Elisabeth Mödden

Automatische Inhaltserschließung in der Deutschen Nationalbibliothek

Erschließungsmaschine EMa und KI-Projekt

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

DEUTSCHE NATIONAL BIBLIOTHEK

Hintergrund

- DNB sammelt jährlich mehr als 2 Mio. Publikationen
  - davon sind ca. 1 ½ Mio. Netzpublikationen (E-Books, E-Journals etc.)
- Instrumente der inhaltlichen Erschließung:

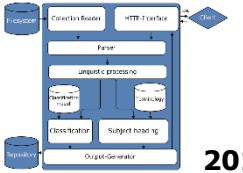
  DDC-Sachgruppen, DDC-Notationen und GND-Schlagwörter

Seit etwa 10 Jahren sind maschinelle Verfahren im Einsatz,

um inhaltsbeschreibende Metadaten zu generieren … und damit

das Finden im Katalog zu unterstützen.

Maschinelle Erschließung Projekte - Routine



2010
Einstellung der intellektuellen Erschließung von Netzpublikationen

2014
Masch. Beschlagwortung und Gründung des Referats AEN

2017
ToC-basierte, masch. Beschlagwortung von Printpublikationen

2019
EMa-Projekt

2022
Start Erschließungsmaschine mit „Annif als Service“ und
Ende EMa-Projekt

2009
PETRUS-Projekt - Zusammenarbeit mit Firma Averbis

2012
Masch. Klassifikation DNB-Sachgruppen

2015
Masch. Klassifikation DDC-Kurznotationen und
Ende PETRUS-Projekt

2018
Masch. Beschlagwortung englischspr. Netzpublikationen

2020
Masch. Beschlagwortung Kinder- und Jugendliteratur

2021
Start KI-Projekt „Automatisches Erschließungssystem“

Klassifkation

DEUTSCHE NATIONAL BIBLIOTHEK

104 DDC-Sachgruppen

000	Allgemeines, Wissenschaft
004	Informatik
010	Bibliografien
…	
500	Naturwissenschaften
510	Mathematik
520	Astronomie, Kartographie
530	Physik
…	
600	Technik
610	Medizin, Gesundheit
620	Ingenieurwissenschaften und Maschinenbau
621.3	Elektrotechnik, Elektronik
…	

DDC-Kurznotationen

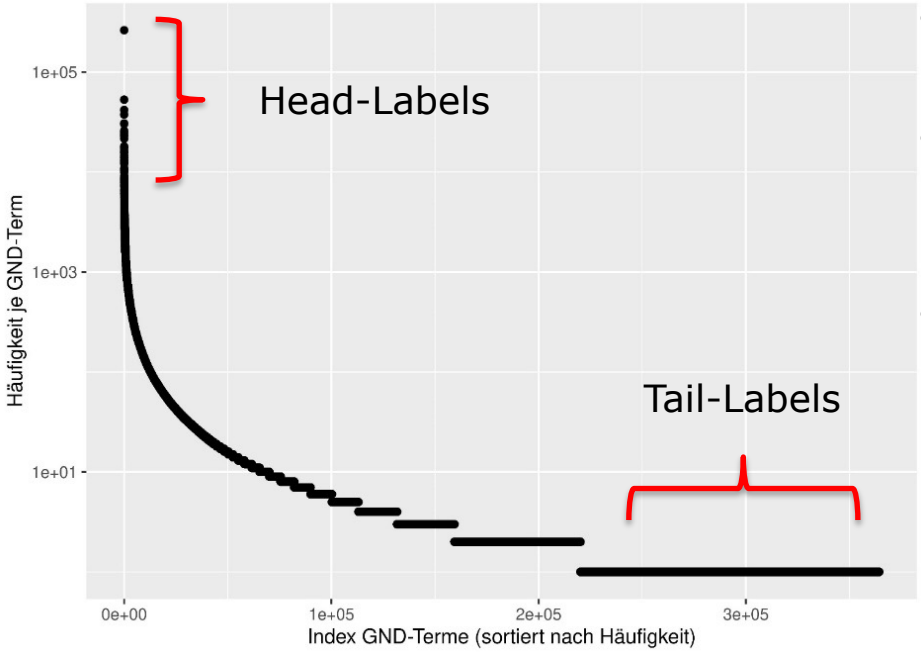
- Entwickelt 2005/2006 für die Sachgruppe Medizin
- Seit 2017 Entwicklung von Kurznotationen für alle Sachgruppen

