Christoph Poley & Sandro Uhlmann

In der Deutschen Nationalbibliothek lesen jede Nacht die Maschinen – Automatische Inhaltserschließung im täglichen Einsatz

7. Online-Workshop Computerunterstützte Inhaltserschließung | 14. & 15. November 2023

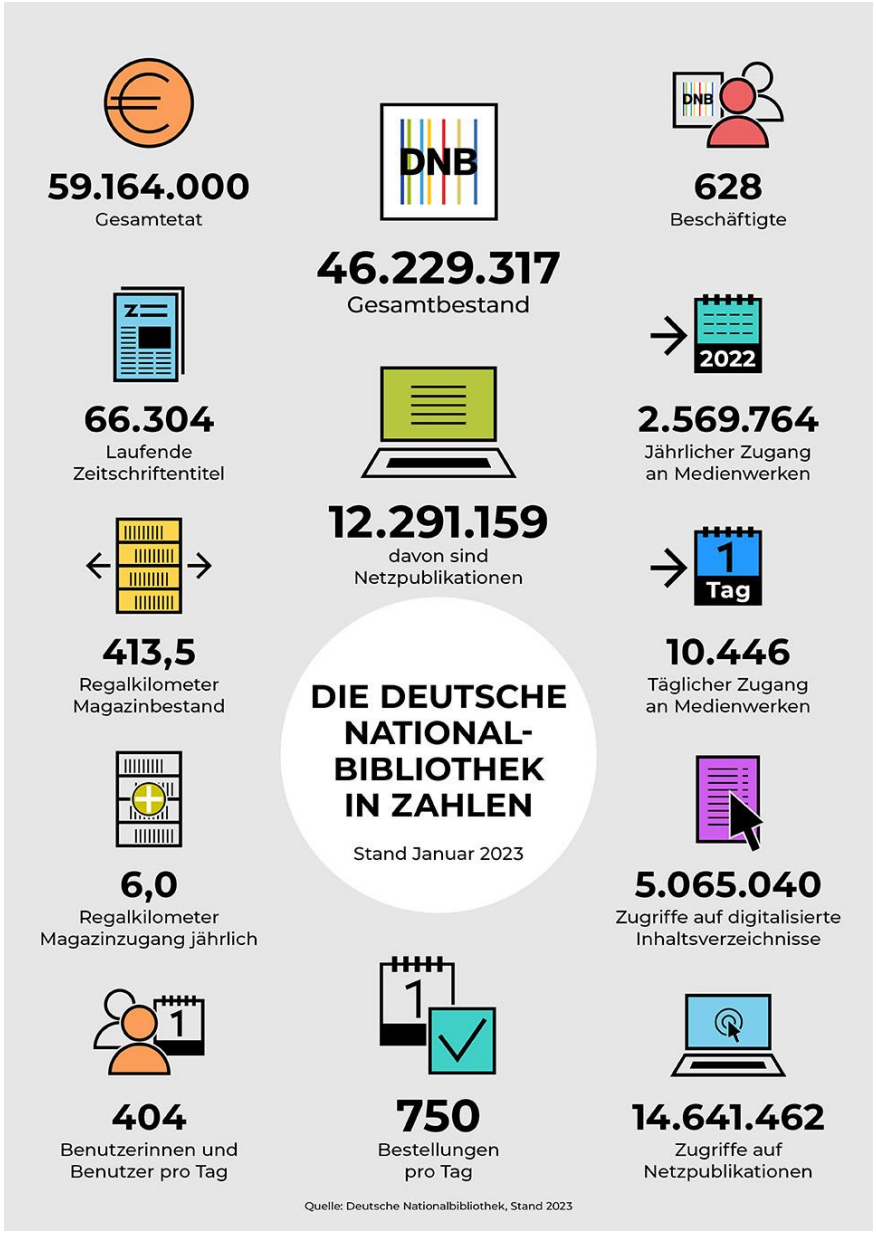
Automatische Inhaltserschließung in der DNB

Agenda:

- DNB und täglich neue Publikationen und Themen
- Automatische Inhaltserschließung an der DNB
- Automatische Indexierung mit GND
- Ausweitung der automatischen Inhaltserschließung
- Weiterentwicklung der Erschließungsmodule

