Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

STAATSBIBLIOTHEK ZU BERLIN – PK

# WORKSHOP SEMI-AUTOMATISCHE SACHERSCHLIEßUNG @ SBB

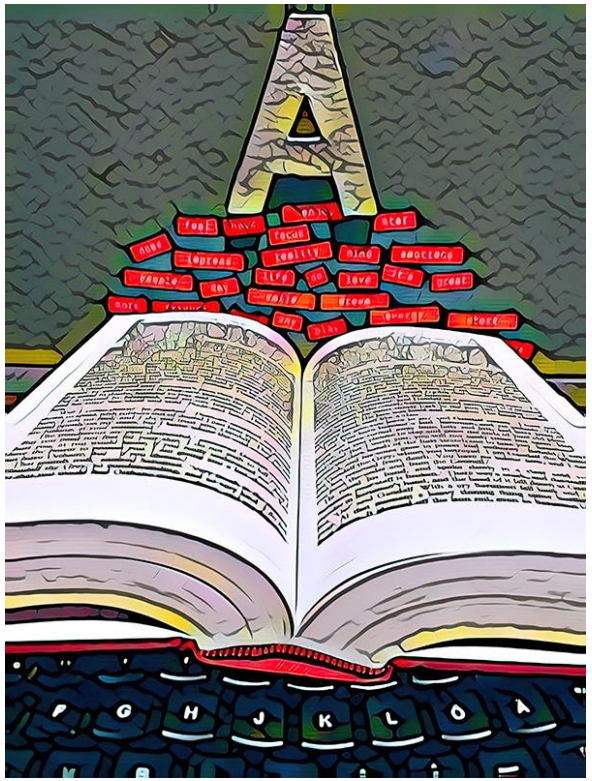
## EINFÜHRUNG IN DAS PROJEKT „MENSCH.MASCHINE.KULTUR“

Clemens Neudecker

# Was ist Künstliche Intelligenz?



- Der Begriff “Künstliche Intelligenz“ wird in Expertenkreisen als durchaus problematisch angesehen und sollte besser vermieden werden, da er einen Anthropomorphismus darstellt - als ob es sich hierbei um eine dem Menschen vergleichbare Intelligenz handelt. Dies ist aber absolut nicht der Fall.
- Besser wäre es entweder allgemein von “maschinellem Lernen” bzw. „deep learning“ zu sprechen oder spezifischer von “stochastischen Vorhersagemodellen” (provokanter „stochastic parrots“, vgl. Bender et al. 2021)
- Grundsätzlich gilt dabei: aus möglichst vielen repräsentativen Ausgangsdaten (Beispielen) werden Wahrscheinlichkeitsmodelle trainiert, um diese dann auf weitere Daten anwenden zu können.
- Die Qualität eines Modells (der “KI”) hängt also maßgeblich davon ab, wie umfangreich, qualitativ hochwertig und vielfältig die zum Training verwendeten Ausgangsdaten sind.



Teresa Berndtsson / Better Images of AI / Letter Word Text Taxonomy / CC-BY 4.0

- Um das „Trainieren“ einer KI / eines Modells zu veranschaulichen können zwei einfache Beispiele dienen:

- Bildklassifikation:
  Eine KI / ein Modell soll trainiert werden um Äpfel von Orangen zu unterscheiden. Der KI werden dazu so lange verschiedene Bilder von Äpfeln und Orangen gezeigt, bis die KI / das Modell für ein noch nicht gesehenes Bild selbst korrekt entscheiden kann, ob es sich dabei um einen Apfel oder eine Orange handelt.

- Sprachmodelle (z.B. ChatGPT):
  Eine KI / ein Modell soll trainiert werden um Texte zu bestimmten Themen zu verfassen oder Antworten auf Fragen zu geben. Die KI bekommt dazu sehr viele Texte gezeigt, in denen einzelne Wörter „maskiert“ bzw. ausgeblendet werden, also etwa: „Menschen gehen gerne in Bibliotheken um dort [MASK] zu lesen“.
  Für die Frage: „Warum gehen Menschen gerne in Bibliotheken?“ macht das Modell dann eine Vorhersage, welches Wort basierend auf den Trainingsdaten an Stelle von [MASK] am wahrscheinlichsten stehen könnte, hier also z.B. „Menschen gehen gerne in Bibliotheken um dort Bücher zu lesen.“