Thema:
Studie
Übergewicht bei Kindern
Kiel
2000-2009

DDC-SG 610

DDC-Notation 618.92398009435123090511
Kurznotation 618.92~~398009435123090511~~

GND–Vokabular für die Beschlagwortung



- 1,35 Mio. GND-Datensätze Qualitätslevel 1 und Teilbestand s
- nur ca. 385.000 GND-Entitäten mit mindestens einer Publikation im Bestand der DNB verknüpft
- Long-Tail Charakteristik für typische DNB Trainingsdaten

+ 1 Million Zero-Shot-Labels
(ohne Trainingsdaten)

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Beschlagwortung als XMLC-Problem

- Maschinelle Beschlagwortung von Texten mit Konzepten aus der GND lässt sich als sogenanntes E**x**treme **M**ulti-**L**abel **C**lassification-Problem abstrahieren
  - Eingehende Text-Dokumente werden mit a-priori fest stehenden Labels (GND-Konzepte) verknüpft. Die Menge der zutreffenden Labels pro Dokument ist nicht beschränkt.
- Charakteristisch für XMLC-Probleme sind[1]:
  - Große „Label-Menge" ~ 10^5 – 10^6 Labels
  - Long-Tail-Charakteristik: Ein Großteil der möglichen Labels kommt in Trainingsdaten selten oder nie vor

[1]vgl. u.a. Jain etal. 2016 "Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications"

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Erschließung in der DNB

Klassifikation

DDC-Sachgruppen und
DDC-Kurznotationen

Maschinell lernende Verfahren

Beschlagwortung

GND
als normiertes Vokabular

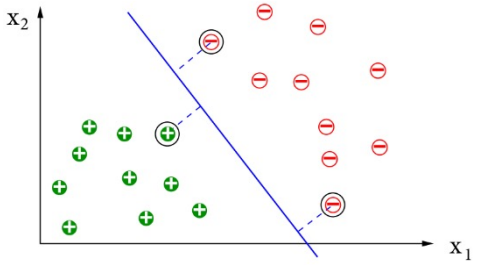
Kombination: Lexikalische und maschinell lernende Verfahren

