DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließung in der DNB

Christoph Poley

EZB-DBIS-Konferenz
Regensburg, 06.10.2026

DEUTSCHE NATIONAL BIBLIOTHEK

Agenda

- Automatische Erschließung in der DNB
- Inhaltserschließung mit der Erschließungsmaschine (EMa)
- Metadatenextraktion mit dem Katalogisierungsagenten (KAi)

Deutsche Nationalbibliothek

Leipzig





Frankfurt

- Die zentrale Archivbibliothek in Deutschland
- In Leipzig 1913 gegründet – aktuell zwei Standorte in Leipzig und Frankfurt
- Rechtliche Grundlage bildet das Gesetz über die Deutsche Nationalbibliothek (DNBG)
- Sammeln alles, was in Deutschland oder in deutscher Sprache veröffentlicht wurde, aus dem Deutschen übersetzt wurde oder sich auf Deutschland bezieht und seit 1913 erschienen ist.

Kurzer Blick ins DNBG

DEUTSCHE NATIONAL BIBLIOTHEK

§ 2 Aufgaben, Befugnisse

Die Bibliothek hat die Aufgabe,

1. a) die ab 1913 in Deutschland veröffentlichten Medienwerke und
   b) die ab 1913 im Ausland veröffentlichten deutschsprachigen Medienwerke, Übersetzungen deutschsprachiger Medienwerke in andere Sprachen und fremdsprachigen Medienwerke über Deutschland

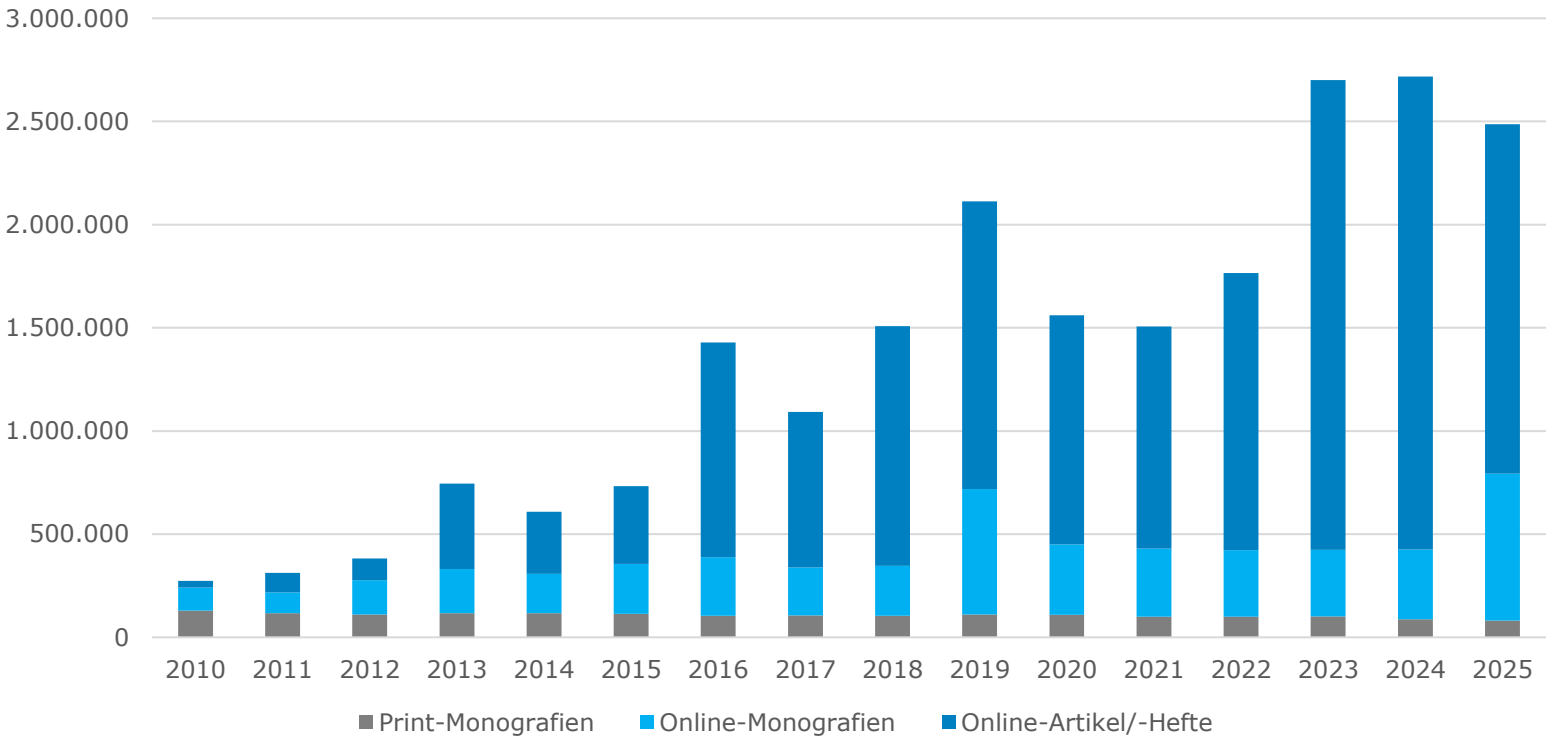
   im Original zu sammeln, zu inventarisieren, zu erschließen und bibliografisch zu verzeichnen, auf Dauer zu sichern und für die Allgemeinheit nutzbar zu machen sowie zentrale bibliothekarische und nationalbibliografische Dienste zu leisten,
2. das Deutsche Exilarchiv 1933 - 1945, die Anne-Frank-Shoah-Bibliothek sowie das Deutsche Buch- und Schriftmuseum zu betreiben,
3. mit den Facheinrichtungen Deutschlands und des Auslands zusammenzuarbeiten sowie in nationalen und internationalen Fachorganisationen mitzuwirken.

