Die Erschließungsmaschine (EMa) – Ergebnisse und Perspektiven der automatischen Inhaltserschließung an der Deutschen Nationalbibliothek

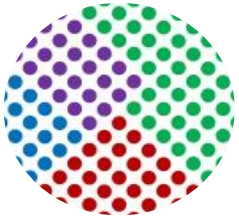
Sandro Uhlmann | Deutsche Nationalbibliothek

EMa – Erschließungsmaschine

DEUTSCHE NATIONAL BIBLIOTHEK

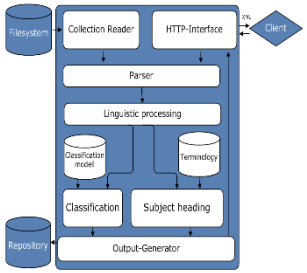
- Projekt EMa: Ausgangslage, Ziele, Vorgehen
- Das Toolkit Annif
- Ergebnisse Klassifizierung & Indexierung
- Automatische Erschließung mit Annif als Service
- Ausblick

Automatische Inhaltserschließung in der DNB



Automatische Klassifizierung von Online- und ausgewählten Printpublikationen mit DDC-Sachgruppen und DDC-Kurznotationen (Support Vector Machine, Assoziative Verfahren)

Automatische Indexierung von Online- und ausgewählten Printpublikationen anhand der normierten Terminologien GND und LCSH (Text Mining, Lexikalische Verfahren)



Software:

Averbis Extraction Platform (AEP)

Ausgangslage

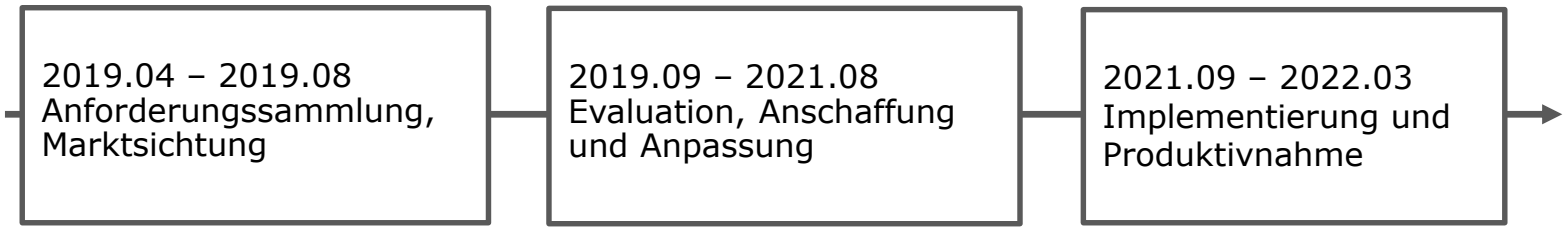
- keine Weiterentwicklung der Erschließungssoftware DNB-AEP* vom Anbieter Averbis
- Einsatz des „Altsystems“ noch für einen Zeitraum von etwa 3 bis 5 Jahren möglich
- eine Neuentwicklung für das gesamte Erschließungssystem, bislang bestehend aus einem Web-Service zur Kommunkiation und Steuerung sowie der Averbis-Software muss geplant und umgesetzt werden

Erschließungsmaschine – Ziele & Organisation

DEUTSCHE NATIONAL BIBLIOTHEK

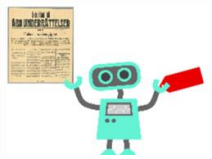
- Aufbau eines modularen Systems zur automatischen Inhaltserschließung
- Ablösung der bislang eingesetzten Averbis-Software (Altsystem)

Stichworte (Auswahl): #Klassifizierung mit unterschiedlichen Klassifikationssystemen #Indexierung mit kontrolliertem Vokabular #Sprachenidentifizierung #Modularität #Erweiterung um neue Funktionen oder Verfahren #kontinuierliche Verbesserung der Erschließungsergebnisse u.a.