Netzpublikationen und ausgewählte Printpublikationen

Deutsch und Englisch

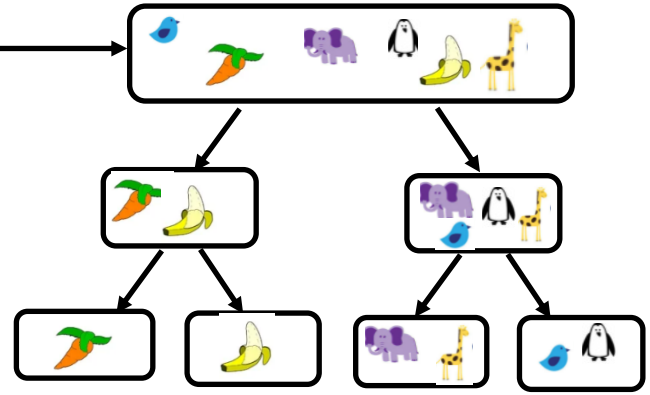
Maschinell lernende Verfahren

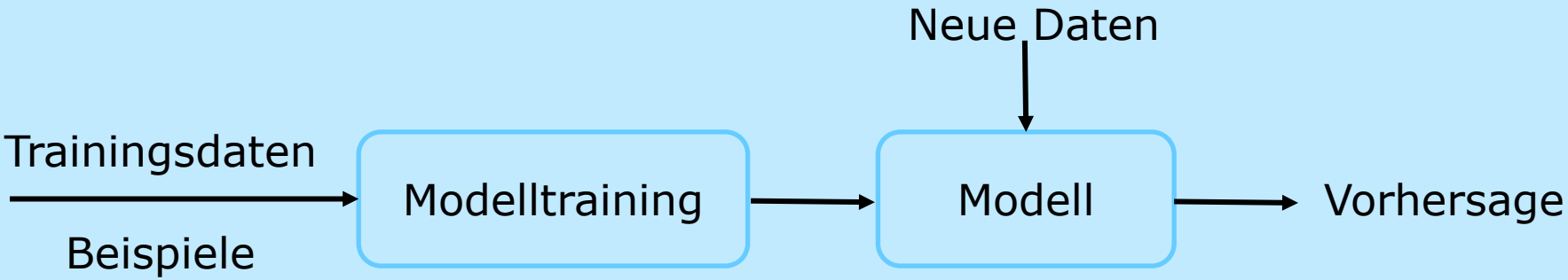
Support-Vektor-Maschine



Lernende Entscheidungsbäume

Wikipedeia:
„Entscheidungsbäume (englisch: decision tree) sind geordnete, gerichtete Bäume, die der Darstellung von Entscheidungsregeln dienen. Die grafische Darstellung als Baumdiagramm veranschaulicht hierarchisch aufeinanderfolgende Entscheidungen. Sie haben eine Bedeutung in zahlreichen Bereichen, in denen automatisch klassifiziert wird oder aus Erfahrungswissen formale Regeln hergeleitet oder dargestellt werden."





Lexikalische Verfahren

DEUTSCHE NATIONAL BIBLIOTHEK

Abgleich zwischen relevanten Termen im Text und Termen im kontrollierten Vokabular

Als Wasserbau werden Maßnahmen, technische Eingriffe und Bauten im Bereich des Grundwassers, der Oberflächengewässer und der Meeresküsten bezeichnet. Heute weniger gebräuchlich ist die Bezeichnung Hydrotechnik für dieses Fachgebiet..* Text aus Wikipedia

Wasserbau
Sachbegriff

GND-Nummer :	4064700-6
Untergliederung :	Allgemeinbegriff [saz]
Systematik :	31.3b Bautechnik
DDC-Notation :	627
Quelle :	M
Varianten :	Hydrotechnik Wasserbauwerk
Thematischer Bezug :	Unterwasserbau (verw
Oberbegriffe :	Bauwesen (Oberbegriff

Grundwasser
Sachbegriff

GND-Nummer :	4022369-3
Untergliederung :	Allgemeinbegriff [saz]
Systematik :	19.3 Hydrologie, Meereskunde
DDC-Notation :	2--1698 551.49 553.79 628.114

Küste
Sachbegriff

GND-Nummer :	4033469-7
Untergliederung :	Allgemeinbegriff [saz]
Systematik :	19.1b Physische Geografie
DDC-Notation :	2--146 551.457 577.699
Quelle :	M
Verwendungshinweis :	Mit einzelnen Gewässern wird, soweit s gebildet
Varianten :	Meeresküste

