DEUTSCHE NATIONAL BIBLIOTHEK

Sandro Uhlmann

Automatische Inhaltserschließung an der Deutschen Nationalbibliothek (DNB)

BID22 8. BIBLIOTHEKS KONGRESS LEIPZIG 2022

Session Indexieren und Sortieren

2022-05-31

Automatische Inhaltserschließung in der DNB

DEUTSCHE NATIONAL BIBLIOTHEK

- Hintergrund
- Projekt Erschließungsmaschine (EMa)
- Das Toolkit Annif
- Automatische Indexierung deutschsprachiger Publikationen mit Schlagwörtern der Gemeinsamen Normdatei (GND)
- EMa mit Annif als Service im produktiven Einsatz
- Ausblick

Automatische Inhaltserschließung in der DNB

Automatische Klassifizierung von Online- und ausgewählten Printpublikationen mit DDC-Sachgruppen und DDC-Kurznotationen

Produktiv seit 2012

Assoziative Verfahren

Automatische Indexierung von Online- und ausgewählten Printpublikationen anhand der normierten Terminologien GND und LCSH

Produktiv seit 2014

Lexikalische Verfahren

Assoziative Verfahren

Junger, Ulrike; Scholze, Frank: Neue Wege und Qualitäten – Die Inhaltserschließungspolitik der Deutschen Nationalbibliothek. In: Qualität in der Inhaltserschließung, herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl und Hans Schürmann, Berlin, Boston: De Gruyter Saur, 2021, S. 55–70. https://doi.org/10.1515/9783110691597-004

Lexikalische vs. Assoziative Verfahren

Lexikalische Verfahren

Abgleich zwischen relevanten Termen im Text und Termen im kontrollierten Vokabular

„Die **Ästhetische Theorie** ist ein posthum erschienenes Werk des Philosophen **Theodor W. Adorno**."

http://d-nb.info/gnd/4122759-1
Ästhetische Theorie

http://d-nb.info/gnd/118500775
Adorno, Theodor W.

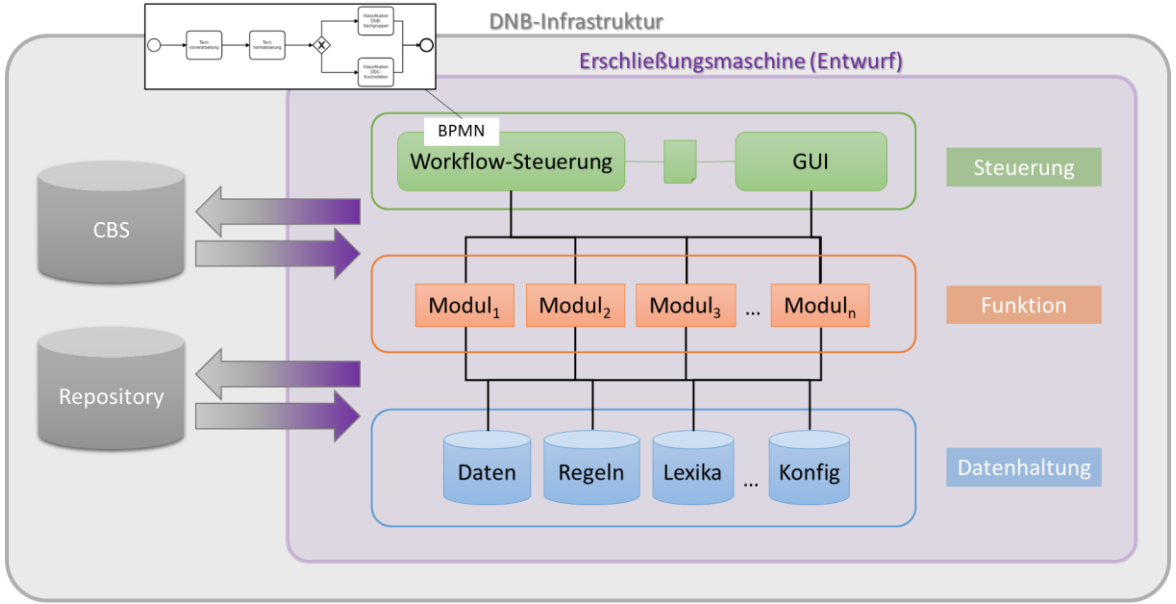
Assoziative Verfahren

Lernen auf der Basis von Trainingsdokumenten: Deskriptoren eines kontrollierten Vokabulars werden mit Begriffen im Dokument assoziiert