Evaluation

annif



DEUTSCHE NATIONAL BIBLIOTHEK

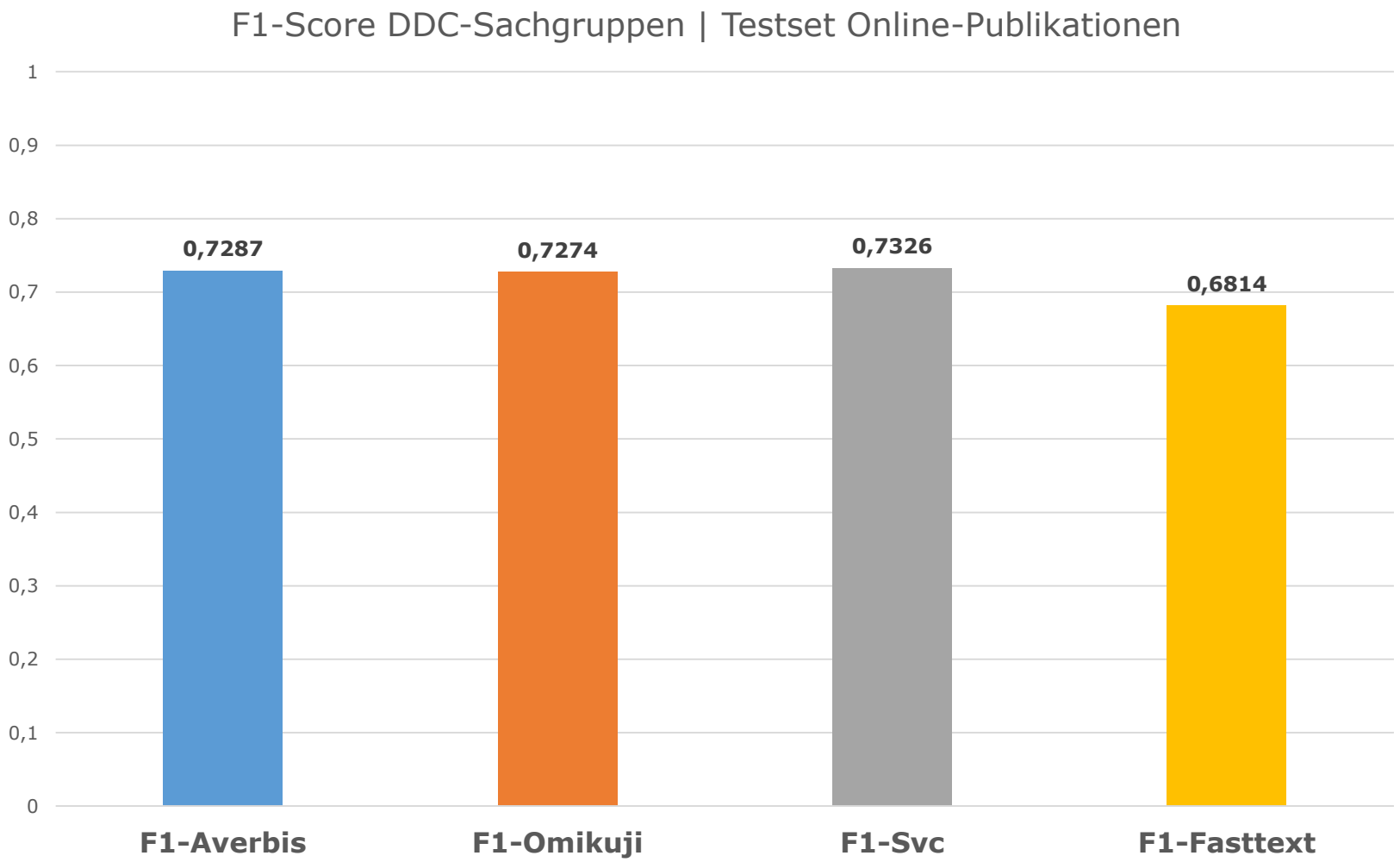
- entwickelt an der Finnischen Nationalbibliothek
- verwendet existierende Werkzeuge
  - zur Verarbeitung natürlicher Sprache
  - zum maschinellen Lernen
- ist multilingual
  - Einsatz des Natural Language Toolkit, NLTK
- kann jedes Vokabular verwenden
  - in SKOS oder einfachem TSV
- ist über Kommandozeile, Web UI und Rest API bedienbar
- ist Open Source und in Python implementiert