Deutsche Nationalbibliothek

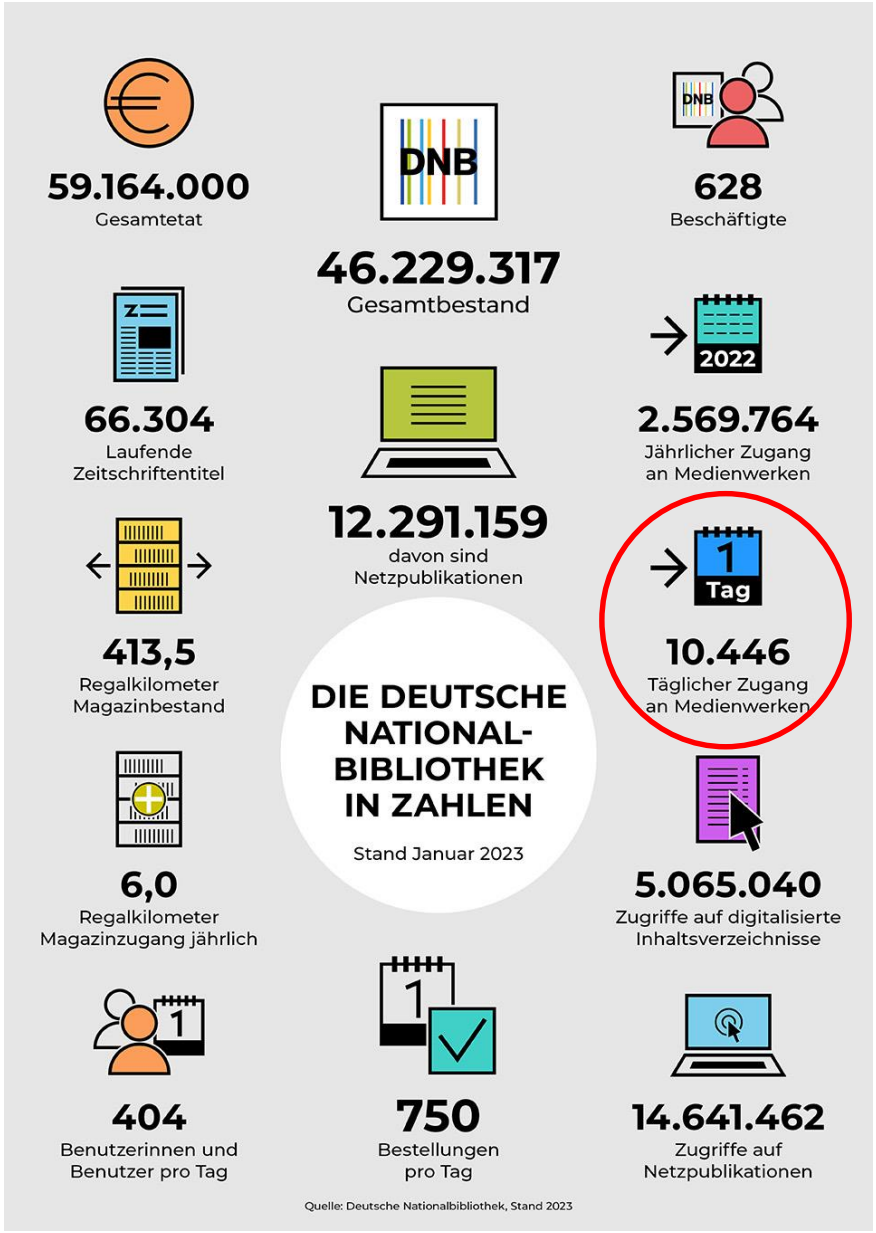
Zentrale Archivbibliothek Deutschlands mit den Standorten Leipzig (gegr. 1912) und Frankfurt am Main (gegr. 1947)



https://www.dnb.de/DE/Ueber-uns/Portraet/portraet_node.html

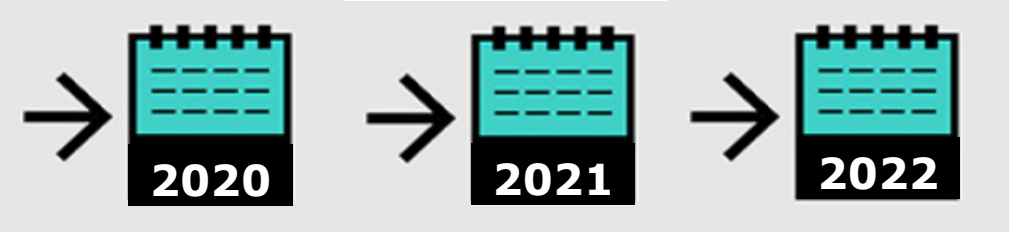
Deutsche Nationalbibliothek

Zentrale Archivbibliothek Deutschlands mit den Standorten Leipzig (gegr. 1912) und Frankfurt am Main (gegr. 1947)



https://www.dnb.de/DE/Ueber-uns/Portraet/portraet_node.html

Täglich neue Publikationen und Themen



Jährlich Ø 297.506 neue digitale monografische Publikationen [1]

Täglich Ø 815 neue digitale monografische Publikationen

- Tagesaktuelle automatische Anreicherung der neu hinzukommenden Publikationen mit DDC-Sachgruppen, DDC-Kurznotationen und Schlagwörtern der GND

[1] Hier unberücksichtigt bleiben bspw. die aktuell 806.733 digitalen Periodika (Ausgaben, Hefte oder Artikel) pro Jahr. Der Gesamtbestand der DNB beträgt derzeit 46,2 Millionen Medienwerke, wovon 12,3 Millionen digitale Medienwerke sind.
Deutsche Nationalbibliothek Jahresbericht 2022: Zahlen und Fakten.
https://jahresbericht.dnb.de/Webs/jahresbericht/SharedDocs/Downloads/DE/2022statistikenGesamt.html

Automatische Inhaltserschließung in der DNB

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

Produktiv seit 2012

Machine Learning

Automatische Indexierung von Online- und ausgewählten Printpublikationen mit normierter Terminologie GND