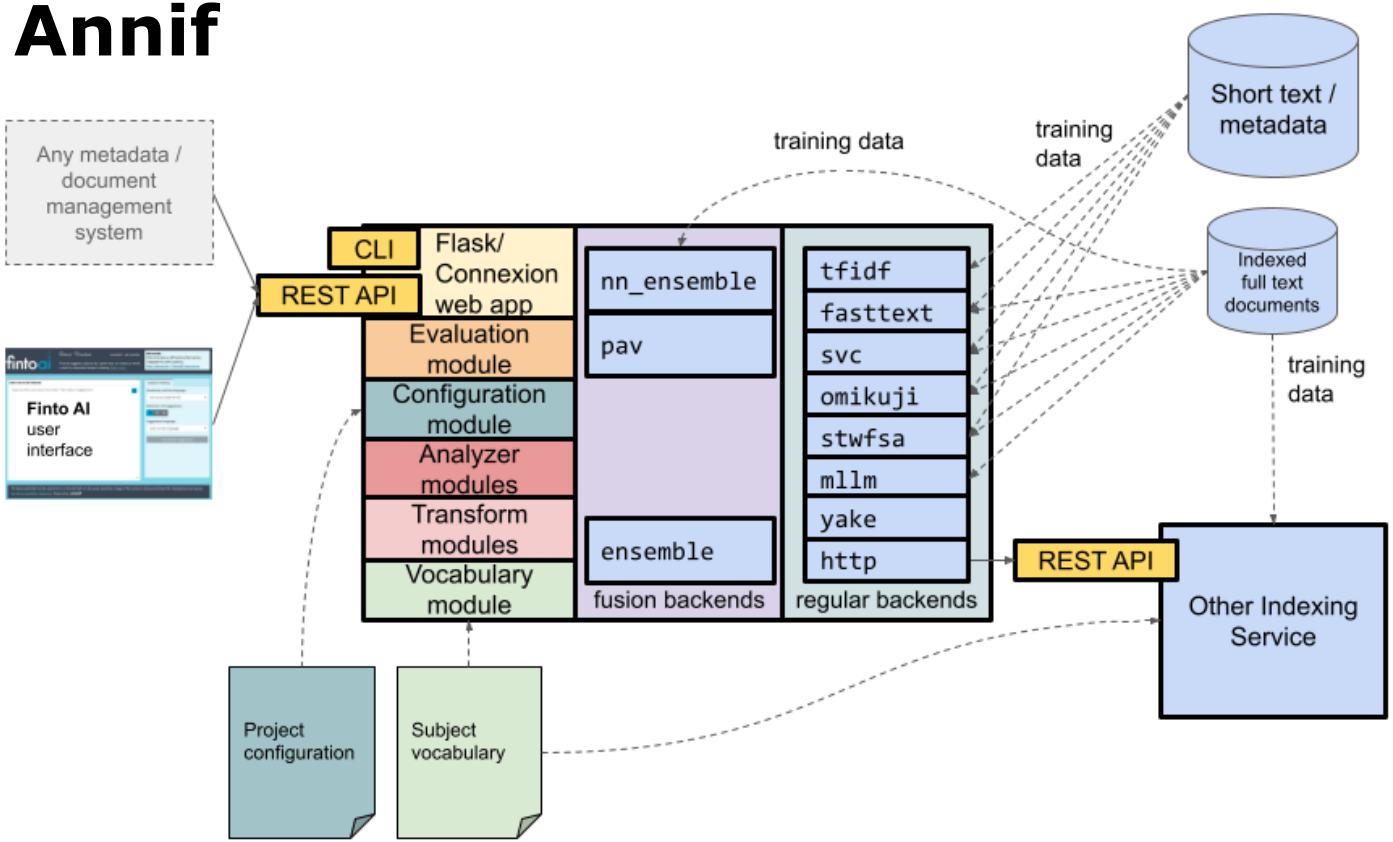
Oberflächengewässer
Sachbegriff

GND-Nummer :	4172246-2
Untergliederung :	Allgemeinbegriff [saz]
Systematik :	19.3 Hydrologie, Meereskunde
Quelle :	Taschenlex. Wasser § 323,4 c RSWK
Thematischer Bezug :	Oberflächenwasser (verwandter Begriff)
Oberbegriffe :	Gewässer (Oberbegriff generisch)

begriff generisch)