Suominen, Osma: Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly, 29 (1) 2019 , pp.1–25. http://doi.org/10.18352

Suominen, Osma; Inkinen, Juho; Lehtinen, Mona: Annif and Finto AI: Developing and Implementing Automated Subject Indexing. PrePrint accepted on 23 June 2021 for publication in JLIS.it https://urn.fi/URN:NBN:fi-fe2021080942632

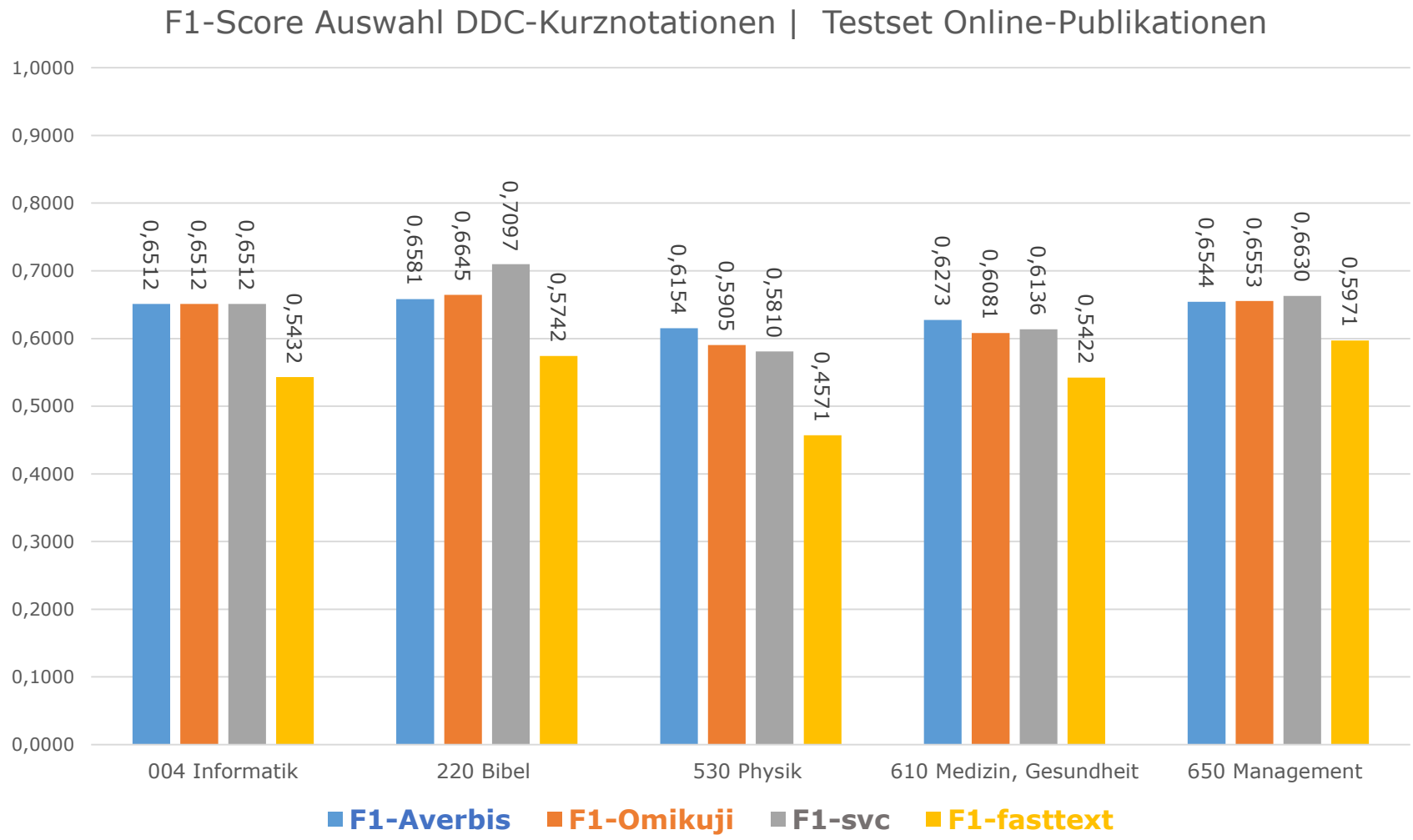
Ergebnisse - DDC-Sachgruppen

DEUTSCHE NATIONAL BIBLIOTHEK



Ergebnisse - DDC-Kurznotationen

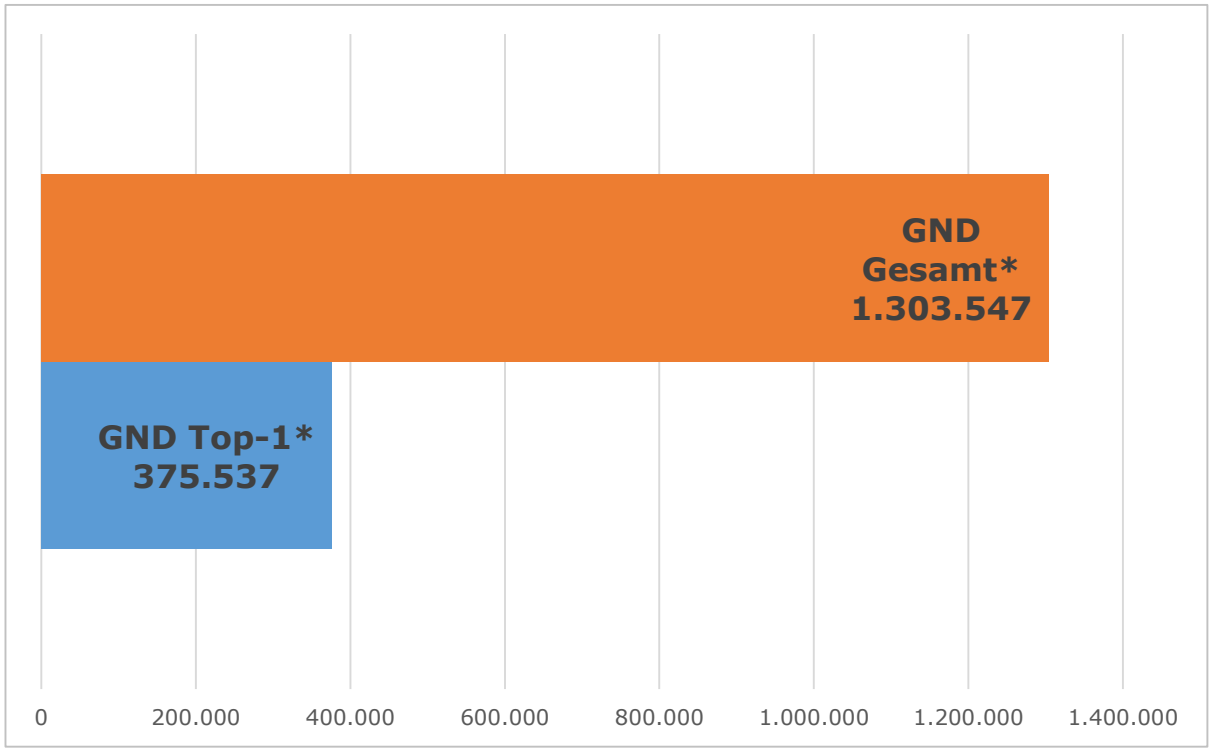
DEUTSCHE NATIONAL BIBLIOTHEK



Indexierung

DEUTSCHE NATIONAL BIBLIOTHEK

Schlagworte der Gemeinsamen Normdatei (GND)