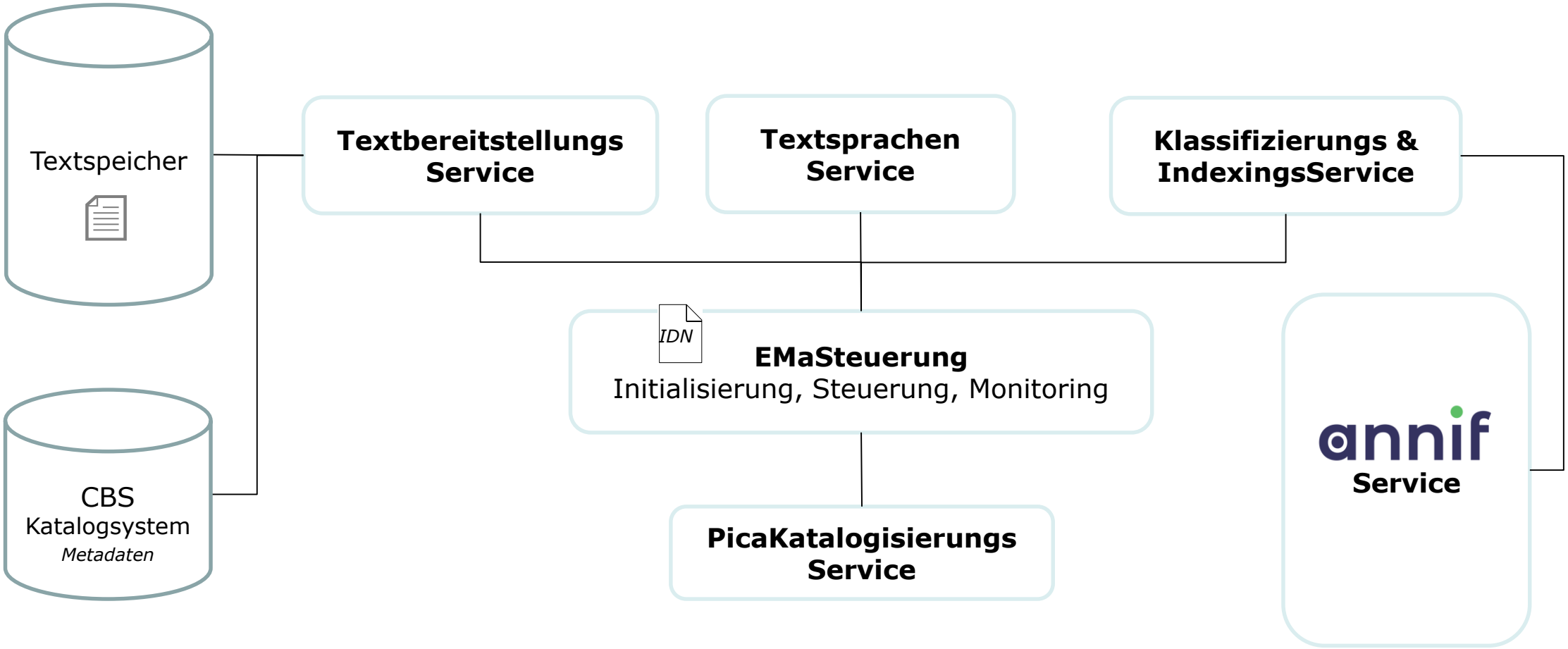
Produktiv seit 2014

Machine Learning

Text Mining

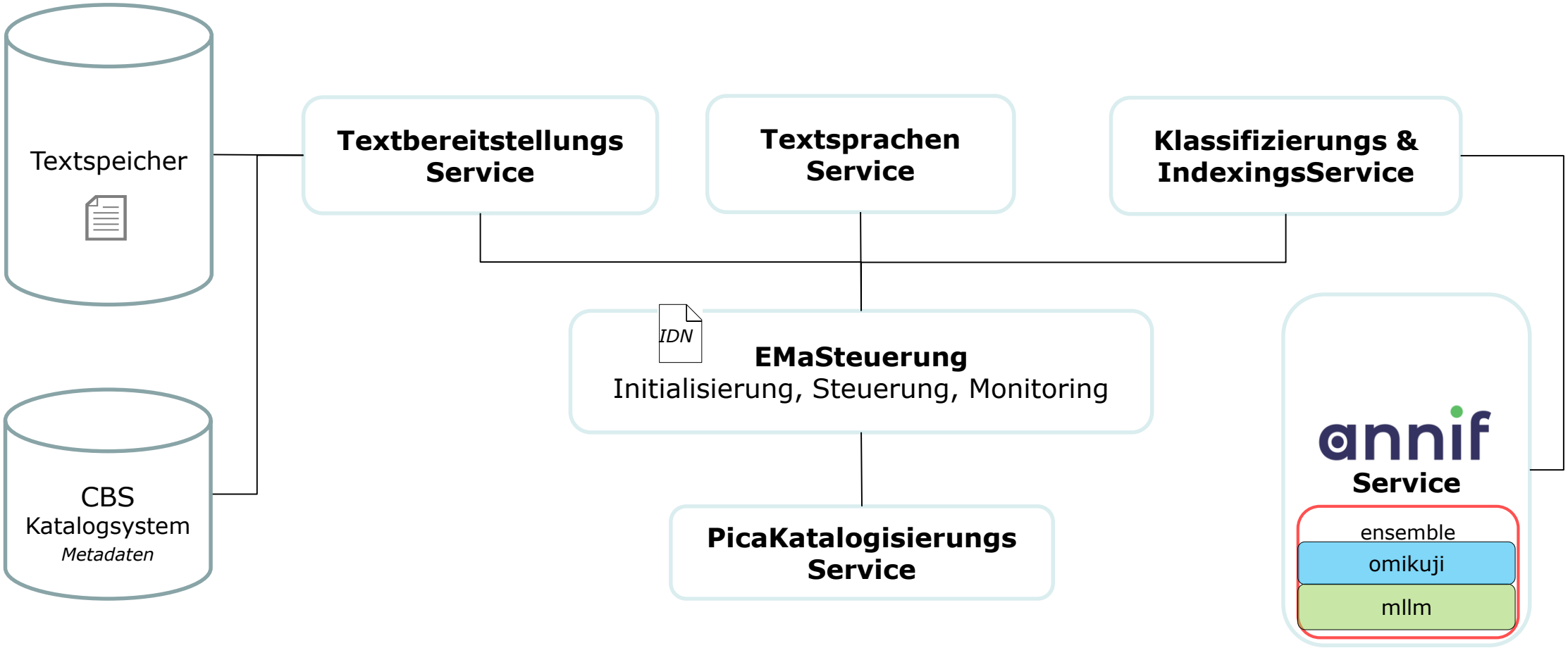
Junger, Ulrike; Scholze, Frank: Neue Wege und Qualitäten – Die Inhaltserschließungspolitik der Deutschen Nationalbibliothek. In: Qualität in der Inhaltserschließung, herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl und Hans Schürmann, Berlin, Boston: De Gruyter Saur, 2021, S. 55–70.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