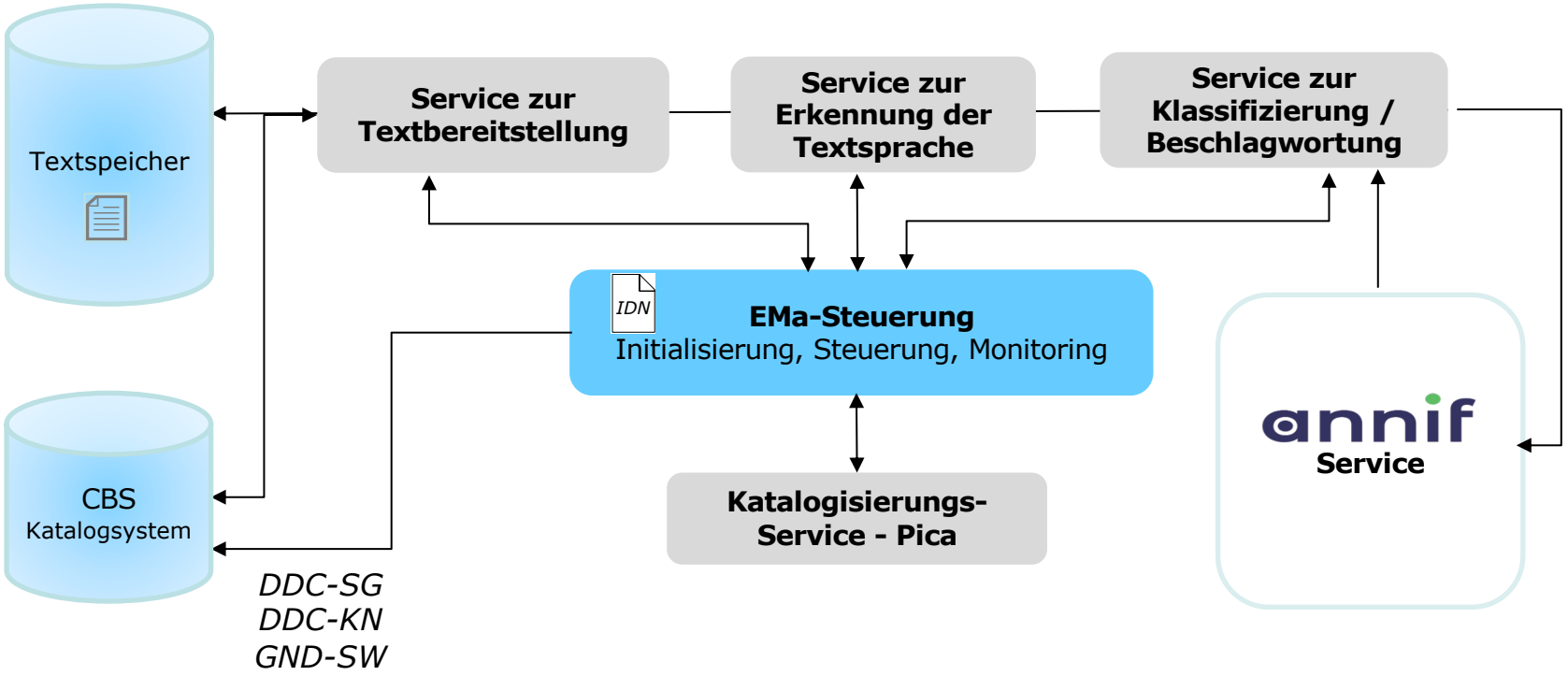
DEUTSCHE NATIONAL BIBLIOTHEK

Annif



https://github.com/NatLibFi/Annif/wiki/Architecture

Modulare Erschließungsmaschine mit Annif als Service



DEUTSCHE NATIONAL BIBLIOTHEK

Beispiel I

Titel: Hunde-Sprechstunde : Alles über Krankheiten, Fellpflege und Ernährung. So bleibt Ihr Hund gesund. Erkrankungen erkennen und vorbeugen, ein informativer Ratgeber für alle Hunderassen

Masch. Sachgruppe: 630 Landwirtschaft, Veterinärmedizin

Masch. Kurznotation:636.70889 Hunde—Tiermedizin

Masch. Schlagwörter:!040261816!Hund [Ts1]
!04026193X!Hundekrankheit [Ts1
!041287746!Tierarzt [Ts1]
!041842170!Symptom [Ts1]
!040328449!Krankheit [Ts1]