§ 3 Medienwerke

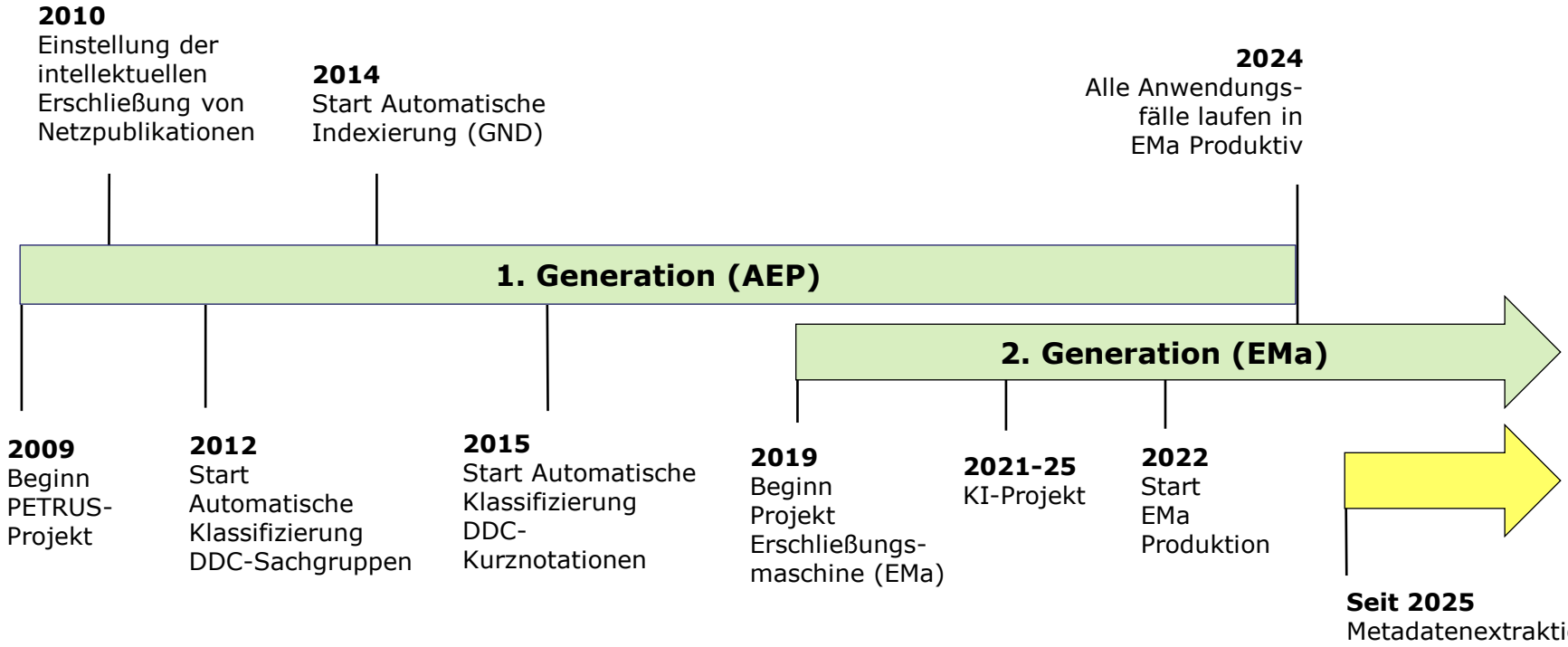
(1) Medienwerke sind alle Darstellungen in Schrift, Bild und Ton, die in körperlicher Form verbreitet oder in unkörperlicher Form der Öffentlichkeit zugänglich gemacht werden.
(2) Medienwerke in körperlicher Form sind alle Darstellungen auf Papier, elektronischen Datenträgern und anderen Trägern.
(3) Medienwerke in unkörperlicher Form sind alle Darstellungen in öffentlichen Netzen.
(4) Filmwerke, bei denen nicht die Musik im Vordergrund steht, sowie ausschließlich im Rundfunk gesendete Werke unterliegen nicht den Bestimmungen dieses Gesetzes.

Vgl: Gesetz über die Deutsche Nationalbibliothek (DNBG) https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html

Automatische Erschließung in der DNB: Motivation



Team Automatische Erschließung - Historie



Automatische Inhaltserschließung in der DNB

- von Netz- und ausgewählten Printpublikationen
- „klassische KI-Verfahren“

Automatische Klassifizierung



- DDC-Sachgruppen (ca. 100)
- DDC-Kurznotationen (ca. 2500)
- Produktiv seit 2012

- *Machinelles (supervised) Lernen*

Automatische Indexierung



- normierte Terminologie GND (>1,4 Mio)
- XMLC-Problem
- Produktiv seit 2014

- *Maschinelles Lernen*
- *Lexikalische Verfahren*

