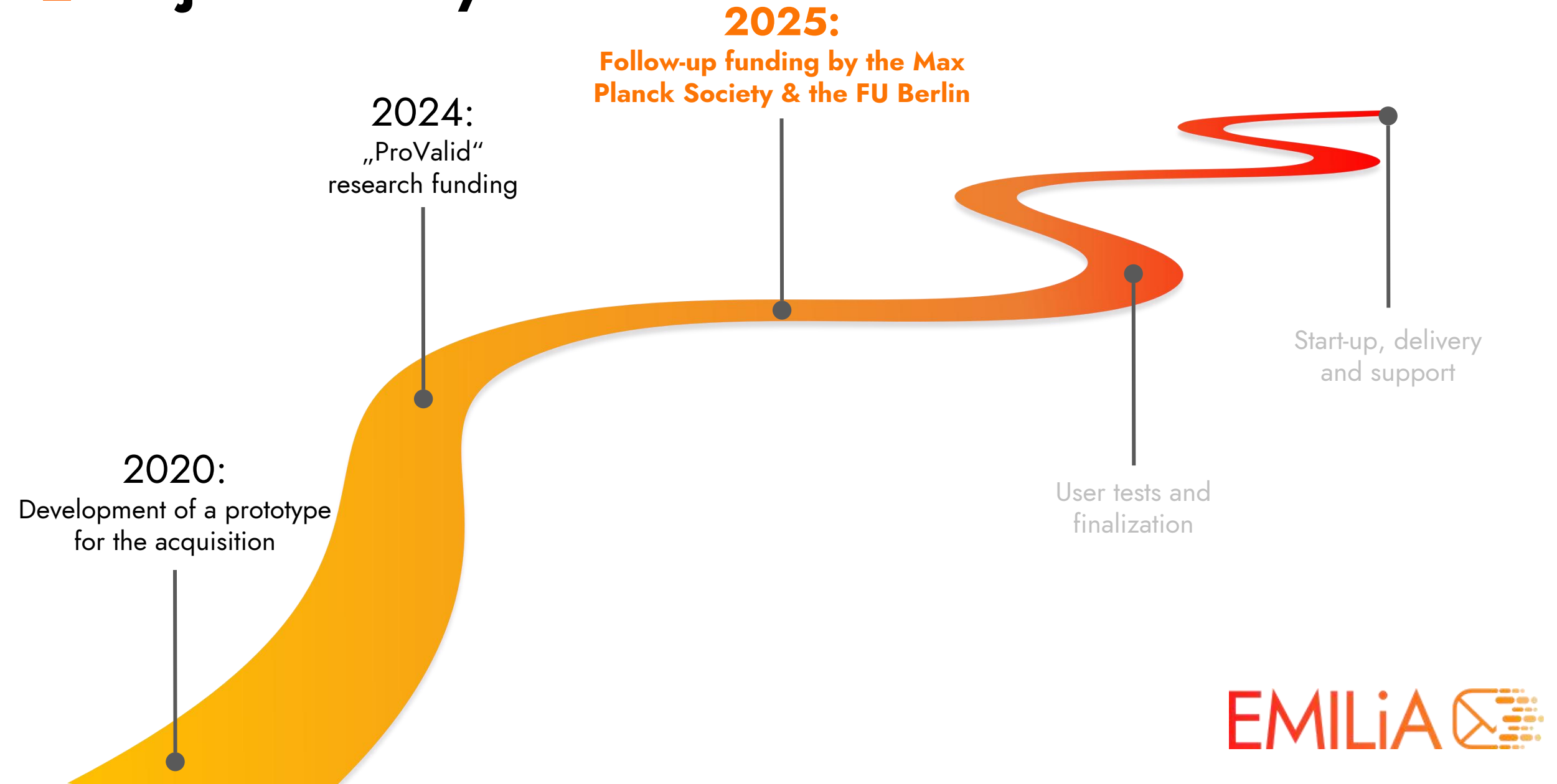




Making Email Archives Accessible through Intelligent Entity Recognition and Automated Anonymization

Nico Beyer & Felix Gericke

Project history



Context

E-mails are an ***integral part*** of most people's everyday professional and private lives.

Emails are a central communication medium

A large part of most mailboxes consists of ***spam, advertising***, or information that is only ***temporarily relevant***.