Projekt Erschließungsmaschine (EMa) 2019 - 2022

DEUTSCHE NATIONAL BIBLIOTHEK

- Aufbau eines modularen Systems zur automatischen Inhaltserschließung



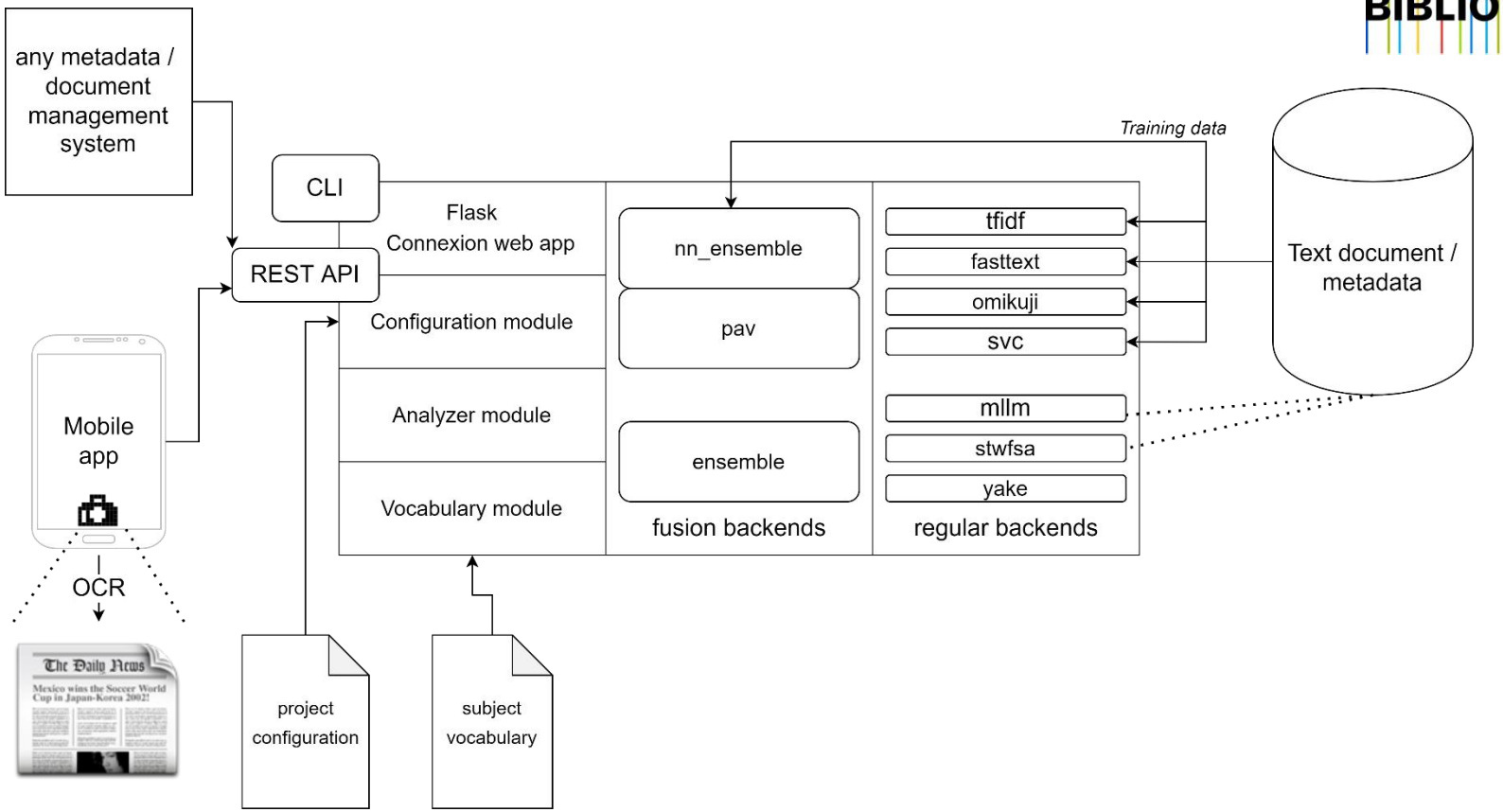
- Ablösung der bislang eingesetzten Software (Altsystem)

Tool for automated subject indexing and classification

- entwickelt an der Finnischen Nationalbibliothek
- verwendet existierende Werkzeuge zur Verarbeitung natürlicher Sprache und 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
- Verfahren können allein oder als Ensemble eingesetzt werden