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

Lexikalische Verfahren

z.B. MLLM

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

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

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

Abgleich zwischen relevanten Termen im Text und Termen im kontrollierten Vokabular

[1]Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction.
[2]Maui-like Lexical Matching https://github.com/NatLibFi/Annif/wiki/Backend%3A-MLLM

Machinelles Lernen

z.B. Omikiji, SVM

Dok_1	Titel: „Bellen und Bellen“ • http://d-nb.info/gnd/4026181-5 Hund
Dok_2	Titel: „Schnurren, Schnurren und Schnurren“ • http://d-nb.info/gnd/4030046-8 Katze
Dok_3	Titel: „Bellen und Schnurren“ • http://d-nb.info/gnd/4026181-5 Hund • http://d-nb.info/gnd/4030046-8 Katze • http://d-nb.info/gnd/4169148-9 Maus
Dok_I	Titel: „Lorem ipsum …“ • http://d-nb.info/gnd/4030046-8 Katze • http://d-nb.info/gnd/4169148-9 Maus

Feature-Matrix (TFIDF, fiktiv)

doc_id	Bellen	Schnurren
Dok_1	0.6	0
Dok_2	0	0.7
Dok_3	0.7	0.7
Dok_I	0.3	0.9

Document-Label-Matrix

doc_id	Hund	Katze	Maus
Dok_1	1	0	0
Dok_2	0	1	0
Dok_3	1	1	1
Dok_I	0	1	1

Beim Maschinellen Lernen werden Zusammenhänge zwischen Feature-Matrix und Document-Label-Matrix gelernt.