„Erschließungsmaschine (EMa)“* mit Annif** als Service



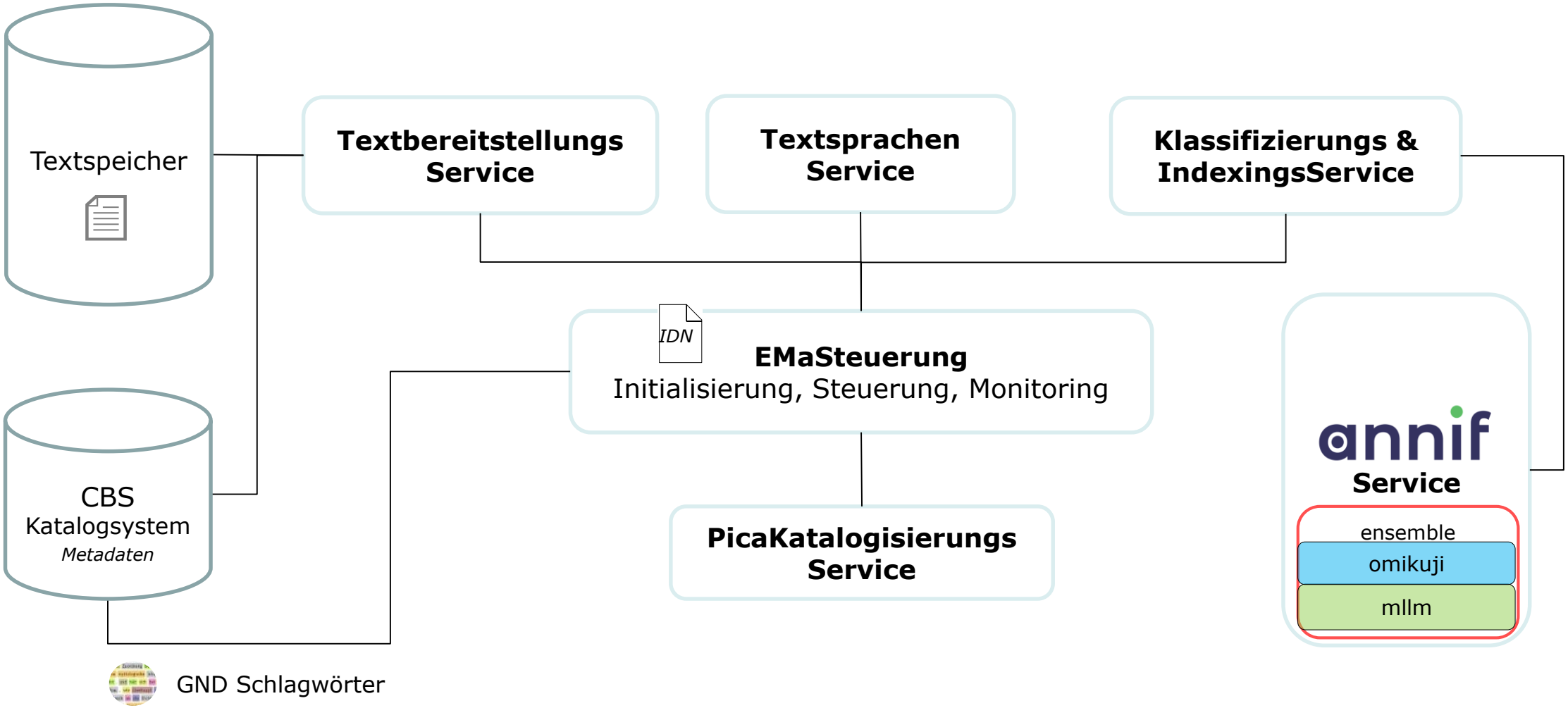
*Busse, Frank et al.: Erschließungsmaschine gestartet. In: DNB Blog 04.05.2022. https://blog.dnb.de/erschliessungsmaschine-gestartet/
**Suominen, Osma; Inkinen, Juho; Lehtinen, Mona: Annif and Finto AI: Developing and Implementing Automated Subject Indexing. JLIS.It, 13(1), 265–282. https://doi.org/10.4403/jlis.it-12740 | github https://github.com/NatLibFi/Annif

„Erschließungsmaschine (EMa)“* mit Annif** als Service



*Busse, Frank et al.: Erschließungsmaschine gestartet. In: DNB Blog 04.05.2022. https://blog.dnb.de/erschliessungsmaschine-gestartet/
**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

„Erschließungsmaschine (EMa)“* mit Annif** als Service



*Busse, Frank et al.: Erschließungsmaschine gestartet. In: DNB Blog 04.05.2022. https://blog.dnb.de/erschliessungsmaschine-gestartet/
**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