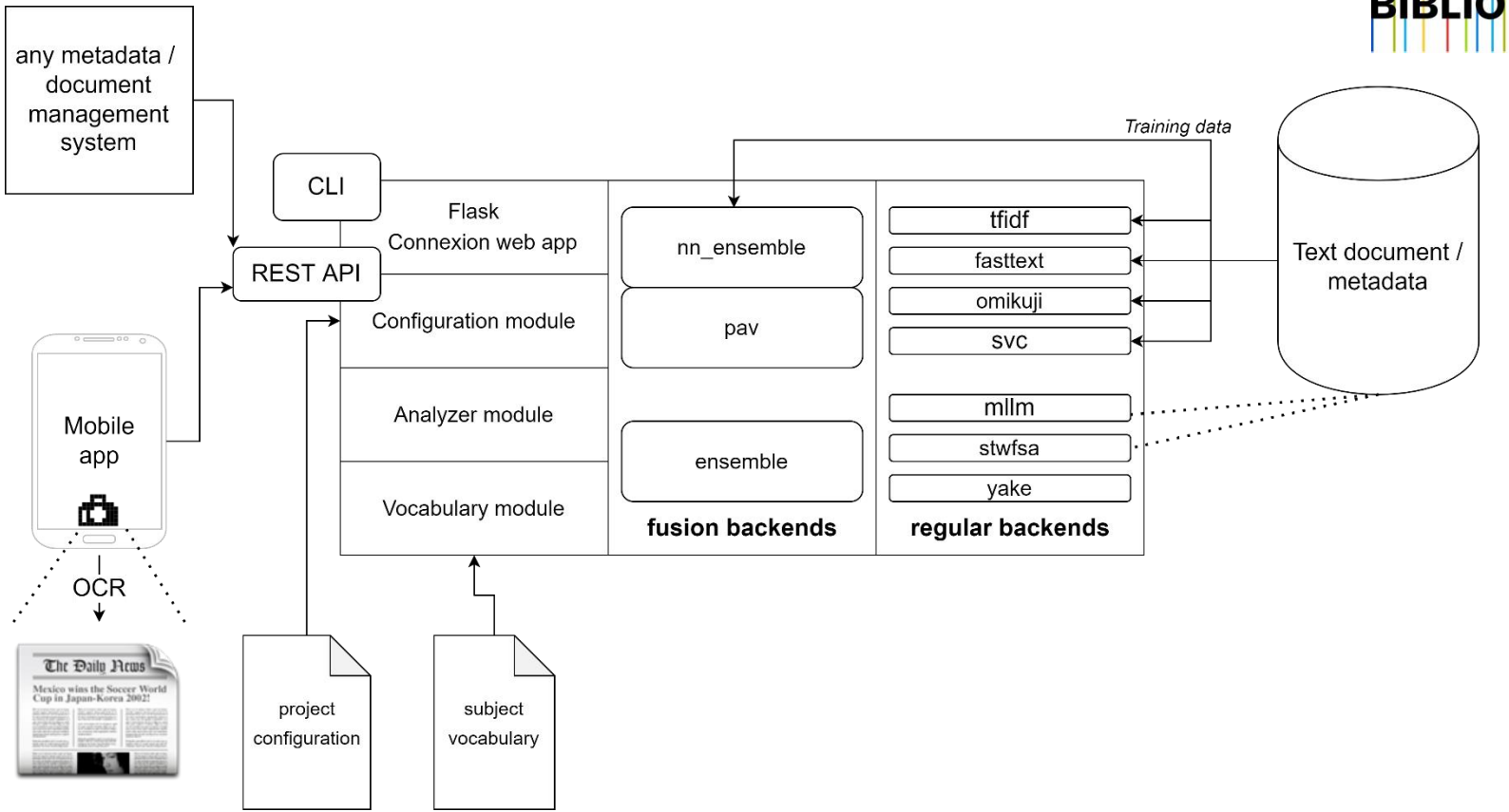
Suominen, Osma; Inkinen, Juho; Lehtinen, Mona: Annif and Finto AI: Developing and Implementing Automated Subject Indexing. JLIS.It, 13(1), 265–282. https://doi.org/10.4403/jlis.it-12740
Github https://github.com/NatLibFi/Annif

