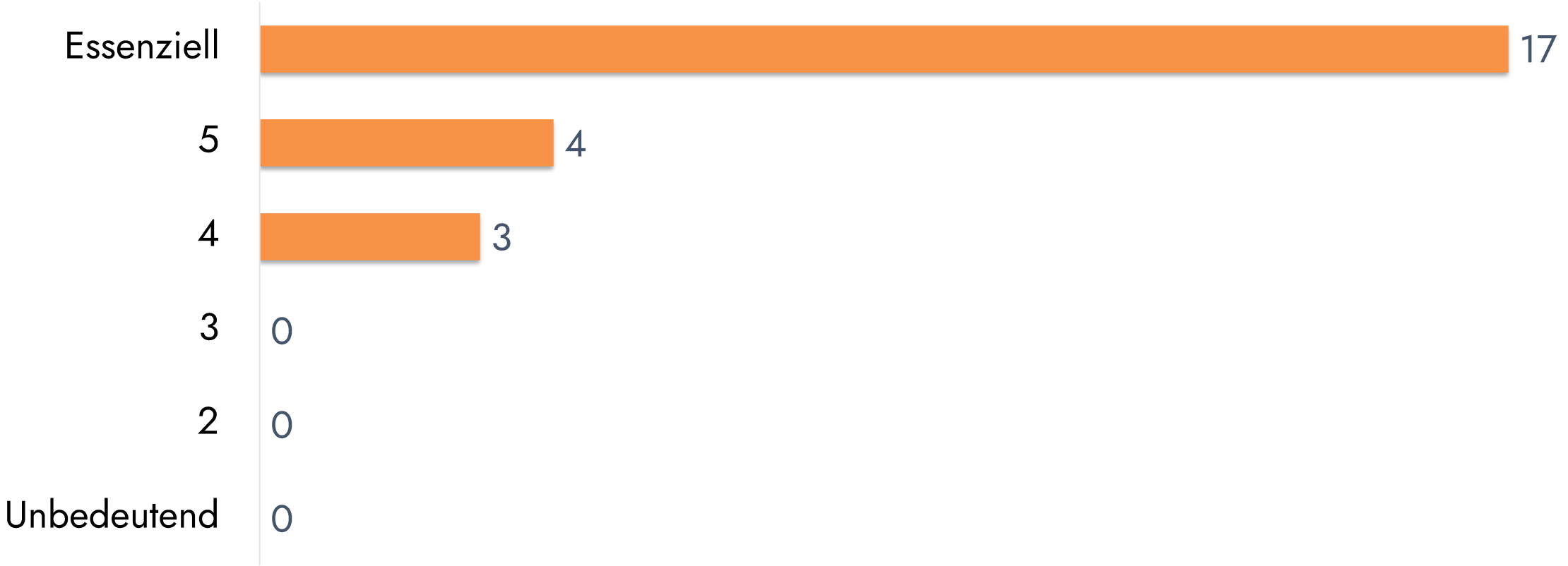


Welchen dieser Bereiche lässt sich Ihre aktuelle Tätigkeit zuordnen?



Informationswert von E-Mails

Wie schätzen Sie den Stellenwert von E-Mails für die künftige Forschung ein?