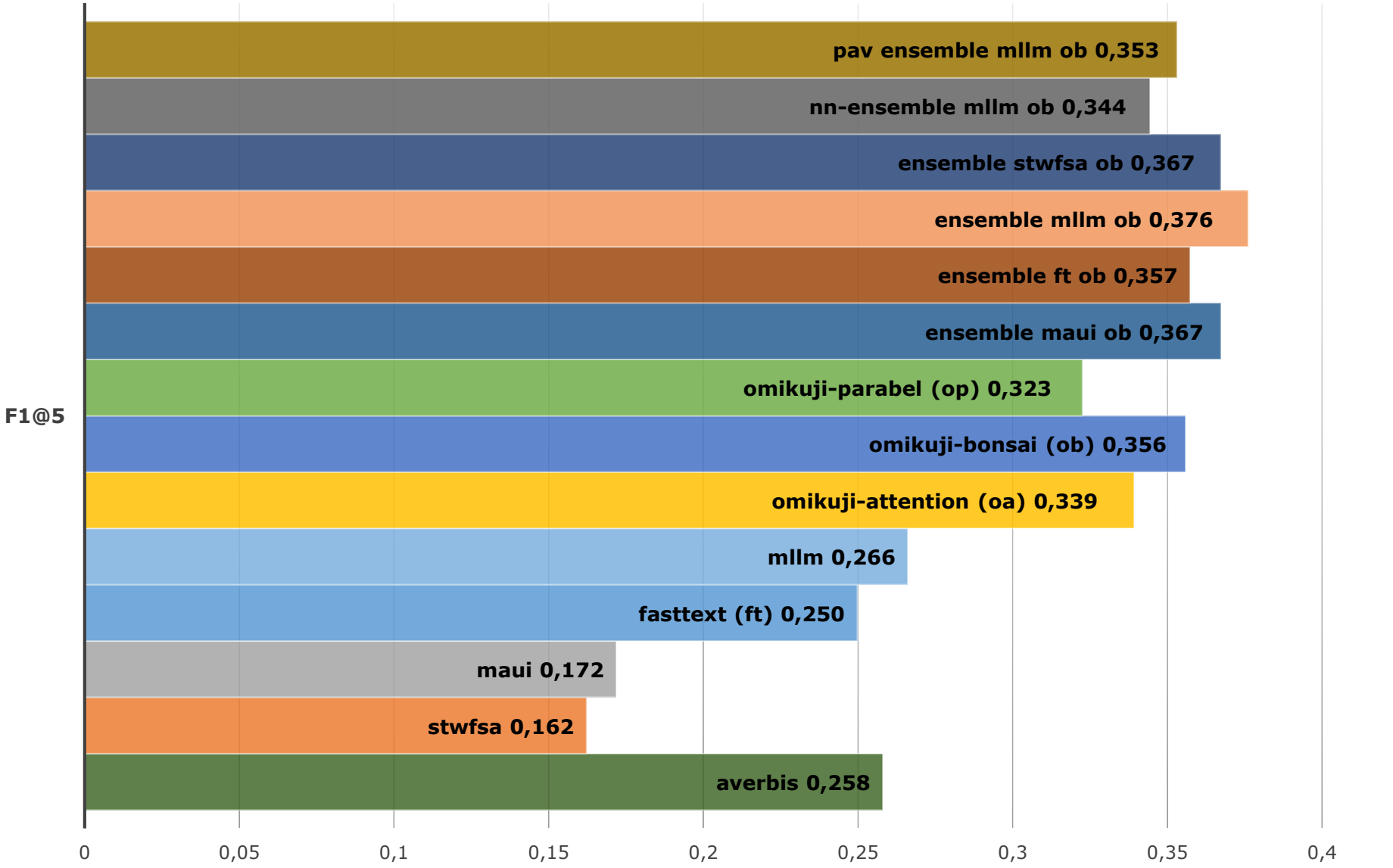
Trainingsmaterial, Deutsch:

- kein vollständiges Trainingsset für alle GND-Schlagworte
- nur 375.537 haben mind. einen Textobjekt
- 928.010 haben kein Textobjekt

*Katalogisierungslevel 1 oder z und aus dem Teilbestand s

Ergebnisse GND-Schlagworte

Testset 1.261 Online-Publikationen | F1-Score (n=5 Schlagworte)