annif - Architektur



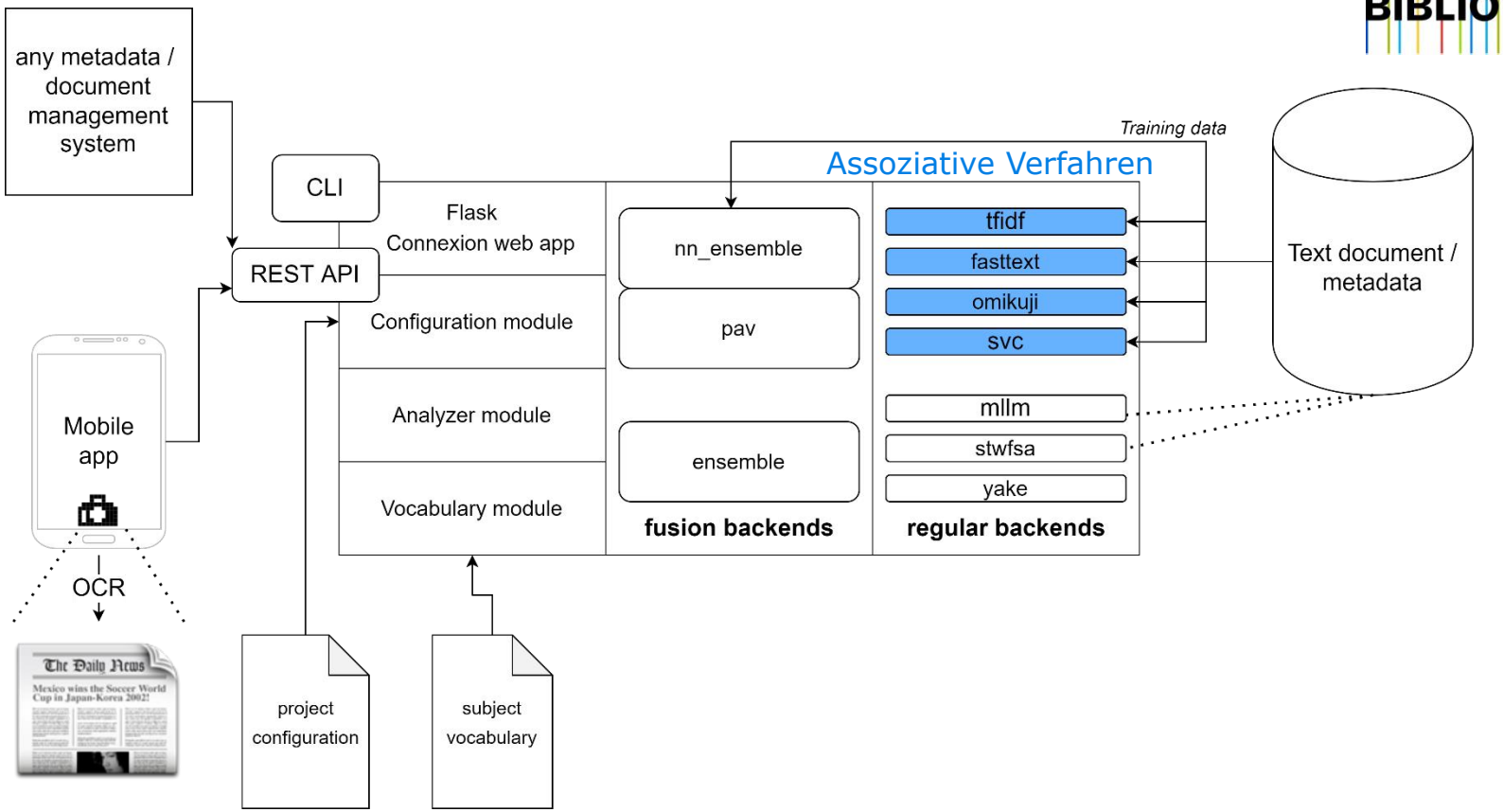
Adaption nach: https://github.com/NatLibFi/Annif/wiki/images/annif-architecture.png

annif - Architektur



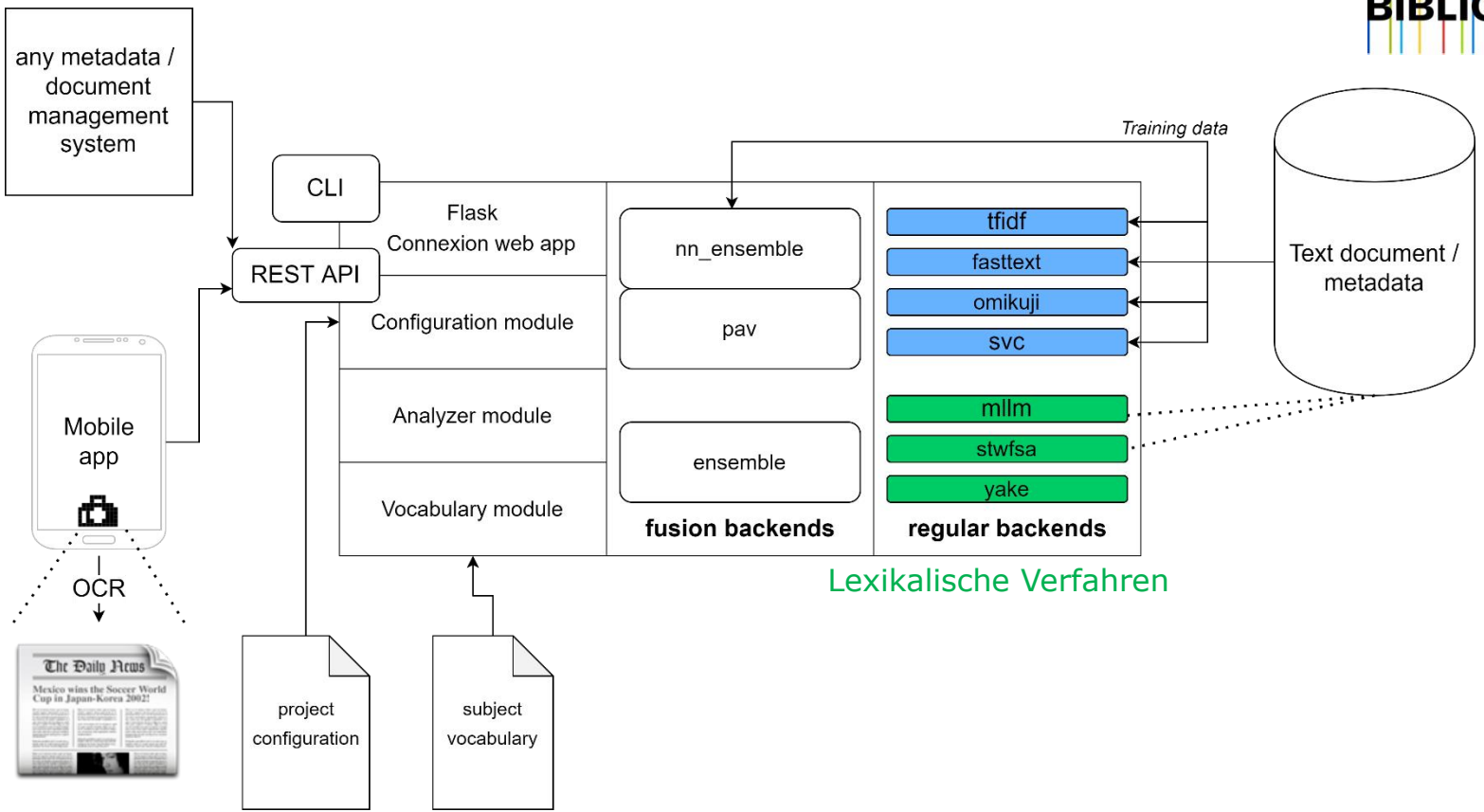
Adaption nach: https://github.com/NatLibFi/Annif/wiki/images/annif-architecture.png