Automatische Indexierung mit GND als XMLC-Problem

E**x**treme **M**ulti-**L**abel **C**lassification

- neue Textdokumente werden mit fest stehenden Labels (GND-Schlagwörtern) verknüpft
- die Anzahl an zutreffenden Labels pro Textdokument ist nicht beschränkt

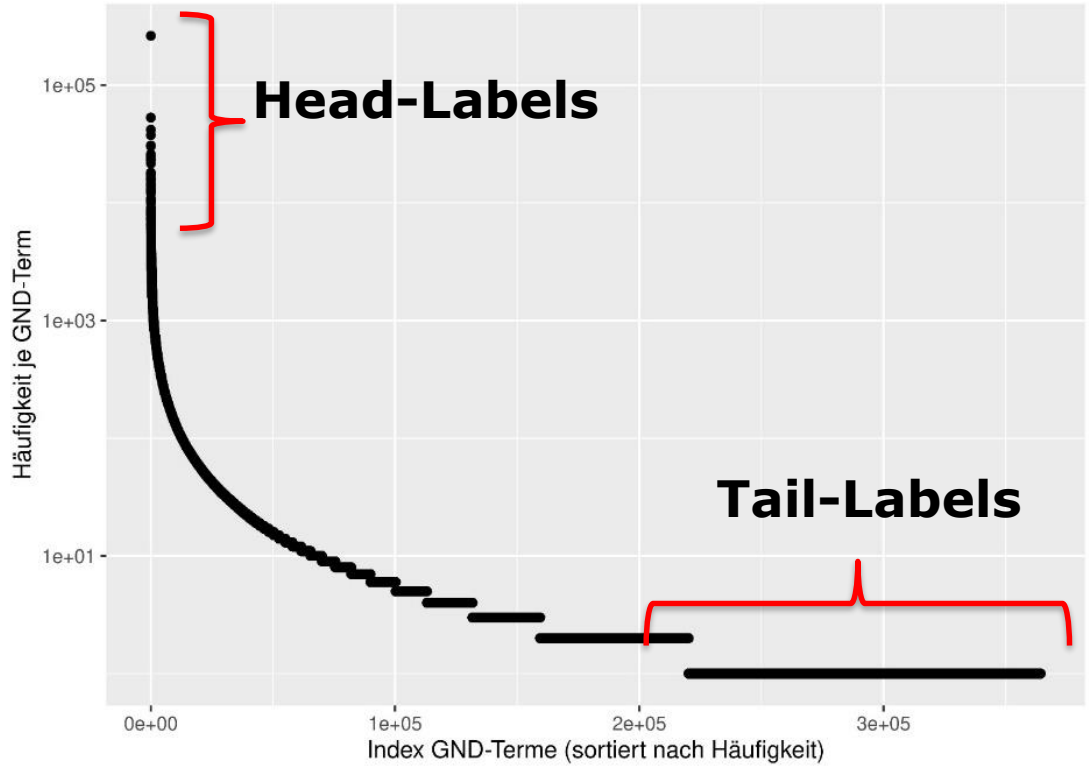
Charakteristisch für XMLC-Probleme sind*

- große „Label-Menge“: ~ 10^5 – 10^6 Labels
- Long-Tail-Charakteristik: Ein Großteil der möglichen Labels kommt in den Trainingsdaten selten oder nie vor

* Vgl. u.a. Jain, H. etal.: Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. KDD 2016, p. 935–944 https://doi.org/10.1145/2939672.2939756

Long-Tail Charakteristik

Long-Tail Charakteristik für typische DNB Trainingsdaten



*GND gesamt**
1.388.832 Schlagwörter

mind. einmal oder häufiger im Trainings- oder Testdatenset verwendet:
233.661 Schlagwörter

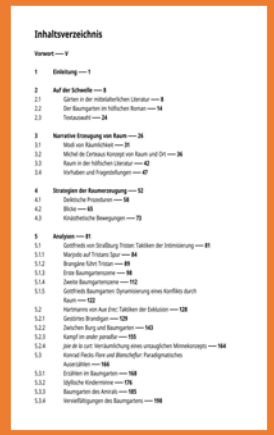
1.155.171 **Zero-Shot-Labels**
(ohne Trainingsdaten)

Anm.: nur Katalogisierungslevel 1 oder z und aus dem Teilbestand s

Trainingsdokumente GND ger

DEUTSCHE NATIONAL BIBLIOTHEK

Titel 1.464.992



Inhaltsverzeichnisse 717.354

Inhaltstexte 291.824

Volltexte 208.915

Ergebnisse GND Indexierung ger

Experiment* | Basis 6.970 Testdokumente (digitale Publikationen ger)

Modell	Precision	Recall	F1-Score
Ensemble	0,4047	0,5598	**0,4162**
MLLM	0,2088	0,3953	**0,2447**
Omikuji 1	0,3823	0,5214	**0,3865**
Omikuji 2	0,3875	0,4893	**0,3759**

*Ausgabe pro Dokument: 7 GND-Schlagwörter (Limit) & Schwellenwert von 0,05 (Threshold)

Ergebnisse GND Indexierung ger

DEUTSCHE NATIONAL BIBLIOTHEK

Experiment* | Basis 6.970 Testdokumente (digitale Publikationen ger)