- Das Projekt „Mensch.Maschine.Kultur – Künstliche Intelligenz für das digitale kulturelle Erbe“ (MMK) wird von BKM mit rund €1.5 Mio. im Rahmen der Eckpunktestrategie KI des Bundes gefördert
- Projektdauer: 36 Monate, von 1. Juli 2022 bis 30. Juni 2025; Projektwebseite: https://mmk.sbb.berlin
- Aufgrund des Umfangs und der Vielfalt der Aufgaben ist das Projekt in 4 Teilprojekte strukturiert:
  - TP1: Intelligente Verfahren für die generische Dokumentanalyse (Dr. Vahid Rezanezhad, Mike Gerber)
  - TP2: Bildanalyseinstrumente zur Erschließung des digitalen kulturellen Erbes (Dr. Kai Labusch, Irina Dumitriu)
  - TP3: KI-unterstützte Inhaltsanalyse und Sacherschließung (Sophie Schneider, Wolfgang Seifert)
  - TP4: Datenbereitstellung und Kuratierung für KI (Dr. Jörg Lehmann)

- TP3 ist wiederum unterteilt in zwei Arbeitspakete
  - *AP 3.1 Semi-automatisierte KI-Verfahren für die Sacherschließung* und
  - *AP 3.2 Voll-automatisierte KI-Verfahren für die Discovery*
- Ziele von AP3.1
  - Erhebung von Anforderungen/Use Cases (bis ca. Ende 2023)
  - Training & Bereitstellung von Modellen für die Generierung von Schlagwörtern auf der Grundlage ausgewählter Klassifikationssysteme (2024)
  - Integration in den Digitalen Assistenten DA-3 und bestehende Workflows (2025)

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

STAATSBIBLIOTHEK ZU BERLIN – PK

# WORKSHOP SEMI-AUTOMATISCHE SACHERSCHLIEßUNG @ SBB

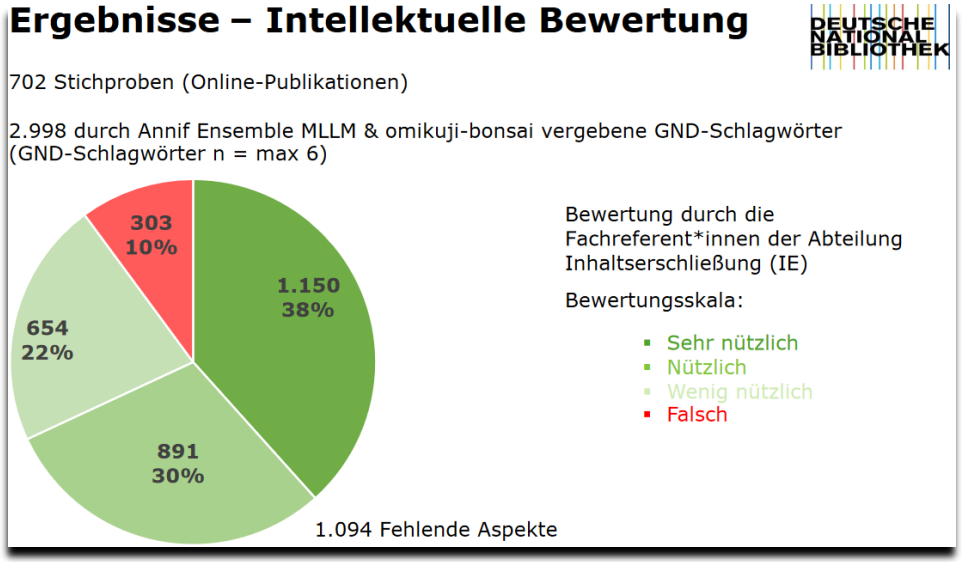
## STAND DER TECHNIK, TRENDS UND PROJEKTE

Wolfgang Seifert

- Deutsche Nationalbibliothek: EMa und Projekt „Automatisches Erschließungssystem“
- ZBW – Leibniz-Informationszentrum Wirtschaft: AutoSE
- Technische Informationsbibliothek: LinSearch
- Bayerische Staatsbibliothek: Yewno

- 2009-15: PETRUS-Projekt
- Netzpublikationen seit 2010 nicht mehr intellektuell, Deutschsprachige seit 2012 nach Sachgruppen, seit 2014 verbal automatisiert erschlossen
- 2019-22: EMa-Projekt (EMa danach verstetigt)
- Seit 2017: ToC-basierte Verschlagwortung von Printpublikationen
- Seit 2018: Automatisierte Verschlagwortung englischsprachiger Netzpublikationen
- Seit 2020: Automatisierte Verschlagwortung
- Seit 2022: Kinder- und Jugendliteratur
- Seit 2021: KI-Projekt „Automatisches Erschließungssystem“