annif - Architektur



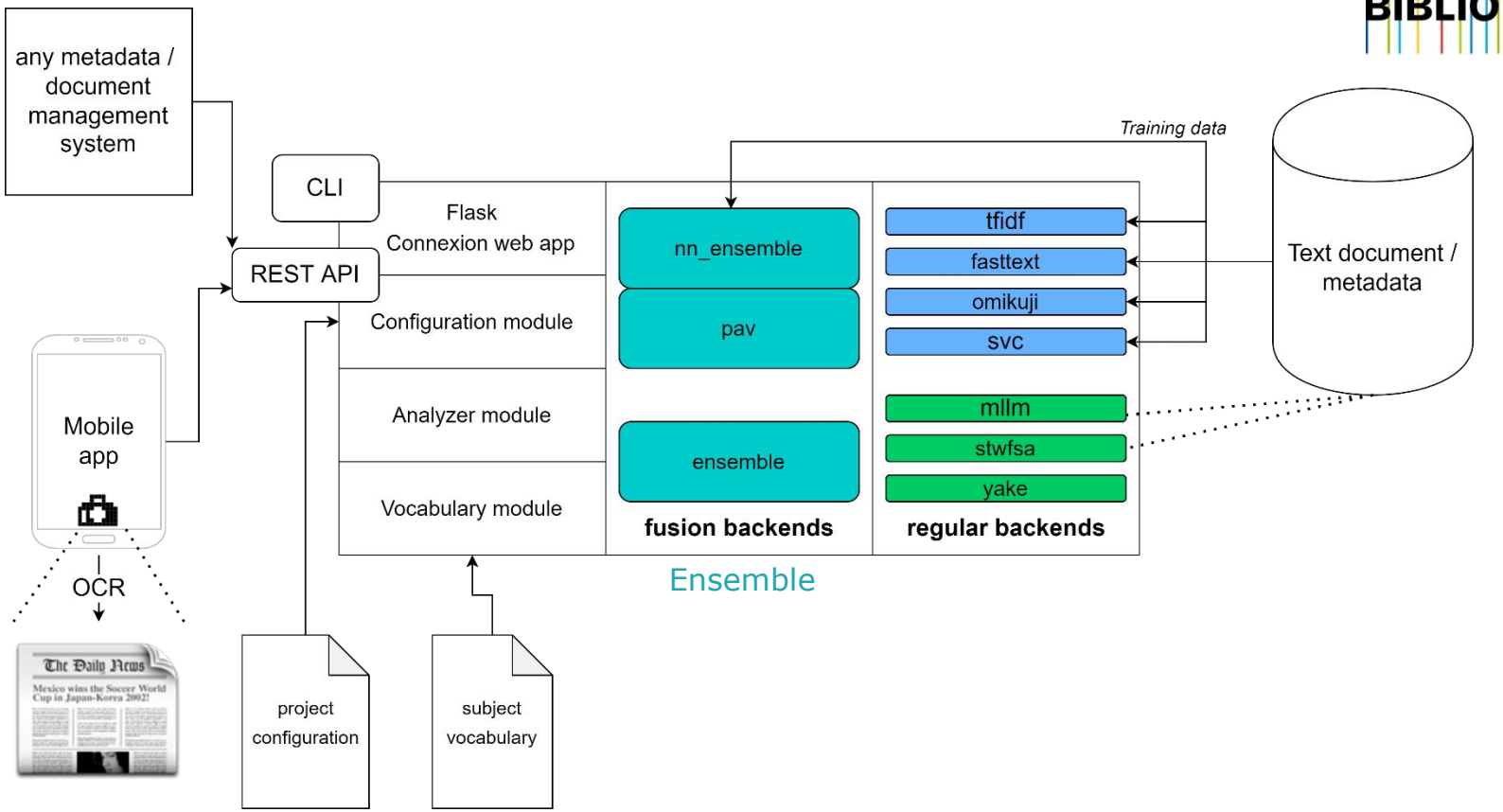
Adaption nach: https://github.com/NatLibFi/Annif/wiki/images/annif-architecture.png