Modell	Precision	Recall	F1-Score
Ensemble	0,4047	0,5598	**0,4162**
MLLM	0,2088	0,3953	**0,2447**
Omikuji 1	0,3823	0,5214	**0,3865**
Omikuji 2	0,3875	0,4893	**0,3759**

Produktiv* | Basis 14.457 Testdokumente (digitale Publikationen ger)

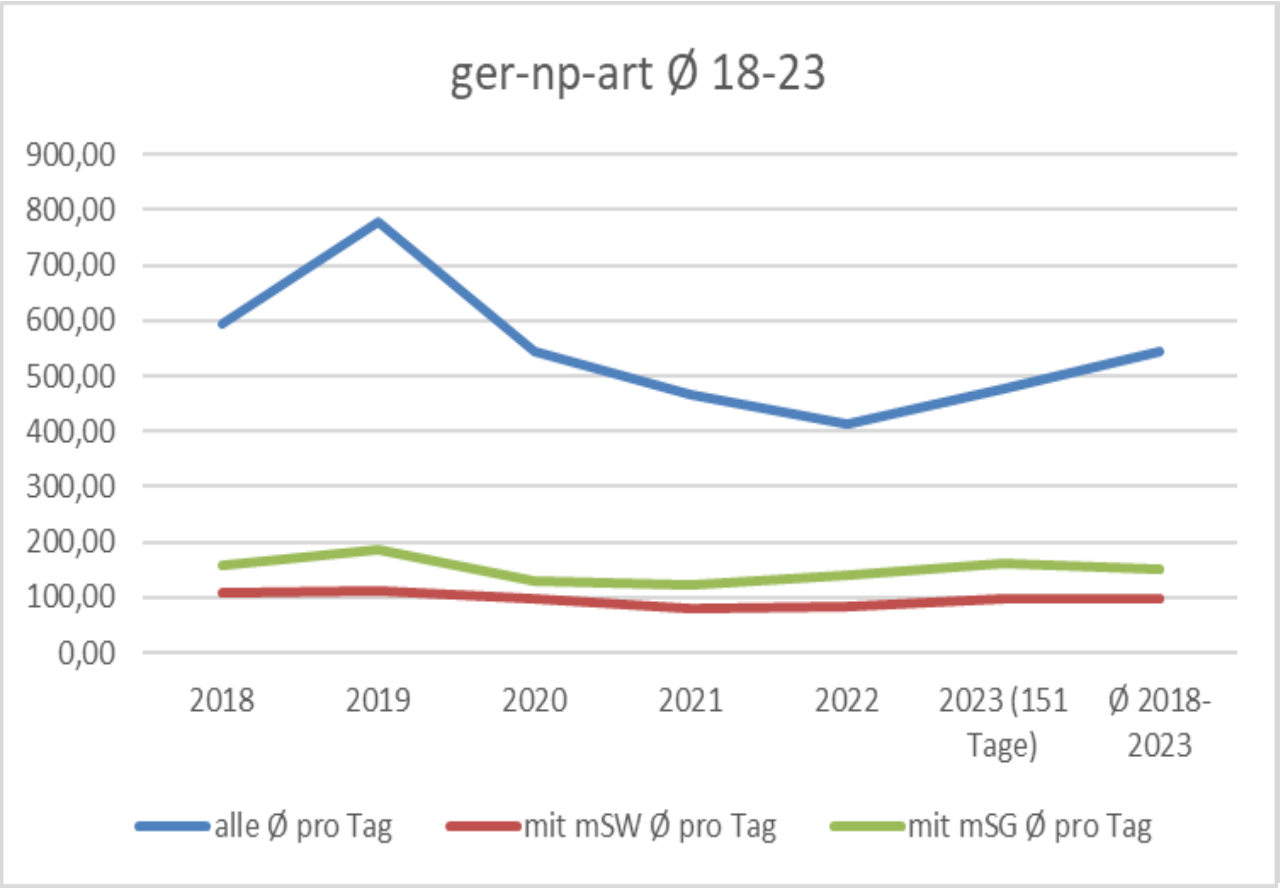
Modell	Precision	Recall	F1-Score
Ensemble	0,3777	0,5461	**0,4466**