Elisabeth Mödden / Automatische Inhaltserschließung in der Deutschen Nationalbibliothek / Folie 16

- BSB: Proprietäre Software „Yewno“

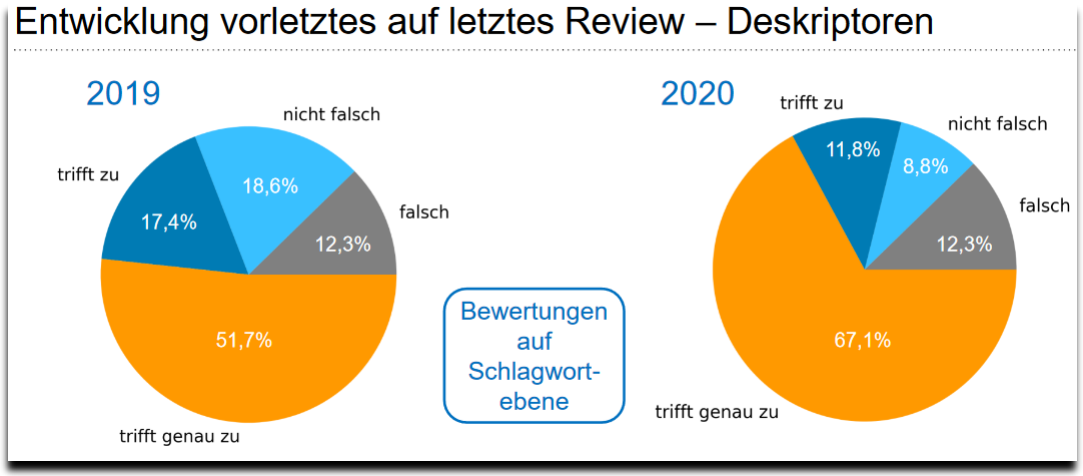
- TIB: Algorithmus „LinSearch“ innerhalb von Annif, um Fachfacetten für das TIB-Portal zu gewinnen (Material: vorhandene Inhaltsmetadaten, Zuordnung von Zeitschriften oder Datenlieferanten, Linguistische Analyse)

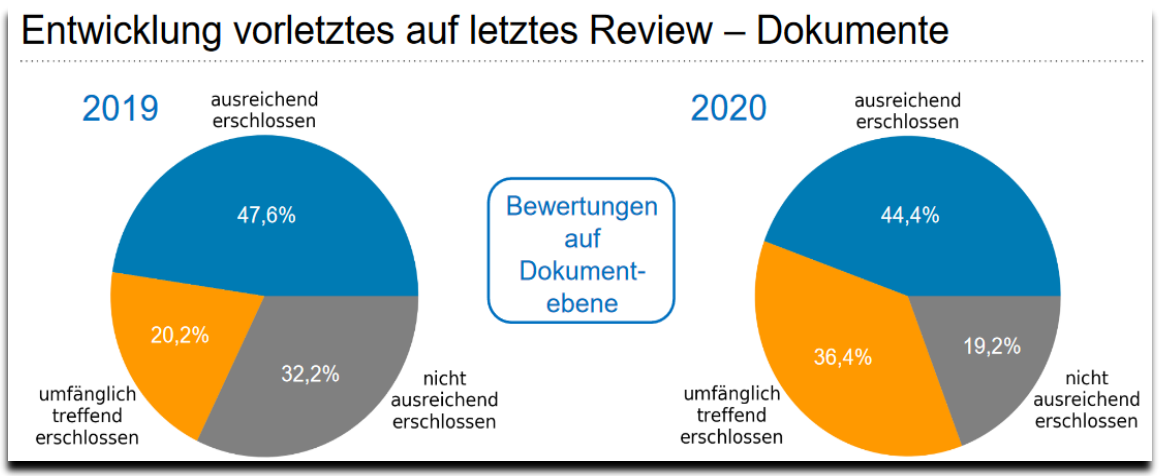
# ZBW: AutoSE und der „Human in the Loop“



- 2002-11 explorative Projekte mit externen Anbietern
- 2014-18 eigene angewandte Forschung und Open-Source Entwicklung
- seit 2019: Verstetigung von AutoSE als Dienst, Einbeziehung von Annif
- Sehr differenziertes Qualitätsmanagement
- Besondere Betonung des „Human in the Loop“