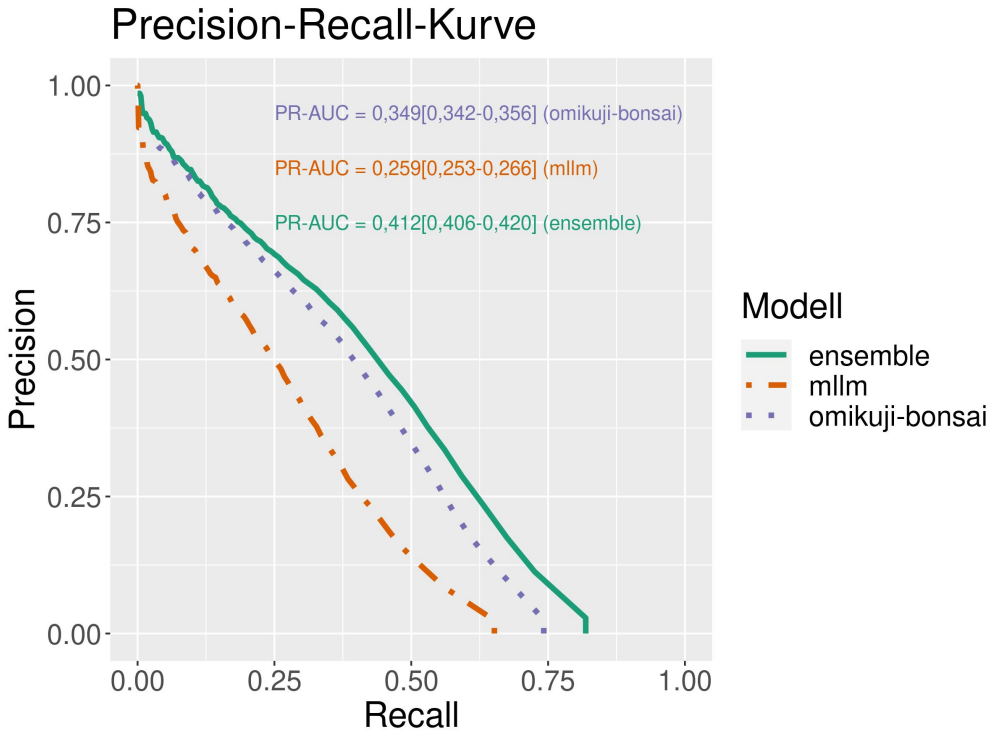
DEUTSCHE NATIONAL BIBLIOTHEK

Beispiel II

Titel: Wet-coffee processing production wastes : quality, potentials, and valorization opportunities

Masch. Sachgruppe: 660 Technische Chemie
Masch. Kurznotation: 663 Getränketechnologie
Masch. Schlagwörter: !041670329! Lebensmittelanalyse [Ts1]
!041302370! Aromastoff [Ts1]
!041361733! Reaktionstechnik [Ts1]
!041629957! Kaffeeanbau [Ts1]
!04178331X! Röstkaffee [Ts1]
!040291898! Kaffee [Ts1]
!04253822X! Nebenprodukt [Ts1]

Ergebnisse maschinelle Beschlagwortung in der Routine



F1@5 Document Average on Test-Set

ensemble	0,376 (0,373-0,380)
mllm	0,275 (0,272-0,278)
omikuji-bonsai	0,349 (0,345-0,352)

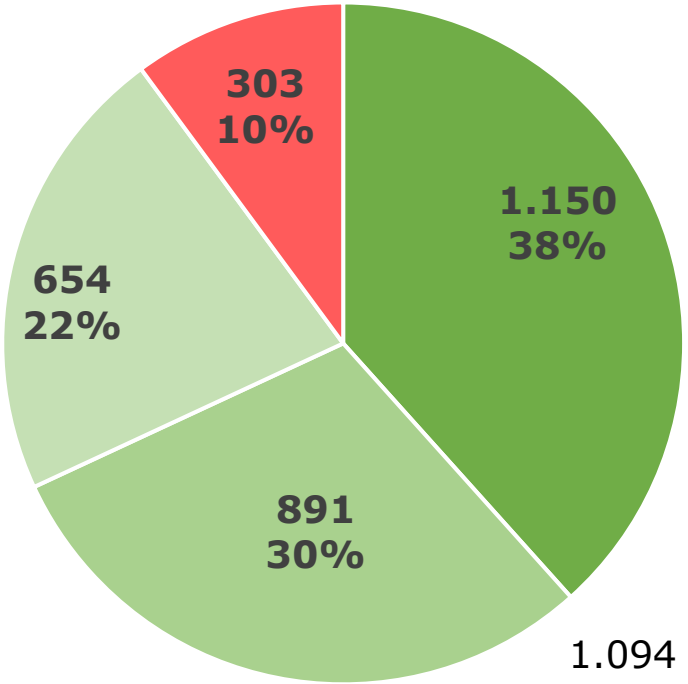
mllm: Maui like lexical matching (vgl. annif.org)
Omikuji-bonsai: Khandagale etal. 2016 "Bonsai: diverse and shallow trees for extreme multi-label classification"