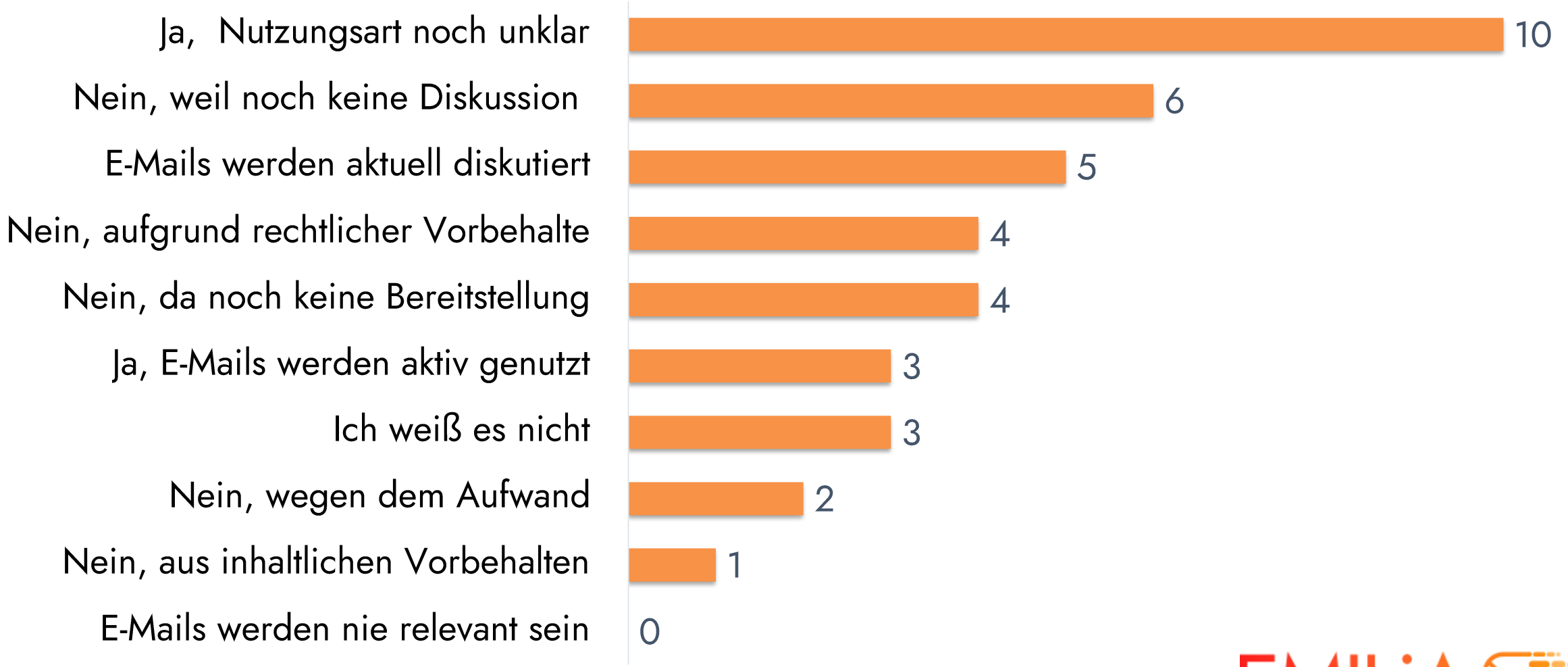


Bitte begründen Sie Ihre Antwort kurz.

- *„E-Mails sind ein zentrales Instrument privater wie beruflicher/wissenschaftlicher Kommunikation“*
- *„Oftmals maßgebliche Kommunikation im Rahmen der Entscheidungsfindung“*
- *„E-Mail (inkl. Anhänge!) als Ersatz des Briefverkehrs, E-Mail als elementares Kontextmaterial zu heterogenen Fileablagen, Fachverfahren, etc.“*
- *„E-Mails sind äquivalent zum Briefverkehr vergangener Jahrhunderte. Interessant für die Netzwerkforschung, den Gedankenaustausch etc.“*
- *„Die gesamte berufliche und wissenschaftliche Kommunikation findet per E-Mail statt. Wer das künftig erforschen will, braucht E-Mails als Quelle.“*
- *„Kommunikationsmittel und teilweise Ersatz für klassische Akten“*

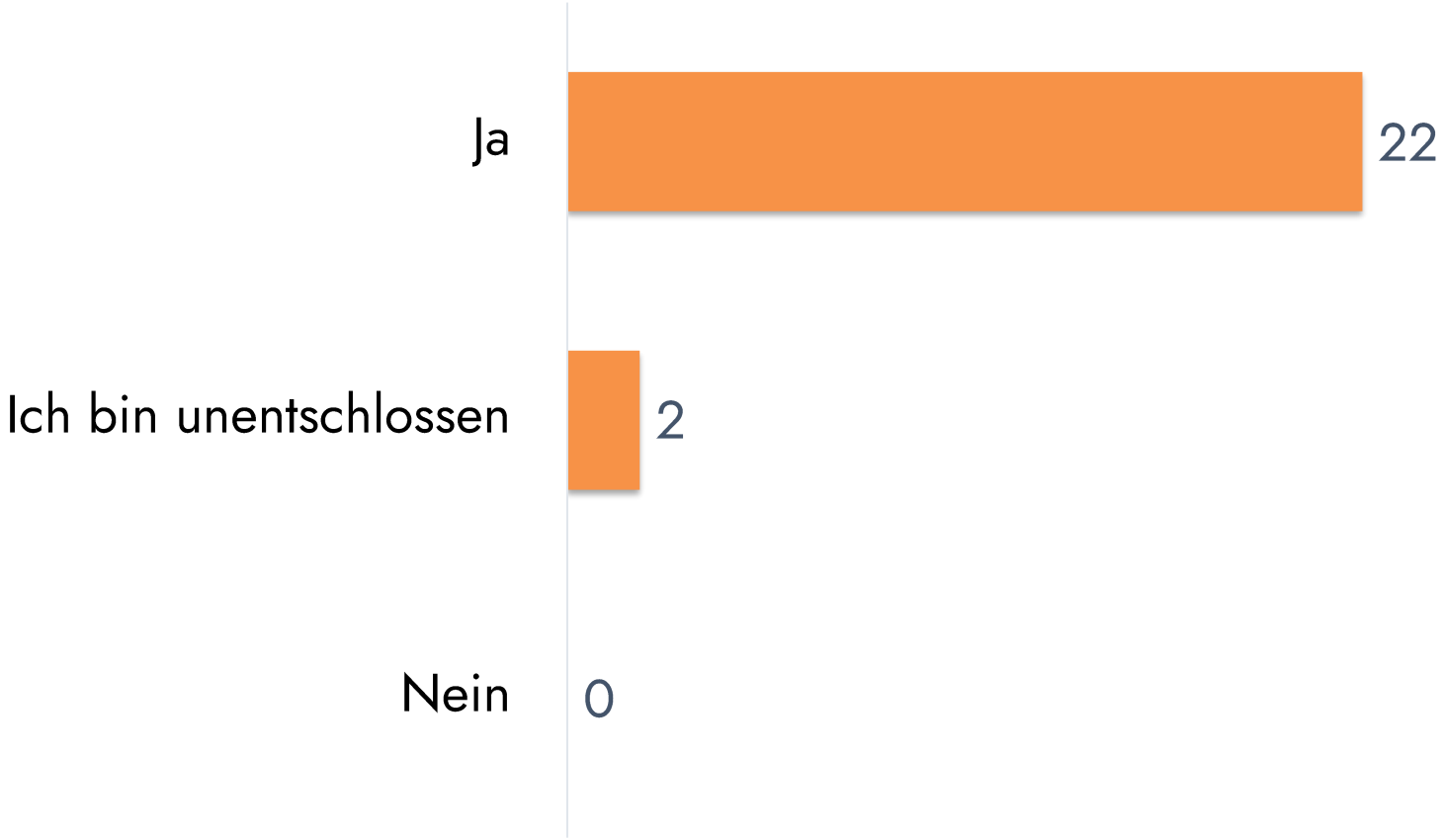
EMILiA

Werden E-Mails in Ihrem Forschungsfeld als potenziell relevante Quellen betrachtet? Welche Aussagen sind in ihrem Umfeld hauptsächlich vertreten?