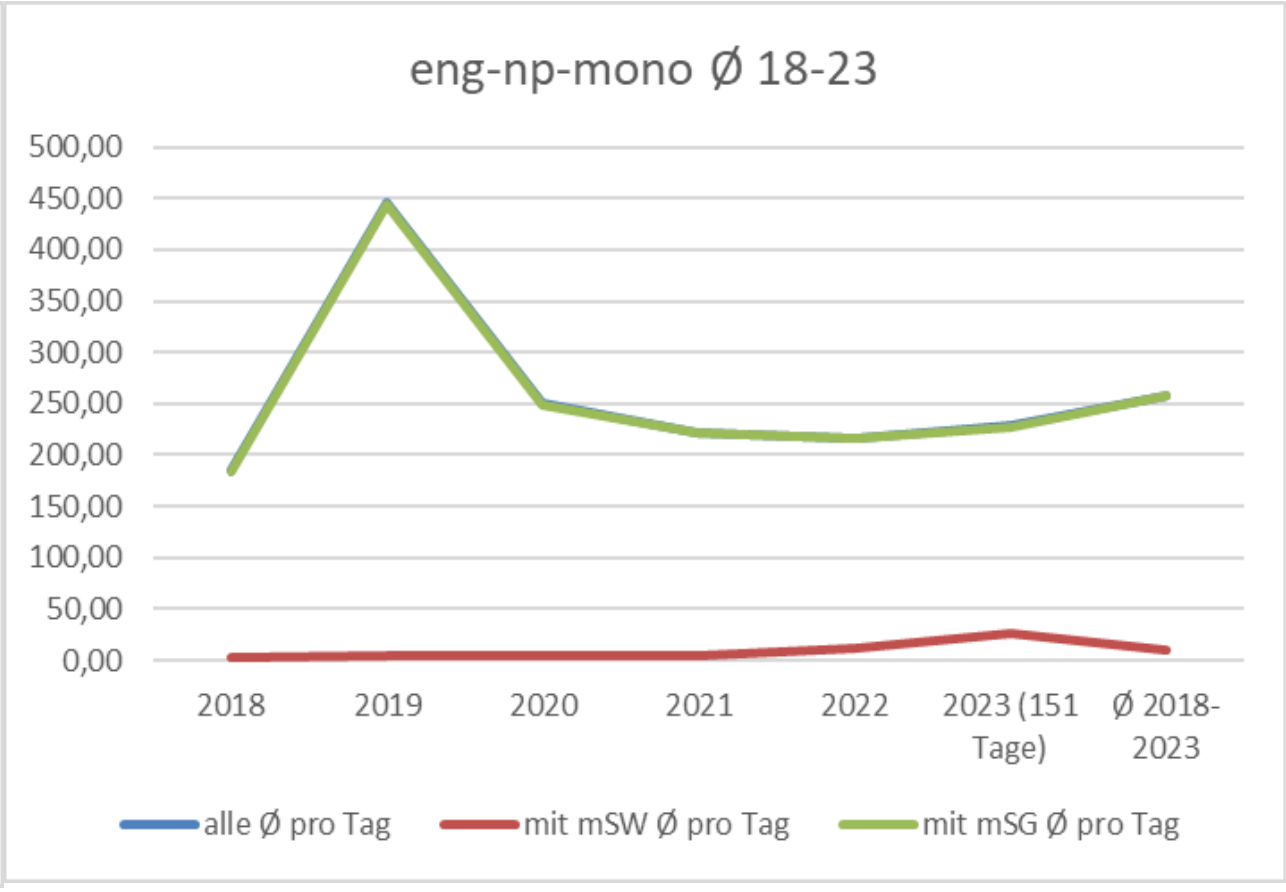
*Ausgabe pro Dokument: 7 GND-Schlagwörter (Limit) & Schwellenwert von 0,05 (Threshold)

Inhaltliche Ausweitung

- Ablösung der Anwenderfälle aus der alten Welt (Averbis Extraction Platform) durch die Erschließungsmaschine – Abschluss bis Q1/2024
- Optimierung aller produktiven Verfahren
- Ausweitung / Ausdifferenzierung Automatische Erschließung
  - alle deutschsprachigen (digitalen) Artikel
  - englischsprachige Monographien und Artikel (derzeit nur digitale Hochschulschriften)
- Verarbeitung von Texten
  - Wo funktioniert die klassische Textextraktion nicht?
  - Welche Verfahren eignen sich da – ist OCR die oder eine Lösung?

Datenanalysen als Voraussetzung





mSW Produktion IST Ø pro Tag 97, SOLL Ø pro Tag 547
mSG Produktion IST Ø pro Tag 151, SOLL Ø pro Tag 547

mSW Produktion IST Ø pro Tag 10, SOLL Ø pro Tag 249
mSG Produktion IST Ø pro Tag 258, SOLL Ø pro Tag 249

Die EMa als System modular weiter entwickeln

DEUTSCHE NATIONAL BIBLIOTHEK

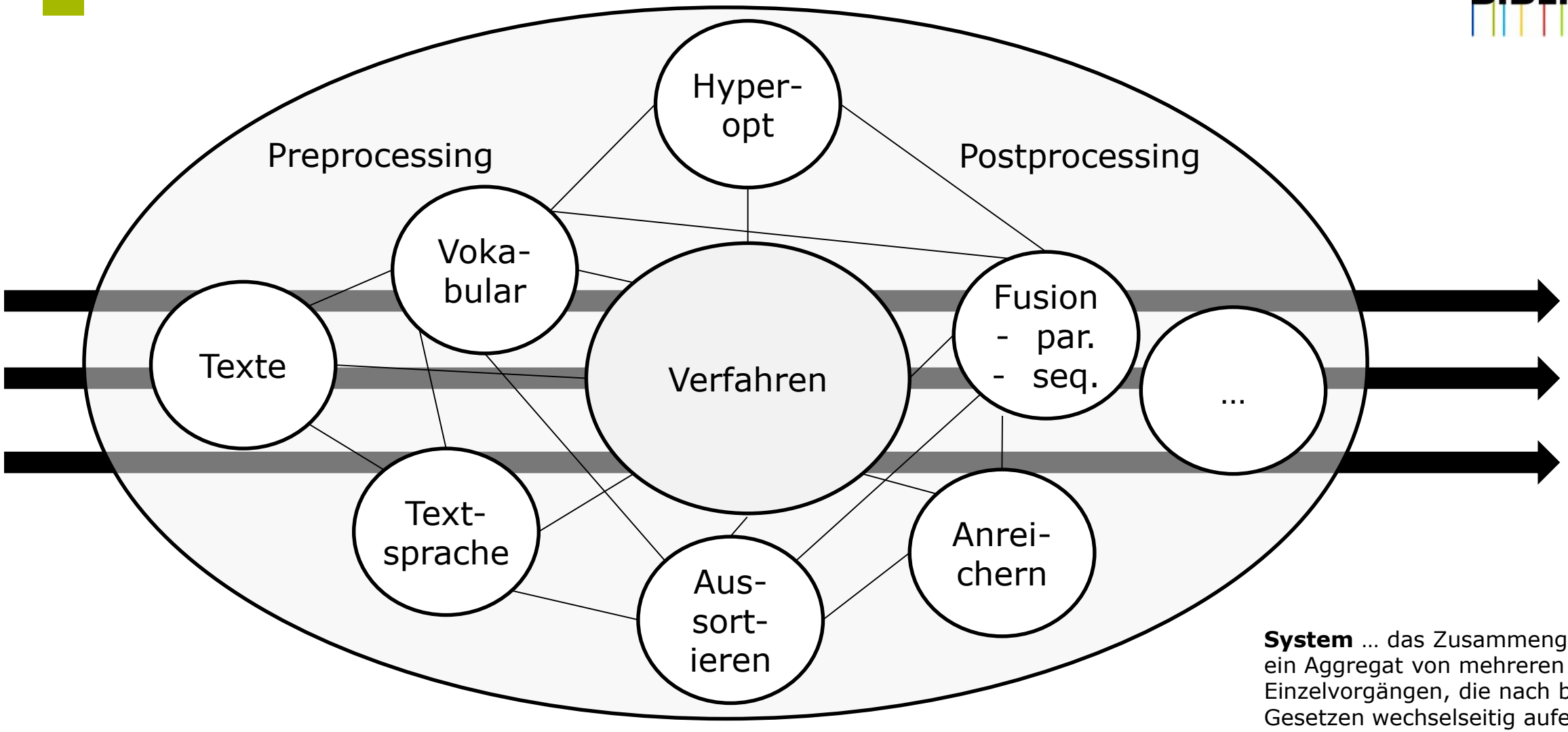
- Verschiedene Anwenderfälle
- Grundlage: Modulare Realisierung der EMa-Komponenten