annif - Architektur



Adaption nach: https://github.com/NatLibFi/Annif/wiki/images/annif-architecture.png

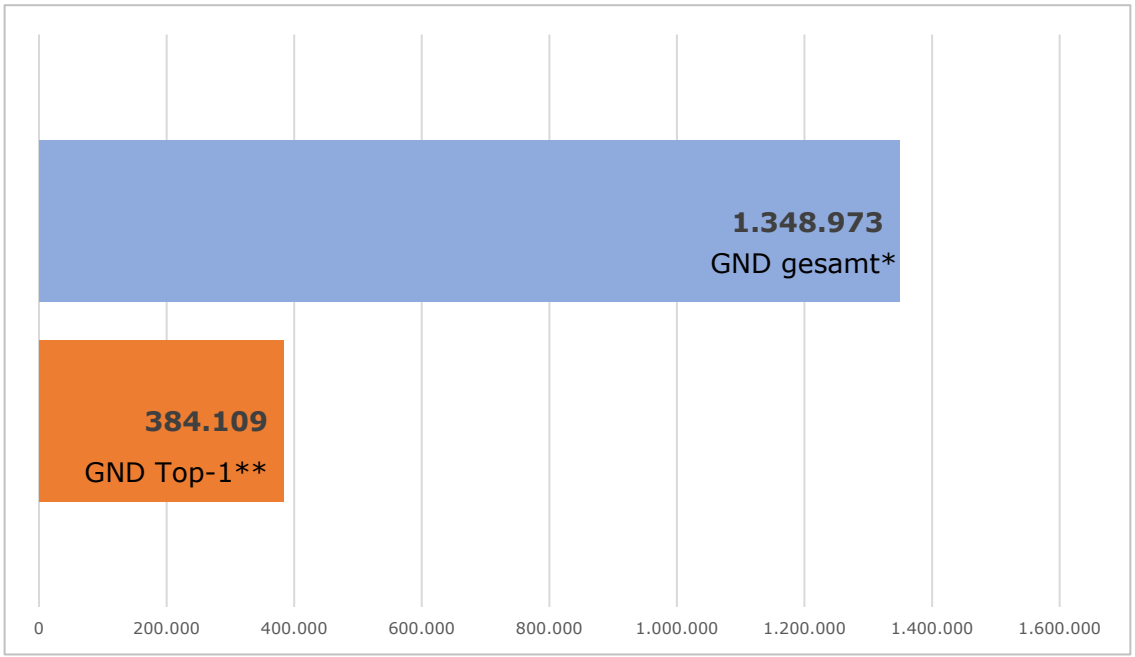
annif - Architektur



Adaption nach: https://github.com/NatLibFi/Annif/wiki/images/annif-architecture.png

Automatische Indexierung

Schlagwörter der Gemeinsamen Normdatei (GND)