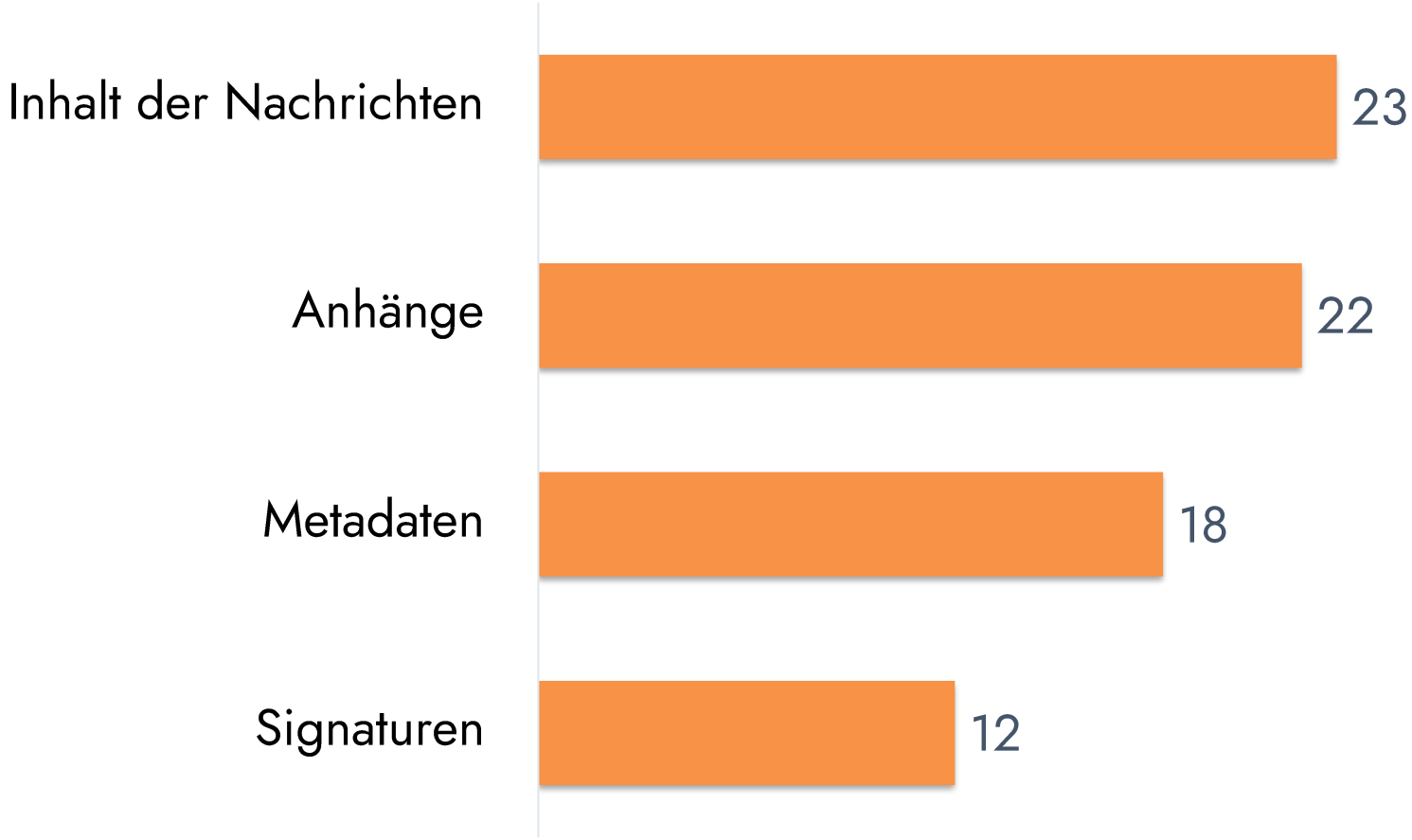


Aufbereitung von E-Mails

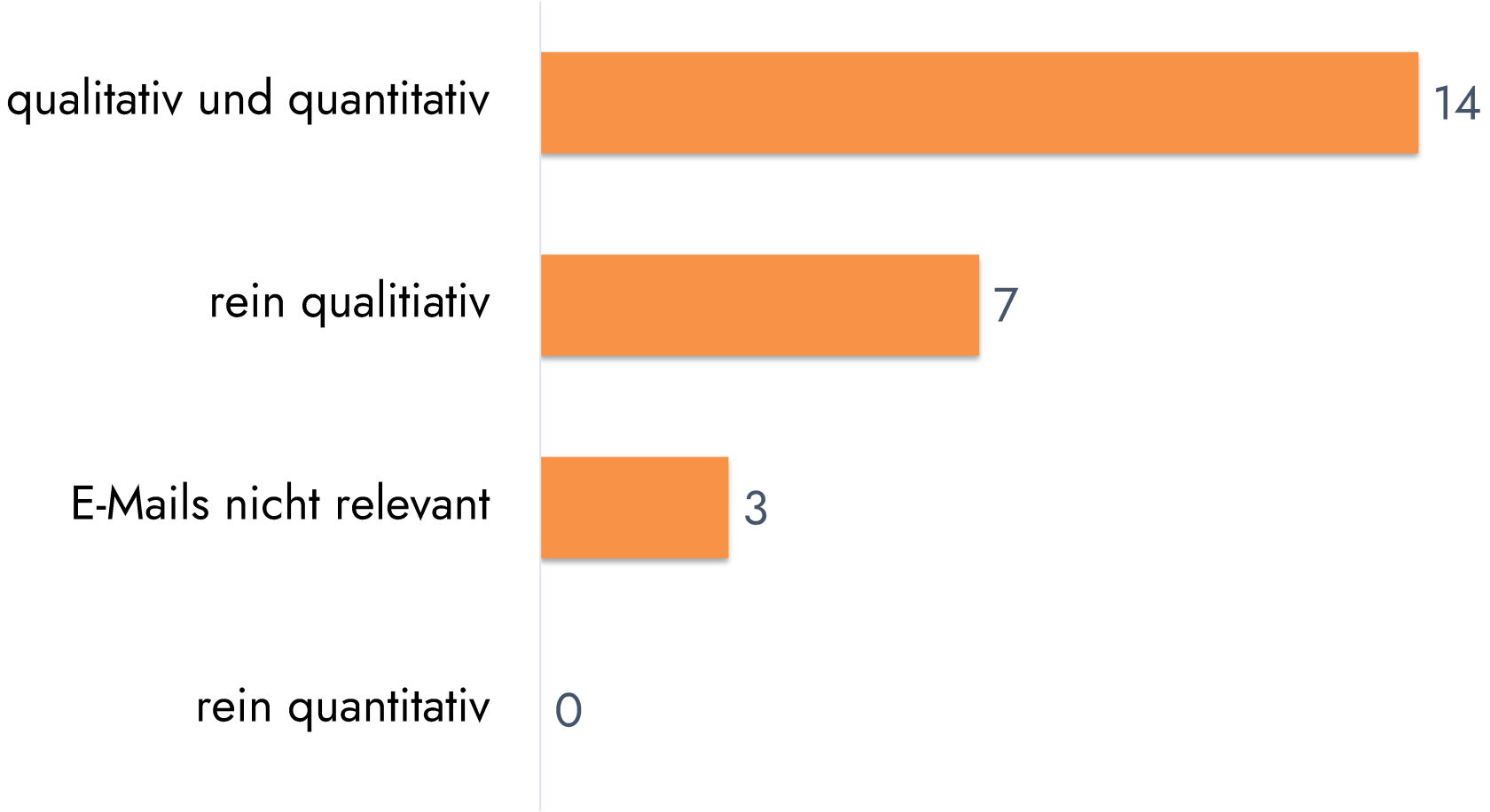
Möchten Sie bei der Nutzung archivierter E-Mails dieselben Funktionalitäten zur Verfügung haben, die ihnen heute verbreitete E-Mail-Clients bieten?



Welche Bestandteile eines E-Mail-Postfachs sollten Ihrer Meinung nach überliefert werden?