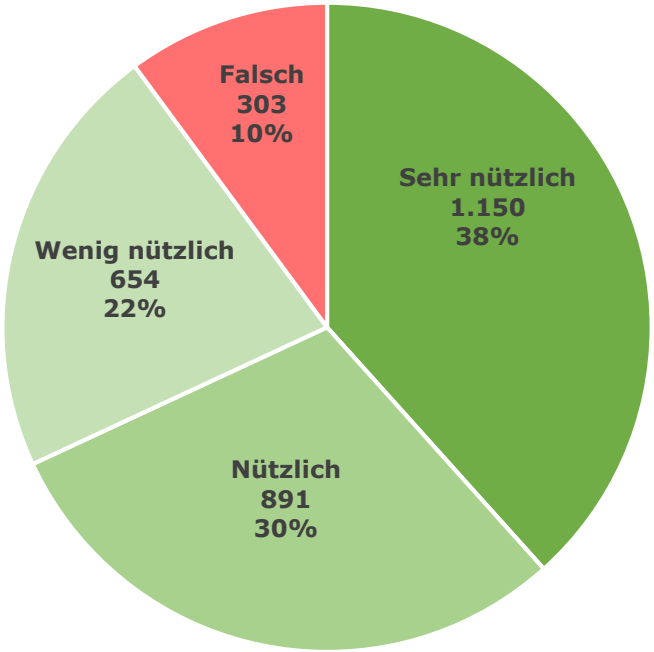
Annif – Intellektuelle Bewertung

Einzelbewertung

DEUTSCHE NATIONAL BIBLIOTHEK

702 Datensätze (Online-Publikationen)

2.998 durch ein Ensemble aus MLLM & omikuji-bonsai vergebene GND-Schlagworte (n = max 6 pro Publikation)