[3]Khandagale, S., Xiao, H., & Babbar, R. (2020). Bonsai: diverse and shallow trees for extreme multi-label classification.
[4]Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., & Varma, M. (2018). Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising
[5]Omikuji Rust-Bibliothek: https://github.com/tomtung/omikuji

Automatische Inhaltserschließung in der Produktion

DEUTSCHE NATIONAL BIBLIOTHEK

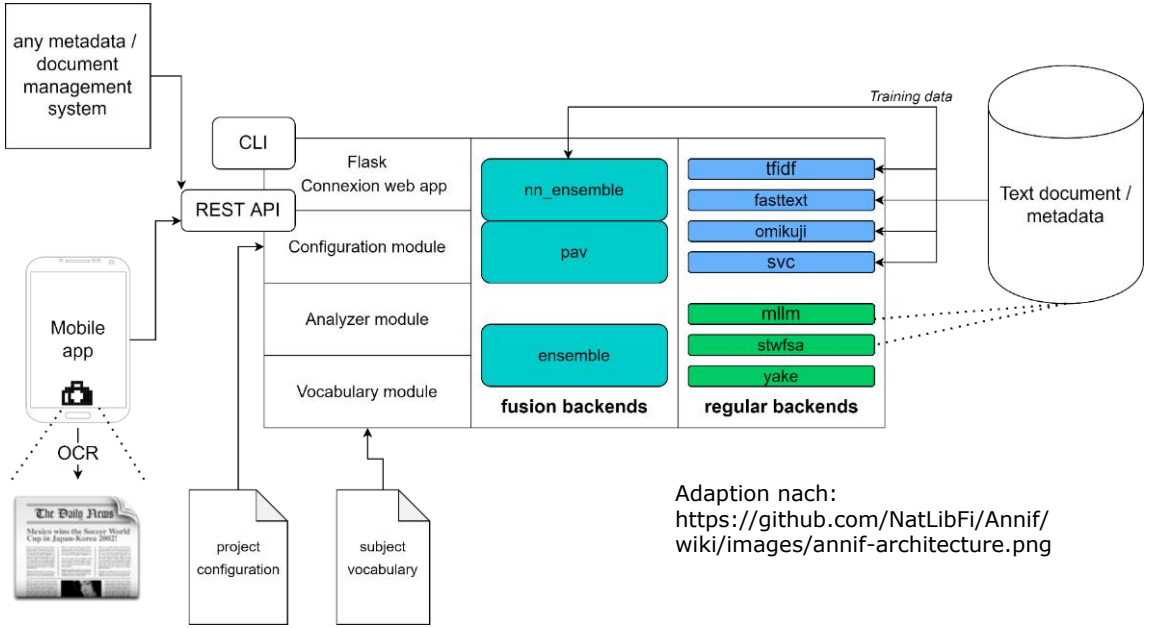
Publikationen

Inhaltserschließung	Netzpublikationen	Printpublikationen		**Anzahl automatisch erschlossener Publikationen** (Stand: 09/2026)
	Reihe O & Artikel	Reihe H	Reihe B	
 DDC-Sachgruppen	⚙	⚙	⚙	9,5 Mio.
DDC-Kurznotationen	⚙	⚙	⚙	6,6 Mio.
 GND-Schlagwörter	⚙	⚙		1,9 Mio.