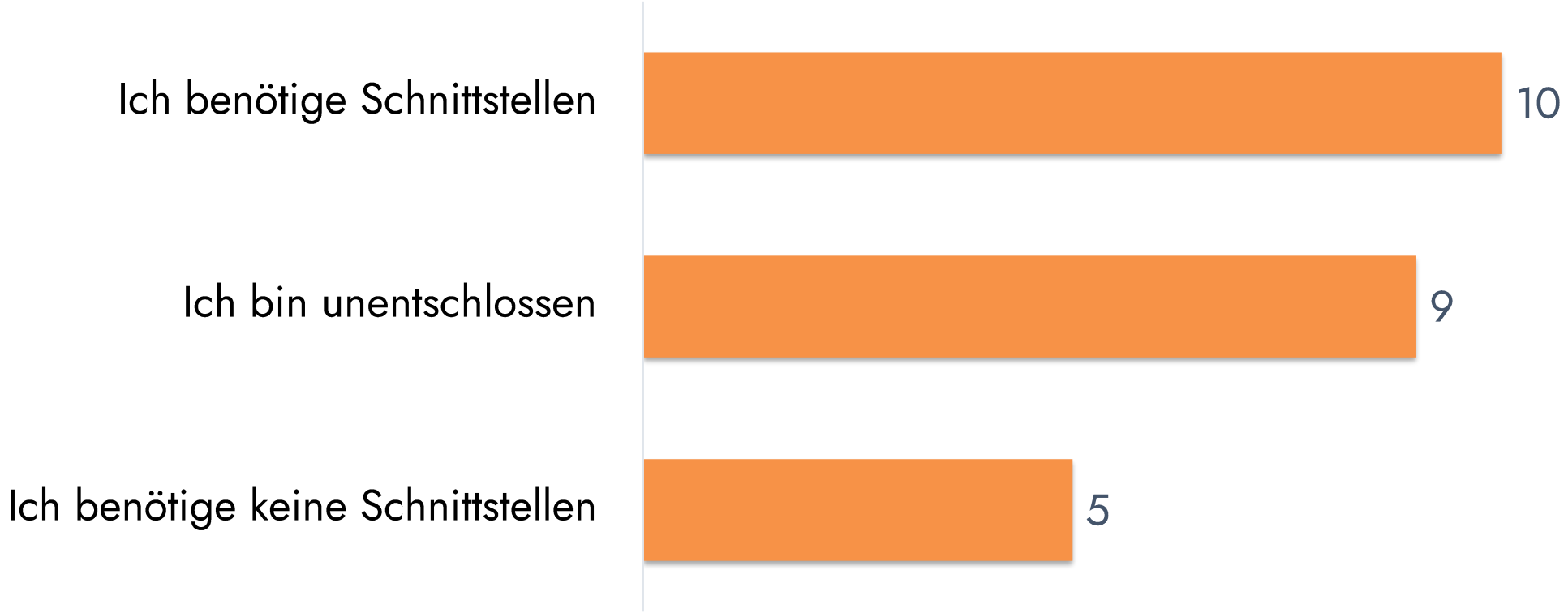


Sind Sie eher an einer quantitativen oder einer qualitativen Auswertung von relevanten E-Mail-Postfächern interessiert?