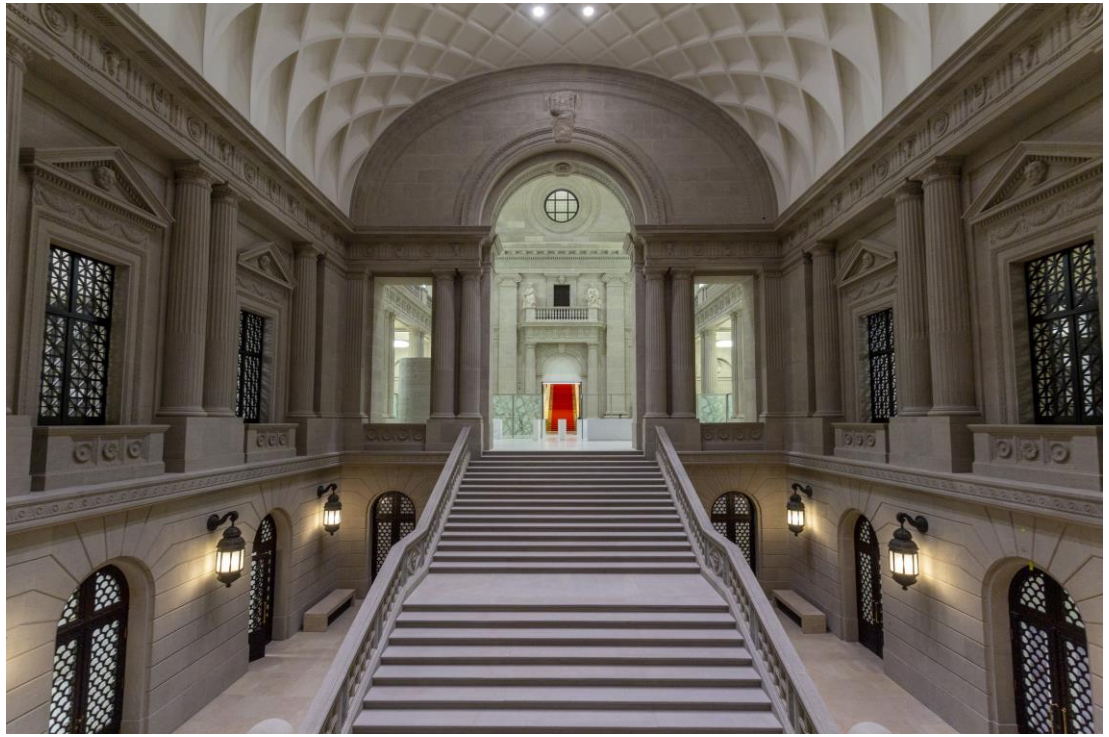

Anna Kasprzik / AutoSE: Automatisierung der Inhaltserschließung mit Machine-Learning-Methoden an der ZBW / Folie 5 / CC-BY




Anna Kasprzik und Christopher Bartz / AutoSE: Automatisierung der Inhaltserschließung mit Machine-Learning-Methoden an der ZBW / Folie 11 (links) und 12 (rechts) / CC-BY

Eine enge Verzahnung zwischen domain experts und technischer Seite wird von allen PraktikerInnen als maßgeblicher Erfolgsfaktor angesehen.