- Nächstes Ziel: weitest gehende Flexibilisierung/Konfigurierbarkeit der Steuerung (Workflow Engine)

-> systematische Verbesserung / Ausdifferenzierung der Erschließungsmaschine

-> mehr Funktionalität -> Komplexität beherrschen

Die EMa als System modular weiter entwickeln



*vgl. https://dorsch.hogrefe.com/stichwort/system

System … das Zusammengestellte … ein Aggregat von mehreren Einzelvorgängen, die nach best. Gesetzen wechselseitig aufeinander wirken, also dynamisch voneinander abhängig sind, i. d. S., dass ein gemeinsamer Effekt erzielt wird.*

Die EMa und das KI-Projekt

- KI-Projekt: Qualität der automatischen Indexierung von deutschsprachigen wissenschaftlichen Netzpublikationen mit der GND durch passende Verfahren / Algorithmen messbar verbessern

KI-Projekt: Forschungslabor
- Wissenschaftliche Untersuchung von Verfahren (XMLC)
- Forschungsergebnisse
- Software (Datenmanagement, PICA-Tool, …)
- Empfehlungen für die EMa

Erschließungsmaschine: Produktiver Service
- Überführung Erkenntnisgewinne in die Produktion
- Evaluierung / Anpassung (Labor vs. Produktion)
- Softwareentwicklung / Betrieb

Fachtagung Netzwerk maschineller Verfahren in der Erschließung 2023

Thema: **„KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?“**

Aktuelle Entwicklungen im Bereich generativer KI und großer Sprachmodelle und deren Bedeutung für die Bibliotheken

- 7. und 8. Dezember 2023 in der DNB Frankfurt am Main
- https://wiki.dnb.de/x/GwfmC

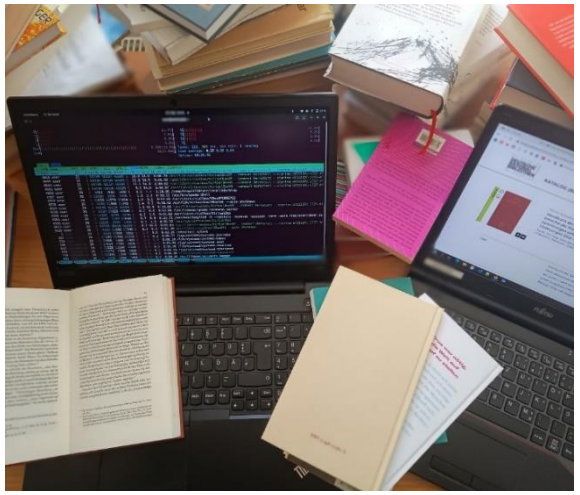
Vielen Dank!

sagen

Christoph Poley (c.poley@dnb.de)
Sandro Uhlmann (s.uhlmann@dnb.de)

… ein möglicher nächster Vortrag könnte lauten

In der Deutschen Nationalbibliothek lesen ~~jede Nacht~~ Tag und Nacht die Maschinen – Automatische Inhaltserschließung im täglichen Einsatz



Bücher und Maschinen (Symbolbild). Quelle: Sandro Uhlmann CC BY-SA 3.0 Lizenz

Weiterführende Informationen im DNB-Blog:

- Uhlmann, Sandro; Jacobs, Jan-Helge; Poley, Christoph, Schumacher, Markus: In der DNB lesen jede Nacht die Maschinen – Ein Blick auf die maschinelle Beschlagwortung.
- Nagelschmidt, Matthias: Text für Maschine.
- Mödden, Elisabeth; Schöning-Walter, Christa; Kähler, Maximilian: Texte erschließen mit KI.