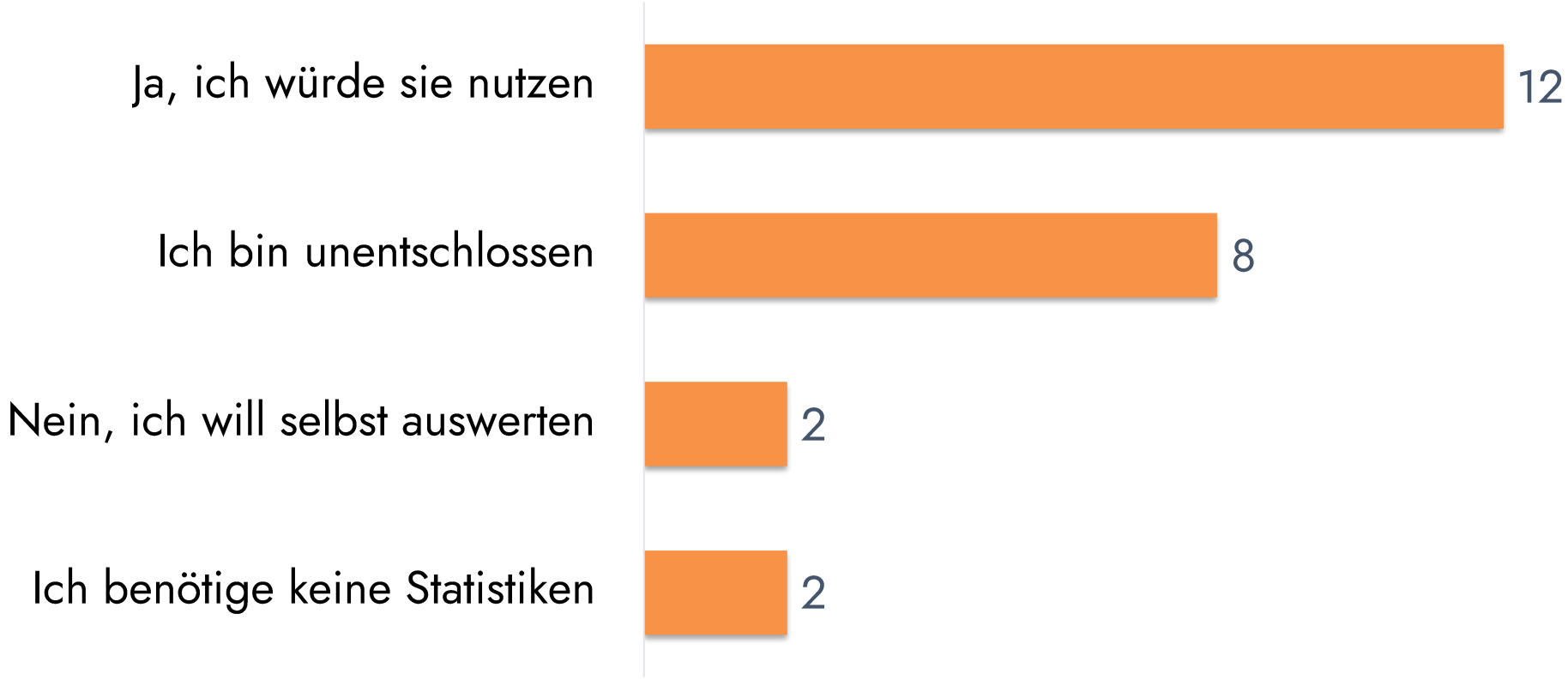


Benötigte Funktionen

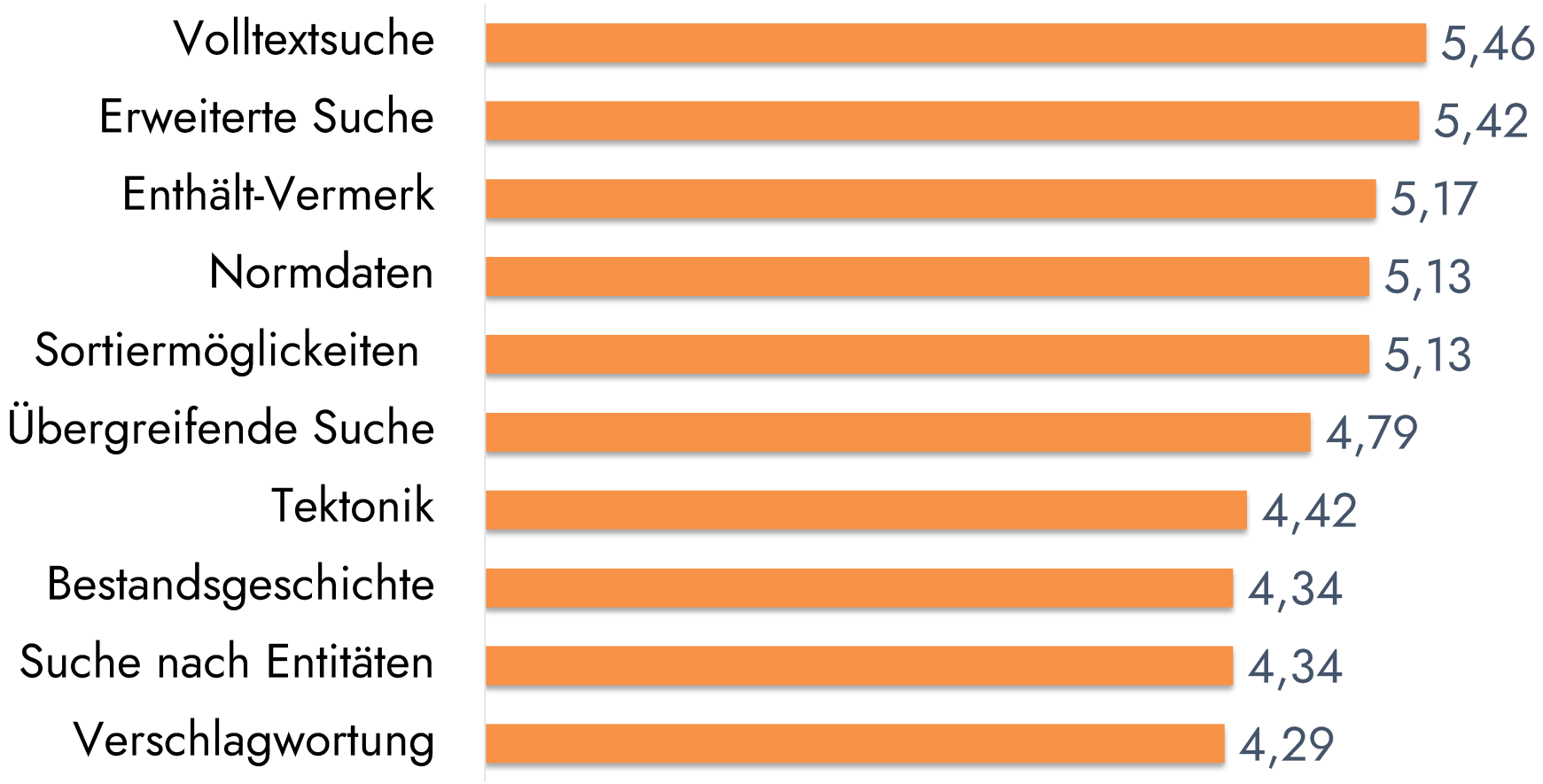
Benötigen Sie Schnittstellen, um selbst mit den archivierten E-Mails zu arbeiten oder genügt Ihnen eine vom Archiv