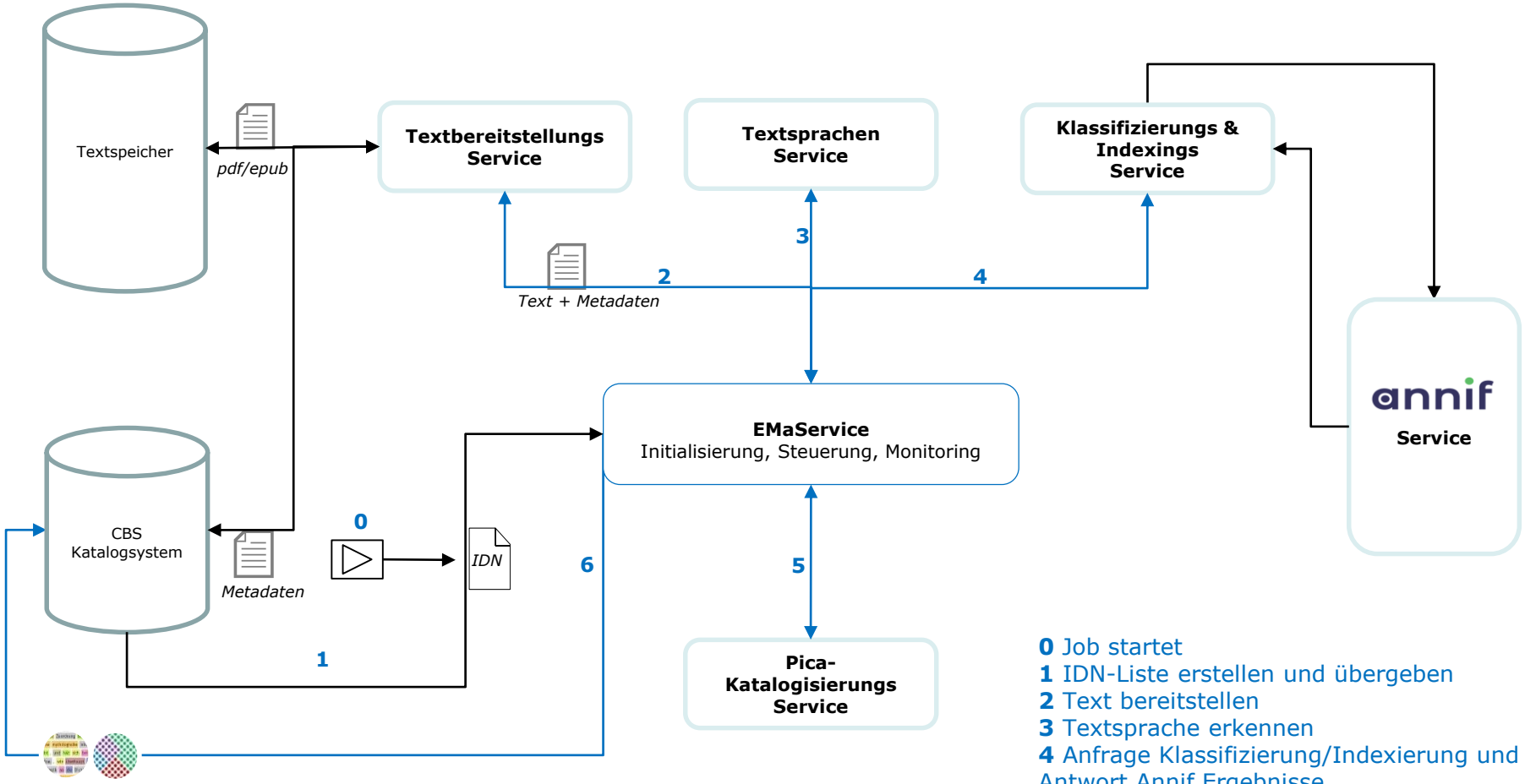
1.094 Fehlende Aspekte

Intellektuelle Bewertung durch die Indexierer* der Abteilung Inhaltserschließung

Bewertungsskala:

- Sehr nützlich
- Nützlich
- Wenig nützlich
- Falsch

Automatische Erschließung mit Annif als Service



0 Job startet
1 IDN-Liste erstellen und übergeben
2 Text bereitstellen
3 Textsprache erkennen
4 Anfrage Klassifizierung/Indexierung und Antwort Annif Ergebnisse
5 Konvertierung der Ergebnisse nach Pica+
6 Zurückschreiben der Ergebnisse

Ausblick

Bis März 2022:

- Test der Verarbeitungskette aller Services
- Produktivgang mind. mit automatischer Indexierung (GND-Schlagworte) für dt.-sprachige Publikationen

Ab April 2022:

- Überführung aller Themen in die Routine und Weiterentwicklung in agiler und iterativer Arbeitsweise
- Retirement des Altssystems werden für den Übergang aus dem Projektstatus in den Produktstatus im Mittelpunkt stehen

Forschungsprojekt

Die Beauftragte der Bundesregierung für Kultur und Medien

DEUTSCHE NATIONAL BIBLIOTHEK

Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit KI

- Förderung: Beauftragte der Bundesregierung für Kultur und Medien (BKM) auf der Grundlage der Nationalen KI-Strategie
- Laufzeit: 2021 – 2024 (3 Jahre)

Forschungsprojekt

Die Beauftragte der Bundesregierung für Kultur und Medien

DEUTSCHE NATIONAL BIBLIOTHEK

Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit KI

- Erforschung verschiedener methodischer Ansätze, um die Resultate der automatischen Erschließung mit dem Vokabular der GND weiter zu verbessern
- Ungelöste Herausforderungen angehen
- Neue Methoden und Algorithmen aus dem Bereich der KI auswählen, untersuchen und adaptieren
- Passende Werkzeuge für die Aufbereitung und Analyse der Texte und Daten entwickeln und zur (Nach-)Nutzung bereitstellen

DEUTSCHE NATIONAL BIBLIOTHEK

EMa – Erschließungsmaschine

EMa-Projektteam:

Erik Brangs, Frank Busse, Claudia Grote, Jan-Helge Jacobs, Matthias Nagelschmidt, Christoph Poley, Kirsten Tröller, Sandro Uhlmann, Nico Wagner

Danke für Ihre Aufmerksamkeit.

s.uhlmann@dnb.de

Tipp **pica-rs** - tool to work with bibliographic records encoded in PICA+ in a fast and efficient way
https://github.com/deutsche-nationalbibliothek/pica-rs