A lot of information without long term value

However, there are also ***emails of historical or legal relevance*** that should be preserved for the long term

An appraisal process is required

Proper ***appraisal, preservation, and analysis*** are only possible with the help of automated processes

Automation as an opportunity

Challenges



The email standard is vague in its specifications



Common email containers are not well suited for archiving purposes



Attachments in a wide variety of formats



Signed and encrypted emails

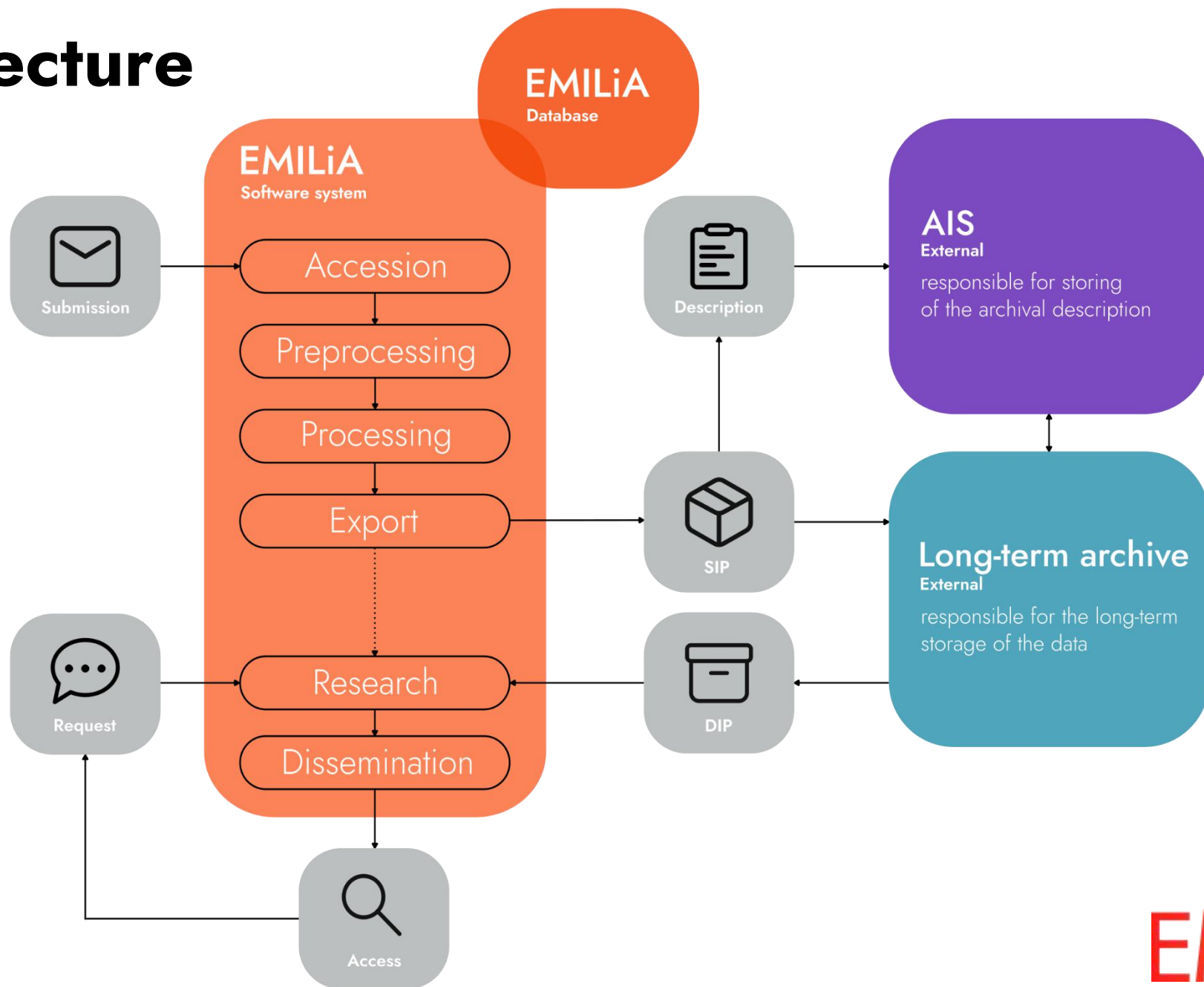


Personal data and copyright-relevant documents



Its hard to appraise which content has archival value

Architecture



Feature overview

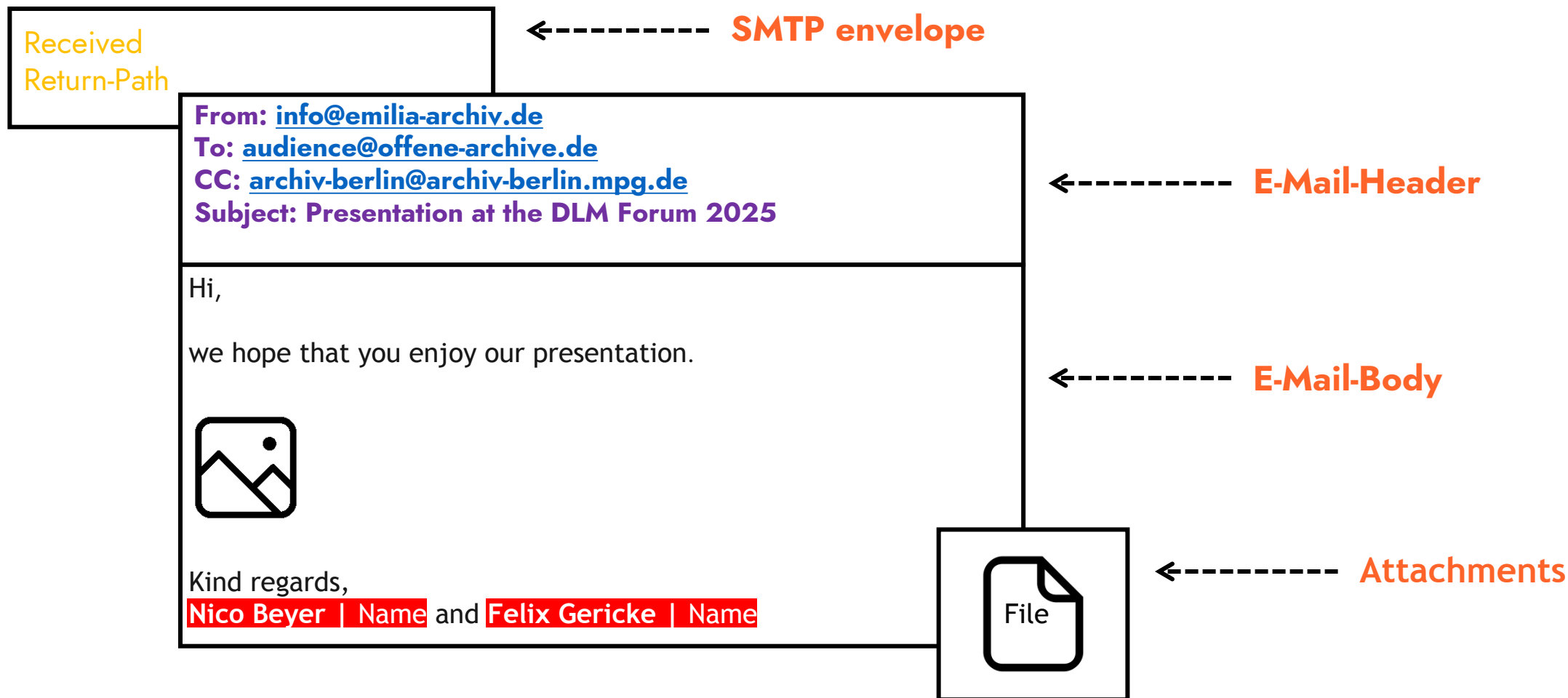