zur Verfügung gestellte Darstellungssoftware?



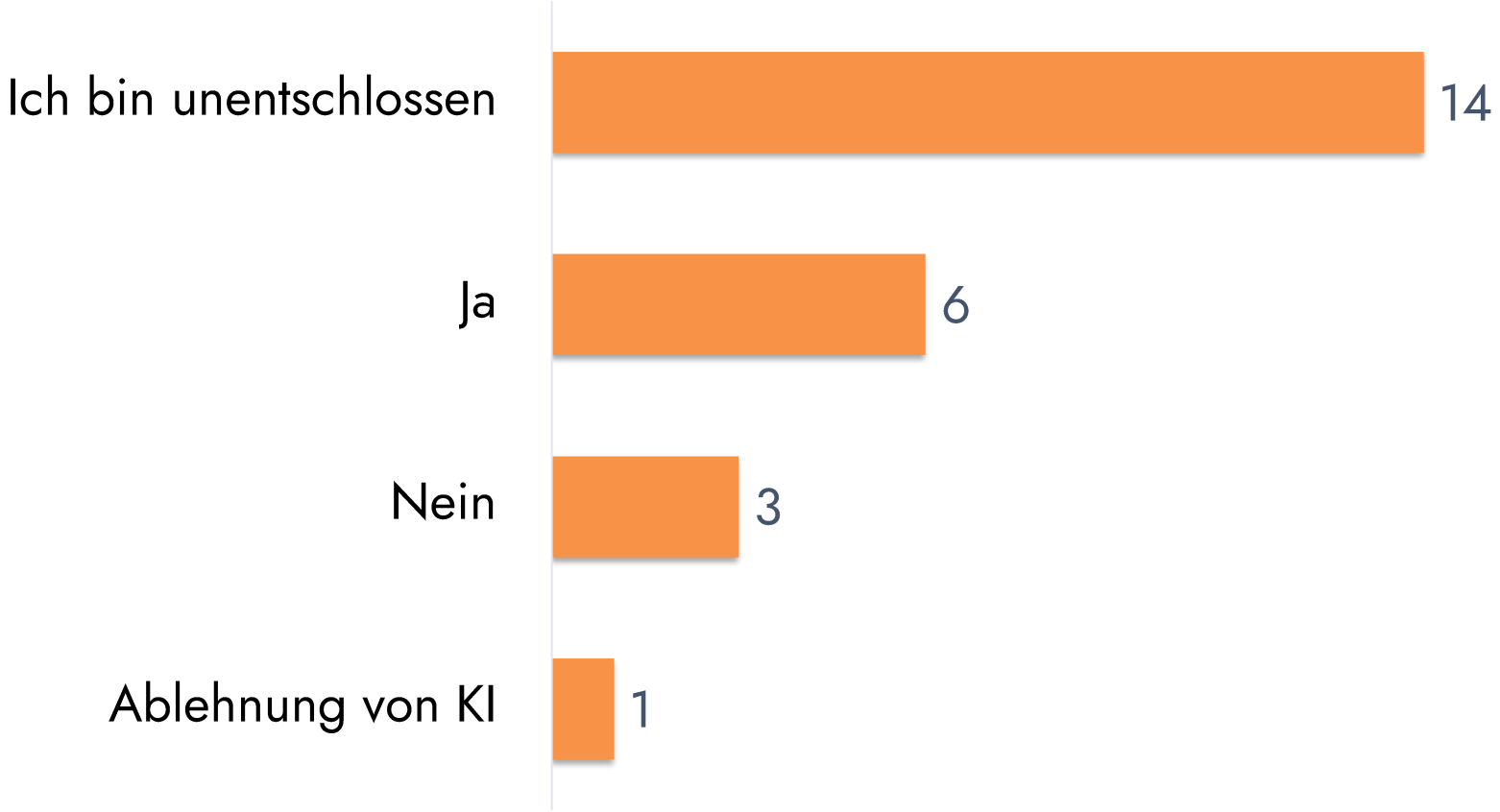
Denken Sie, dass die von uns entwickelte Software Statistiken für eine quantitative Auswertung zur Verfügung stellen sollte?