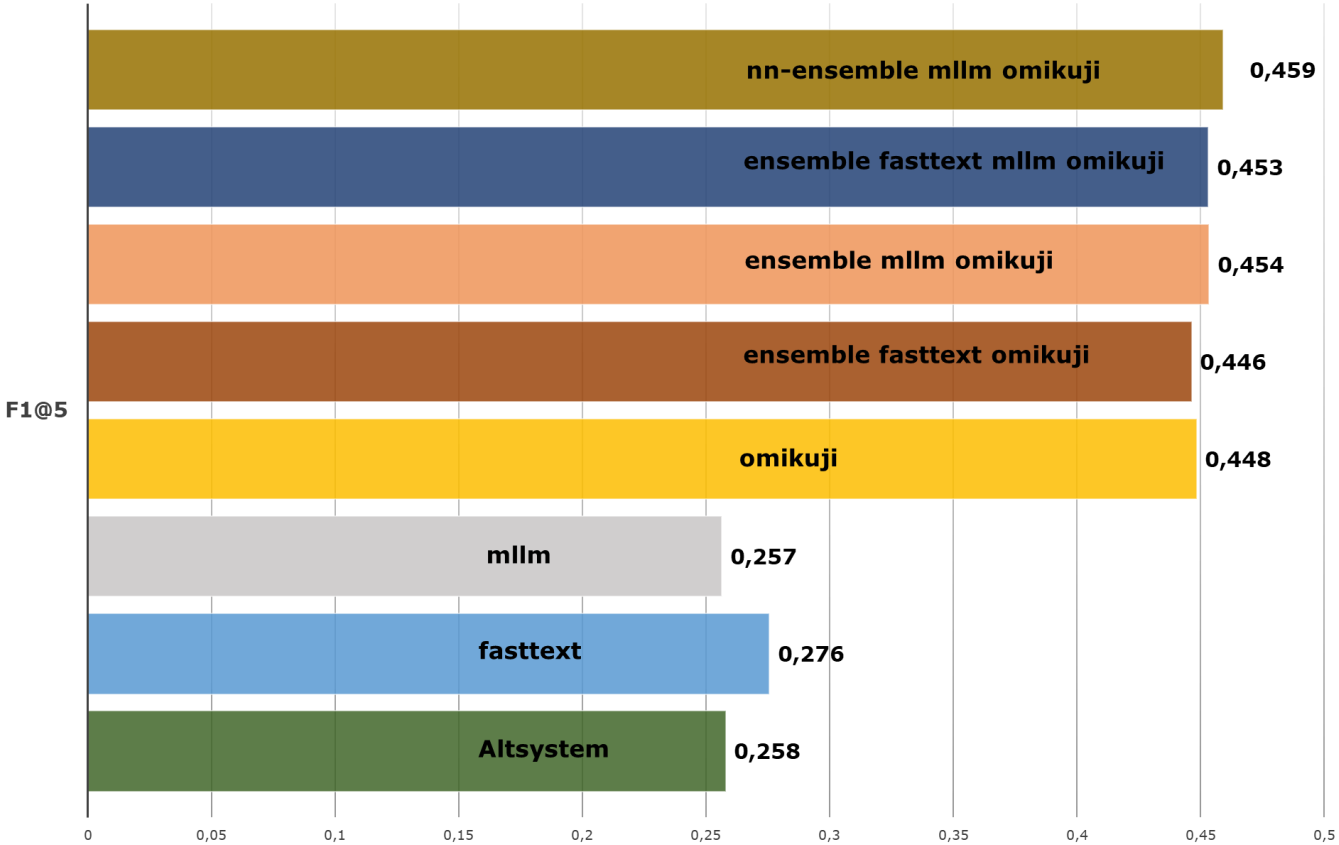


Trainingsmaterial, Deutsch:

- kein vollständiges Trainingsset für alle GND-Schlagwörter
- nur 384.109 haben mind. einen Textobjekt
- 964.864 haben kein Textobjekt

* Katalogisierungslevel 1 oder z und aus dem Teilbestand s
** alle aus GND gesamt* aber nur jene mit mind. einem Dokument als Trainingseinheit

Ergebnisse GND Indexierung



Testdaten:

1.261 Online Publikationen

Ausgabe:

n = 5 Schlagwörter

Metrik:

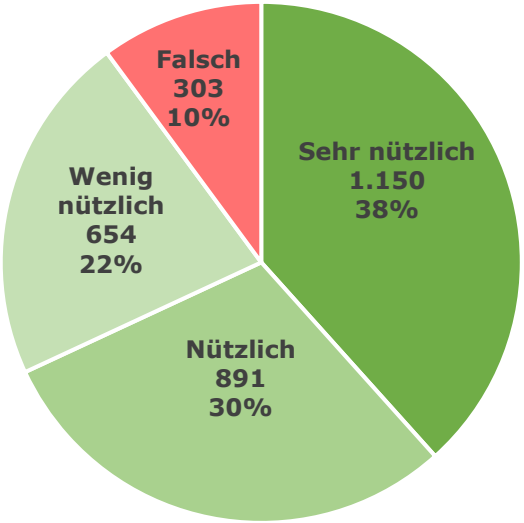
F1-score

Intellektuelle Bewertung

DEUTSCHE NATIONAL BIBLIOTHEK

702 Stichproben (Online Publikationen)

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



1.094 Fehlende Aspekte

Intellektuelle Bewertung durch die Indexierer* der Abt. Inhaltserschließung

Bewertungsskala:

- Sehr nützlich
- Nützlich
- Wenig nützlich
- Falsch

Go-Live* Erschließungsmaschine

Automatische Klassifizierung von Online- und ausgewählten Printpublikationen mit DDC-Sachgruppen und DDC-Kurznotationen