Transfer: Secure transfer, conversion, virus check, format recognition, integrity check

Appraisal & Description: Language recognition, spam- , duplicate and mailing list detection, topic extraction, identification of personal data, appraisal tool, bookmarks, sentiment analysis, contextualization

Research: Viewer, optional anonymization, full text search, filter functions, statistics and networks, summary of threads with LLM

Export: Structured BagIt container with emails, attachments and additional metadata or MBOX

Automated Entity Recognition



Live Demonstration

Thank you for your attention!

E-Mail: info@emilia-archiv.de

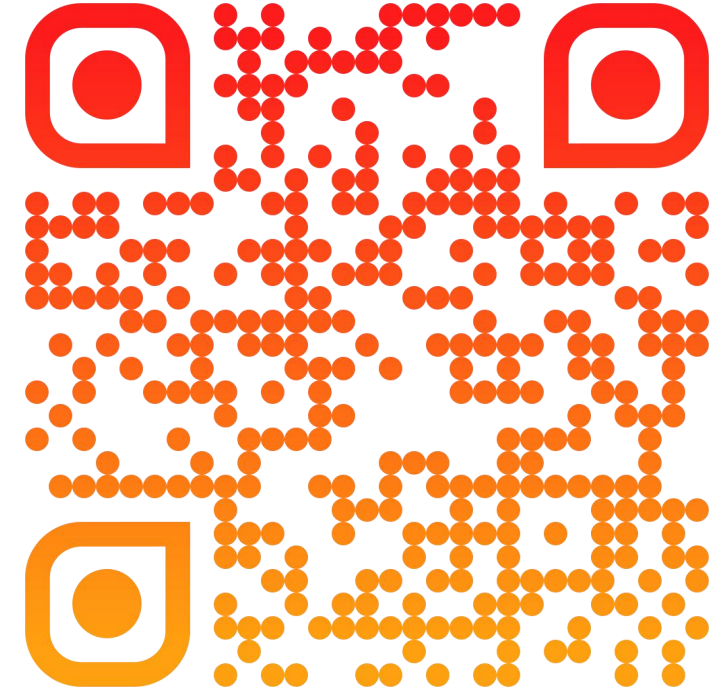
Telephone: +49 30 841 337 15

Freie Universität Berlin

Fachbereich Mathematik und Informatik

Takustraße 9

14195 Berlin-Dahlem



www.emilia-archiv.de