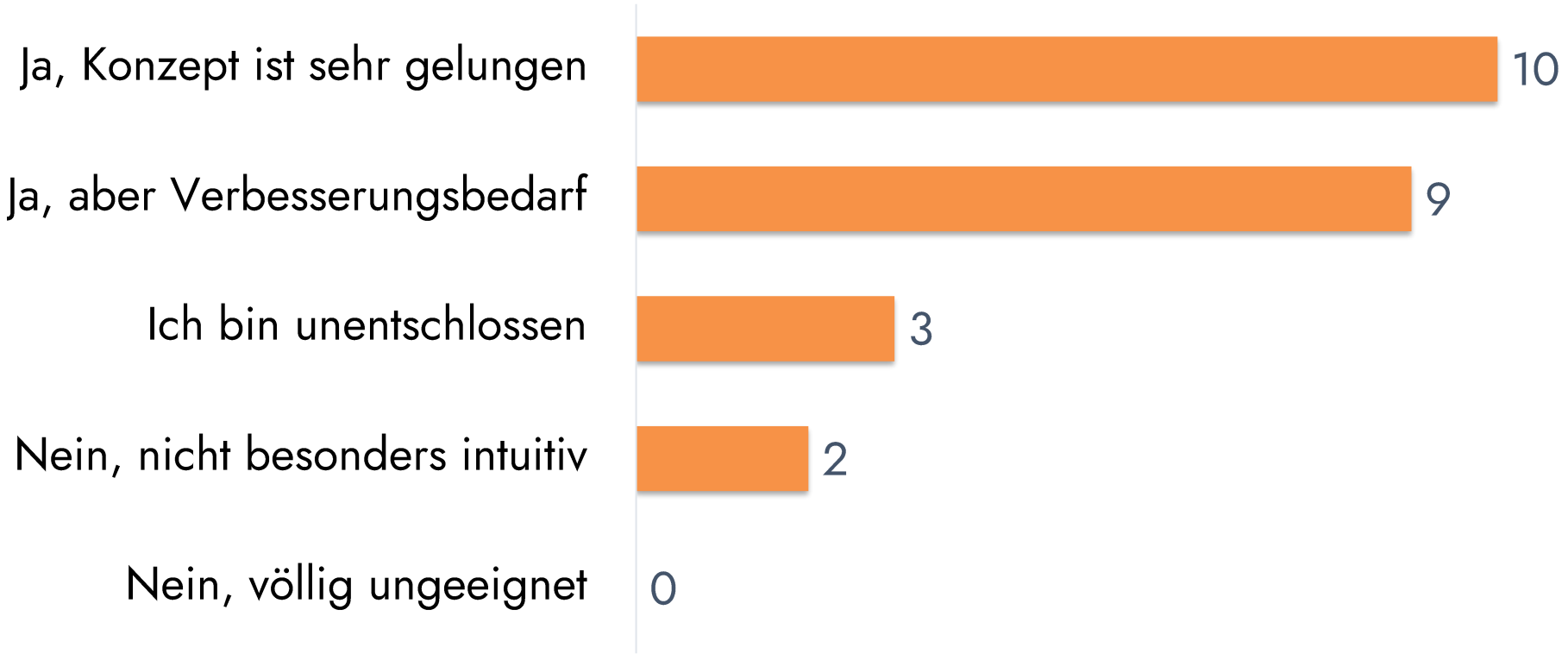
Bewerten Sie die folgenden Suchmöglichkeiten in Hinblick auf Ihr Suchverhalten.



Würden Sie eine Funktion nutzen, mit der Sie sich E-Mail-Verläufe mithilfe einer künstlichen Intelligenz zusammenfassen lassen könnten?



Denken Sie, dass sich ohne weitere Erklärungen in unserer grafischen Oberfläche zurecht finden würden?