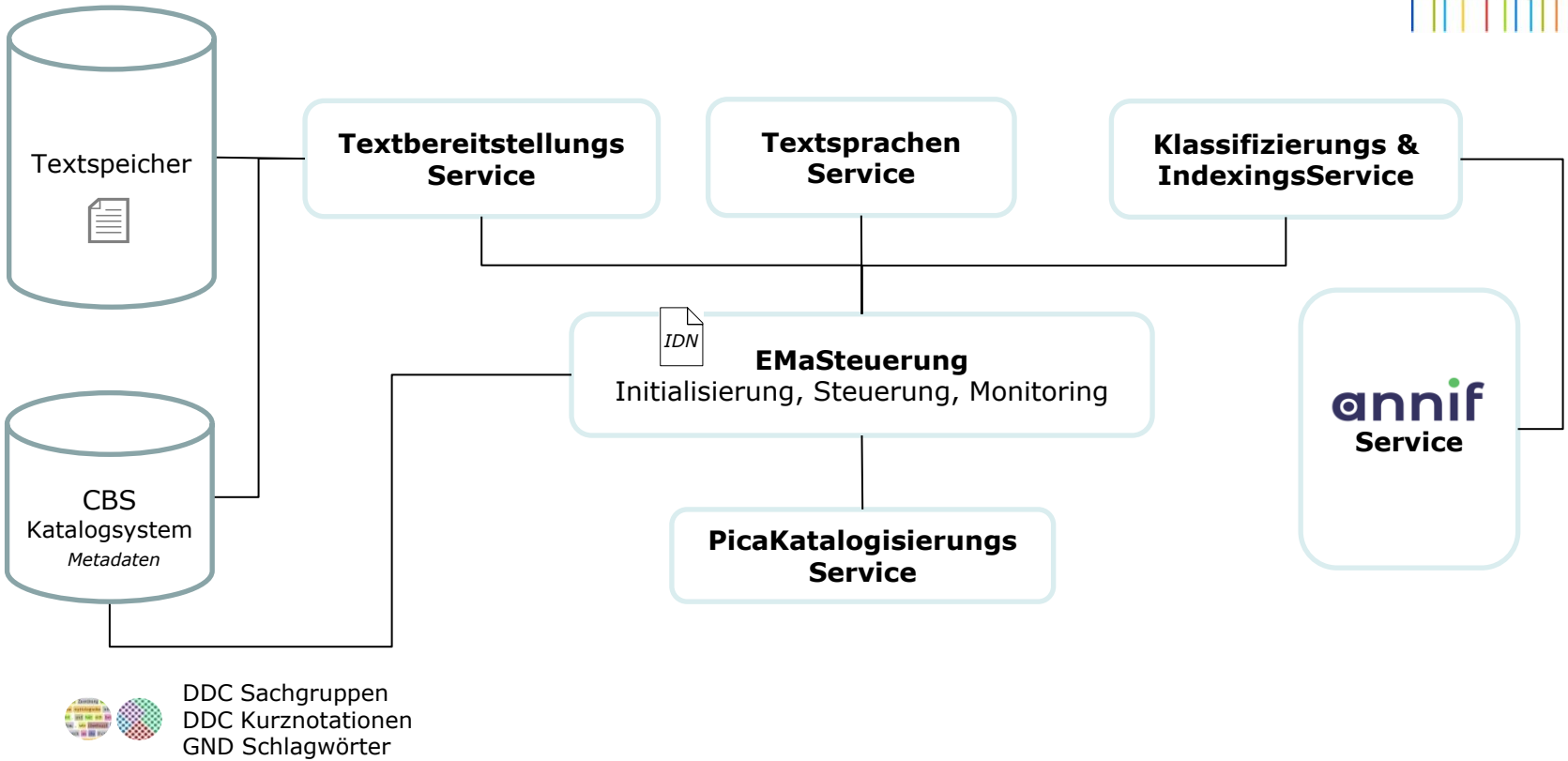
- *DDC-Sachgruppenvergabe ger*
- *DDC-Kurznotationsvergabe für die Sachgruppe Medizin ger eng*

Automatische Indexierung von Online- und ausgewählten Printpublikationen anhand der normierten Terminologien GND und LCSH

- *Beschlagwortung GND ger*

* https://blog.dnb.de/erschliessungsmaschine-gestartet/

Automatische Erschließung mit Annif als Service



Ausblick

- Überführung aller EMa-Themen in die Routine und Weiterentwicklung
- Umsetzung der bislang fehlenden Anwendungsfälle:
    - Sprachencodevergabe
    - Sachgruppenvergabe eng
    - Beschlagwortung GND eng
    - Beschlagwortung Kinder- und Jugendliteratur
    - Kurznotationsvergabe eng ger für alle weiteren Kurznotationssysteme
- sukzessive Außerbetriebnahme des Altsystems

Ausblick

- Projekt *„Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit KI"** gefördert durch das BKM im Rahmen der Nationalen KI-Strategie
- Ziel: Neue Methoden und Algorithmen aus dem Bereich der KI auswählen, untersuchen, adaptieren und zum Einsatz bringen

* https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/KI/ki_node.html

Danke für Ihre Aufmerksamkeit.

s.uhlmann@dnb.de

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