Ergebnisse – Intellektuelle Bewertung

DEUTSCHE NATIONAL BIBLIOTHEK

702 Stichproben (Online-Publikationen)

2.998 durch Annif Ensemble MLLM & omikuji-bonsai vergebene GND-Schlagwörter
(GND-Schlagwörter n = max 6)



Bewertung durch die Fachreferent*innen der Abteilung Inhaltserschließung (IE)

Bewertungsskala:

- Sehr nützlich
- Nützlich
- Wenig nützlich
- Falsch

1.094 Fehlende Aspekte

Maschinelle Erschließung in der Routine

Was wird wie erschlossen?	Publikationen					Maschinell erschlossene Publikationen + Artikel Anzahl (Stand: 11/2022)
	Print				Digital	
	Reihe A		Reihe B	Reihe H	Reihe O	
	buchaffin	nicht buchaffin				
DNB-Sachgruppen						4,8 Mio.
DDC-Kurznotation (nur maschinell)						2,7 Mio.
DDC-Notation (nur intellektuell)						
GND-Schlagwörter						1,0 Mio.

DEUTSCHE NATIONAL BIBLIOTHEK

KI-Projekt
Automatisches Erschließungssystem

- Förderung im Rahmen der nationalen KI-Strategie
  - Beauftragte der Bundesregierung für Kultur und Medien (BKM)
- Laufzeit: 3 1/2 Jahre (Oktober 2021 – März 2025)
- Ausstattung
  - Personal (ca. 4 VZÄ)
  - Server mit 2 * 24 Prozessorkernen, 1 ½ TB RAM, 3.8 TB SSD-Festplatte
  - Mittel für Forschungskooperationen
- https://www.dnb.de/ki-projekt

DEUTSCHE NATIONAL BIBLIOTHEK

Projektziele

- Qualität der maschinellen Beschlagwortung mit der GND durch passende Methoden messbar verbessern
  (F-Score = 0,4 oder besser)
- Technologie- und Wissenstransfer in die bibliothekarische Praxis

Projekt ist unser Labor für die Erforschung –

EMa das modulare (Steckkasten-)System für die Produktion

DEUTSCHE NATIONAL BIBLIOTHEK

Projektziele

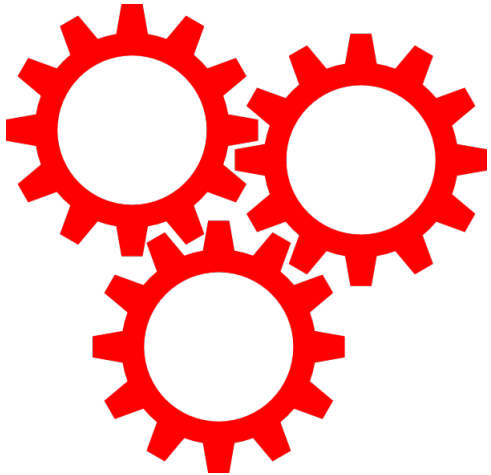
- vielversprechende neue Methoden/Algorithmen finden, systematisch untersuchen und ggf. adaptieren
  - die Qualität der GND-Verknüpfungen verbessern
  - (später) auch semantische Konzepte ohne GND-Repräsentation finden
  - Evaluation mit deutschsprachigen wissenschaftlichen Netzpublikationen
- daraus Werkzeuge entwickeln und bereitstellen
  - Services für die Einbindung in Erschließungssysteme mit modularer Architektur (EMa und andere)

DEUTSCHE NATIONAL BIBLIOTHEK

Forschungsschwerpunkte des Projektes

Methoden für die Textanalyse und thematische Einordnung

(semantische Verknüpfung der Texte mit den Konzepten der GND)



***Methoden für die* Extraktion und Aufbereitung der Texte und bibliografischen Daten**

***Methoden für die* Extraktion und Aufbereitung des Vokabulars**
(1,35 Mio. potenzielle Konzepte der GND)

DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Inhaltserschließung in der Deutschen Nationalbibliothek

Herzlichen Dank – Fragen?

Dipl.-Ing. Elisabeth Mödden

Leitung
Automatische Erschließungsverfahren, Netzpublikationen
Deutsche Nationalbibliothek
+49-69-1525-1533
e.moedden@dnb.de