Feedback zum Viewerkonzept

Konstruktiv

- *„Fehlende Filtermöglichkeiten“*
- *„Etwas unübersichtlich“*
- *„Erschließt sich mir nicht unmittelbar“*
- *„Mich würde noch interessieren, ob Text/Emails/Verläufe aus den Ergebnissen herauskopiert werden können?“*
- *„Balkendiagramm nimmt zu viel Platz auf der Seite ein“*
- *„Ein highlighten der gefundenen Stellen wäre auch interessant“*

Positiv

- *„Look & Feel sehr nahe an gängigen E-Mail-Clients“*
- *„Sieht auf den ersten Blick ganz übersichtlich aus.“*
- *„Intuitiv zu erfassen, da Aufbau sehr an die Ansichten üblicher Mail-programme erinnert.“*
- *„Die Timeline ist intuitiv gelungen“*
- *„Eine gute Mischung aus Mail-Postfach-Ästhetik und Bibliotheks-Nutzeroberfläche“*
- *„übersichtlich mit allen wichtigen Funktionen“*

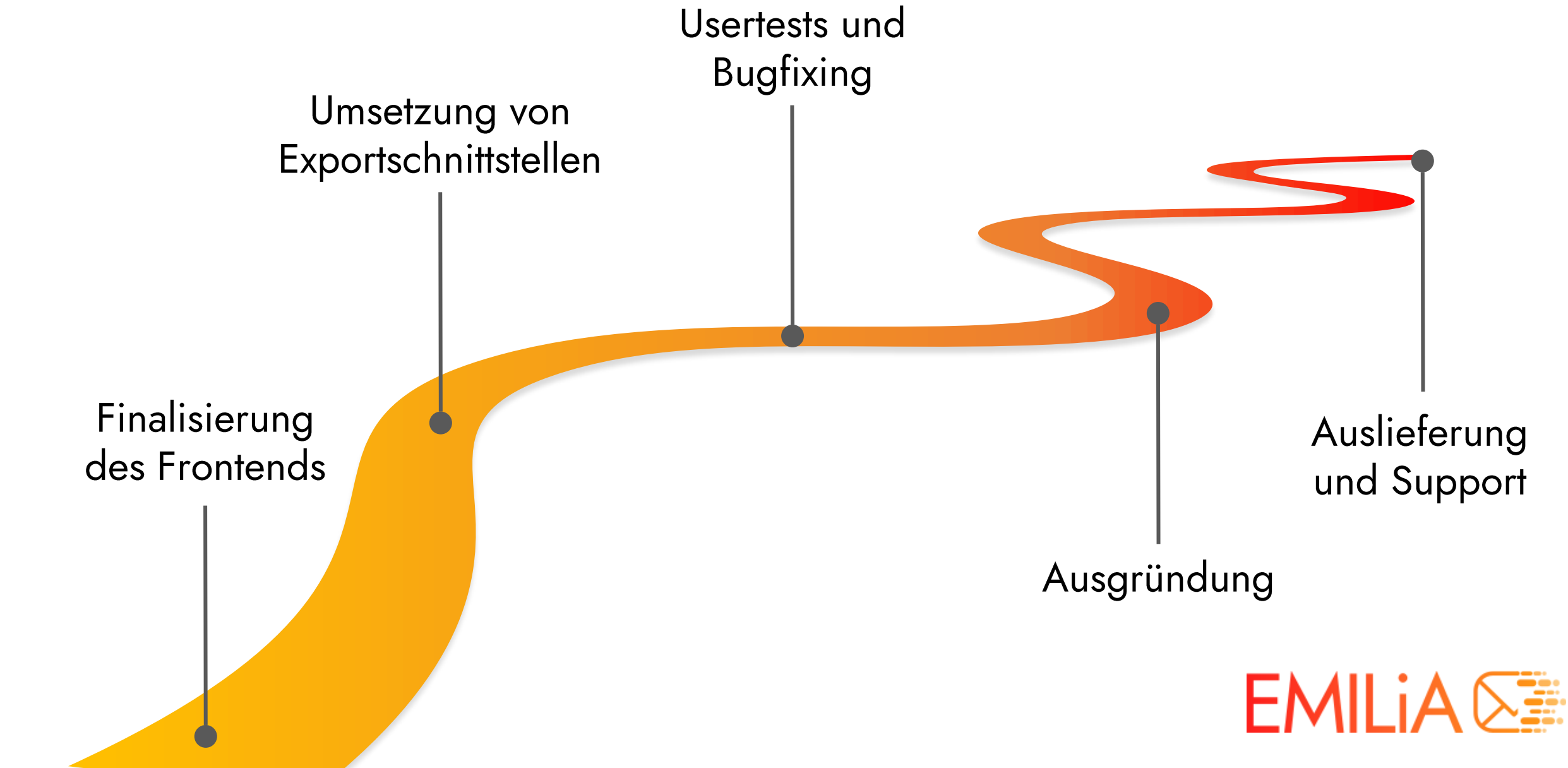
Umfrage zur Nutzung von E-Mails Quellen

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

Nutzungsszenarien

Bereitstellungs-Workflow

Wie geht es weiter?


Usertests und
Bugfixing
Umsetzung von
Exportschnittstellen
Finalisierung
des Frontends
Auslieferung
und Support
Ausgründung

Bleiben Sie auf dem Laufenden

- Um auf dem Laufenden zu bleiben, können Sie unseren Newsletter abonnieren und unsere Webinare besuchen.

EMILiA

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

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