SBB-PK / Ralf Stockmann

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

STAATSBIBLIOTHEK ZU BERLIN – PK

# WORKSHOP SEMI-AUTOMATISCHE SACHERSCHLIEßUNG @ SBB

## STAND TP3 & DEMO ANNIF

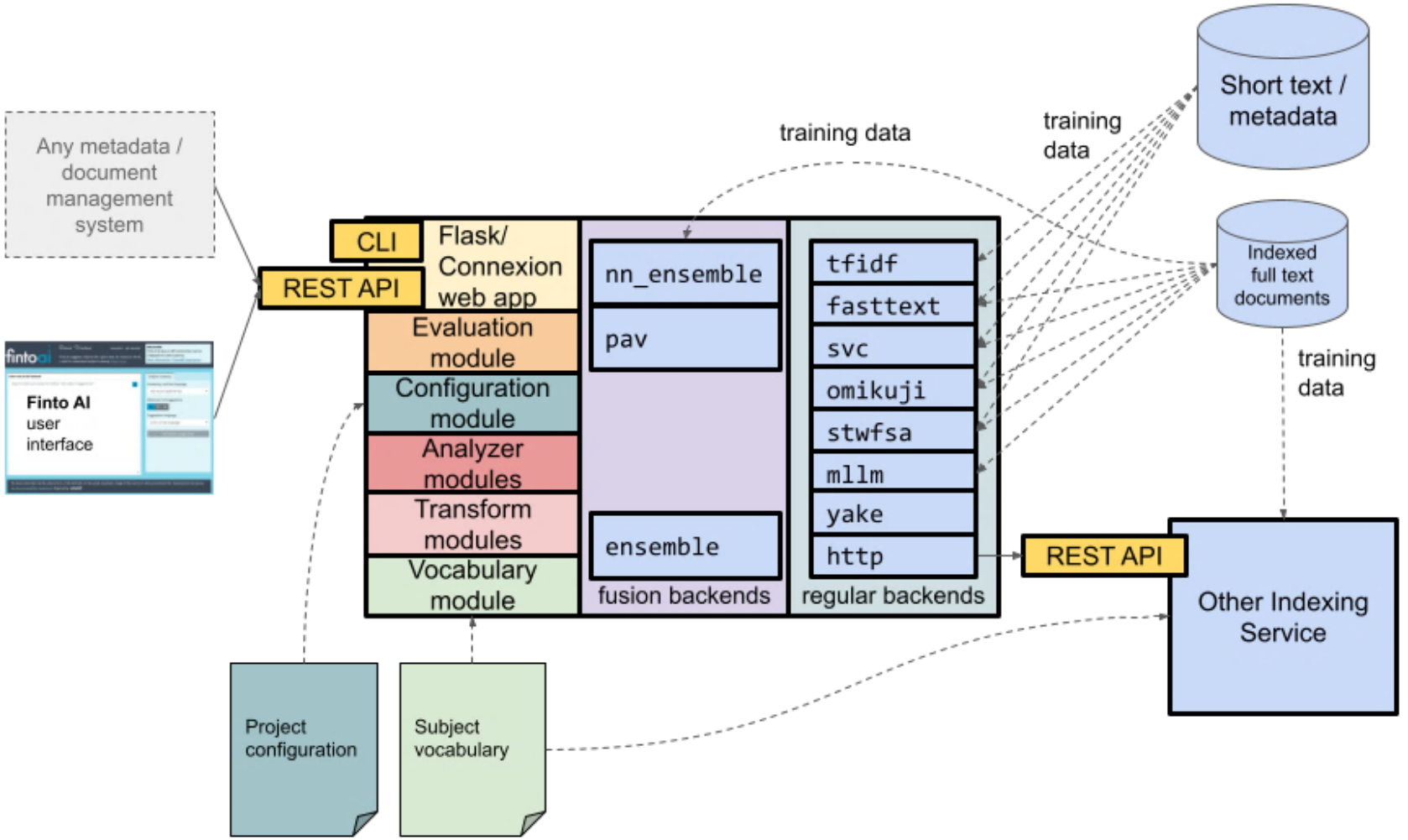
Sophie Schneider

- TP3 ist unterteilt in zwei Arbeitspakete
  - *AP 3.1 Semi-automatisierte KI-Verfahren für die Sacherschließung* und
  - *AP 3.2 Voll-automatisierte KI-Verfahren für die Discovery*
- Ziele von AP3.1
  - Erhebung von Anforderungen/Use Cases (bis ca. Ende 2023)
  - Training & Bereitstellung von Modellen für die Generierung von Schlagwörtern auf der Grundlage ausgewählter Klassifikationssysteme (2024)
  - Integration in den Digitalen Assistenten DA-3 und bestehende Workflows (2025)

- Aktueller Stand
  - Aufsetzen einer eigenen Annif-Testinstanz: http://annif.b-lx0053.sbb.spk-berlin.de/
  - Vorbereitend zur Erstellung der Trainingsdaten: Skript für den Abruf von Katalogdaten via SRU-Schnittstelle + geplante Integration der Konversion in entsprechendes Datenformat
  - erster Erfahrungsaustausch (SBB intern, DNB, weitere stehen an) und Sammlung von Ideen

- Was ist Annif?
  - „an open source toolkit for automated subject indexing and classification“ (Suominen et al. 2022, S. 265)
  - modular und universell einsetzbar (z.B. können verschiedene Sprachen und Vokabulare hinzugefügt und genutzt werden)
  - umfassend dokumentiert (Annif Wiki, Annif Tutorial), internationale Community von Nutzer:innen (Mailingliste Annif Users)




Osma Suominen / Annif Architecture

Bitte diskutieren Sie in Gruppen die folgenden Fragen und notieren Sie ihre Antworten:

1. Welche Hoffnungen und Erwartungen haben Sie für eine automatische Unterstützung bei der SE?
2. Was sind die Besonderheiten und Herausforderungen bei der SE speziell für Ihren Fach- bzw. Arbeitsbereich?