annif Werkzeugkasten für automatische Klassifikation und Indexierung

DEUTSCHE NATIONAL BIBLIOTHEK

- Wird Open Source an der Finnischen Nationalbibliothek entwickelt
- Kooperative Entwicklung, die DNB ist ein Teil davon
- Unterschiedliche Schnittstellen: (client, www, REST API)
- Multilingual, kann jedes Vokabular verwenden
- Statistische Verfahren: *tfidf*, *fasttext*, *omikuji*, *svc*
- Lexikalische Verfahren: *mllm*, *stwfsa*, *yake*
- In Arbeit: *ebm, x-transformer (GPU mandatory)*
- Verfahren können allein oder fusioniert eingesetzt werden



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

Erschließungsmaschine (EMa, Produktion)

Metadaten und Texte einsammeln | Erschließen | Erschließungsdaten schreiben

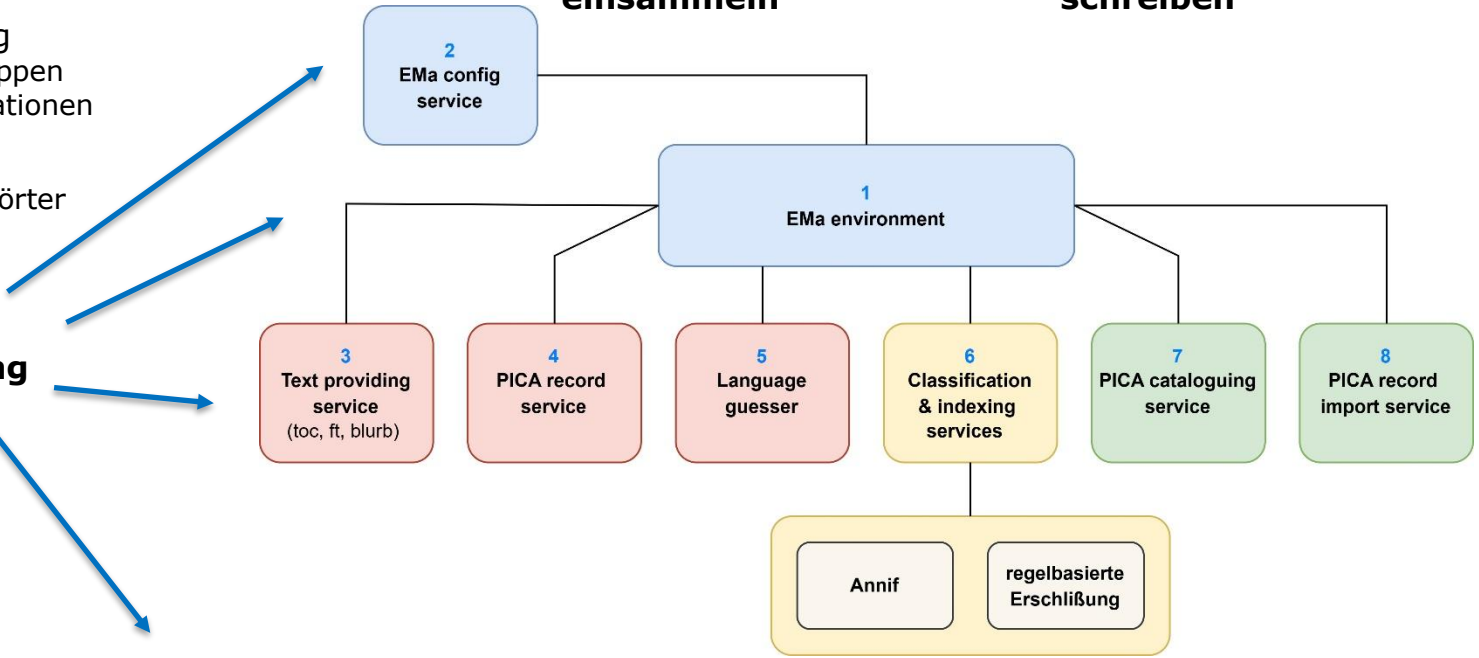


Klassifizierung
- DDC-Sachgruppen
- DDC-Kurznotationen



Indexierung
- GND-Schlagwörter

Produktentwicklung



Erzeugen, Testen, Verbessern von ML-Modellem und Regeln

Planung → **Datenmanagement** → **Modellierung**

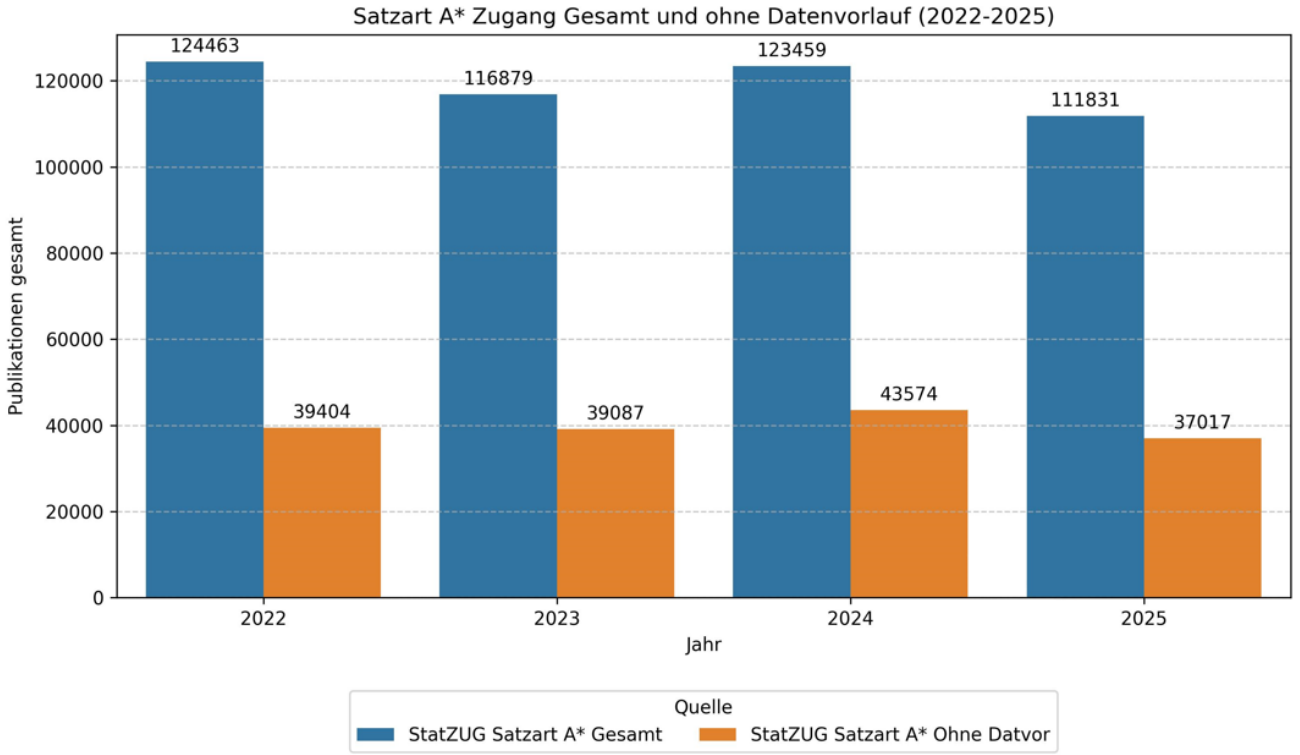
Metadatenextraktion mit dem Katalogisierungsagenten (KAi)



Die Grafik wurde mit Hilfe von KI generiert.

Metadatenextraktion ist Herausforderung in Bibliotheken weltweit.

Mengen ohne Datenvorlauf (p.a.)



KAi: Ziel/Use Case

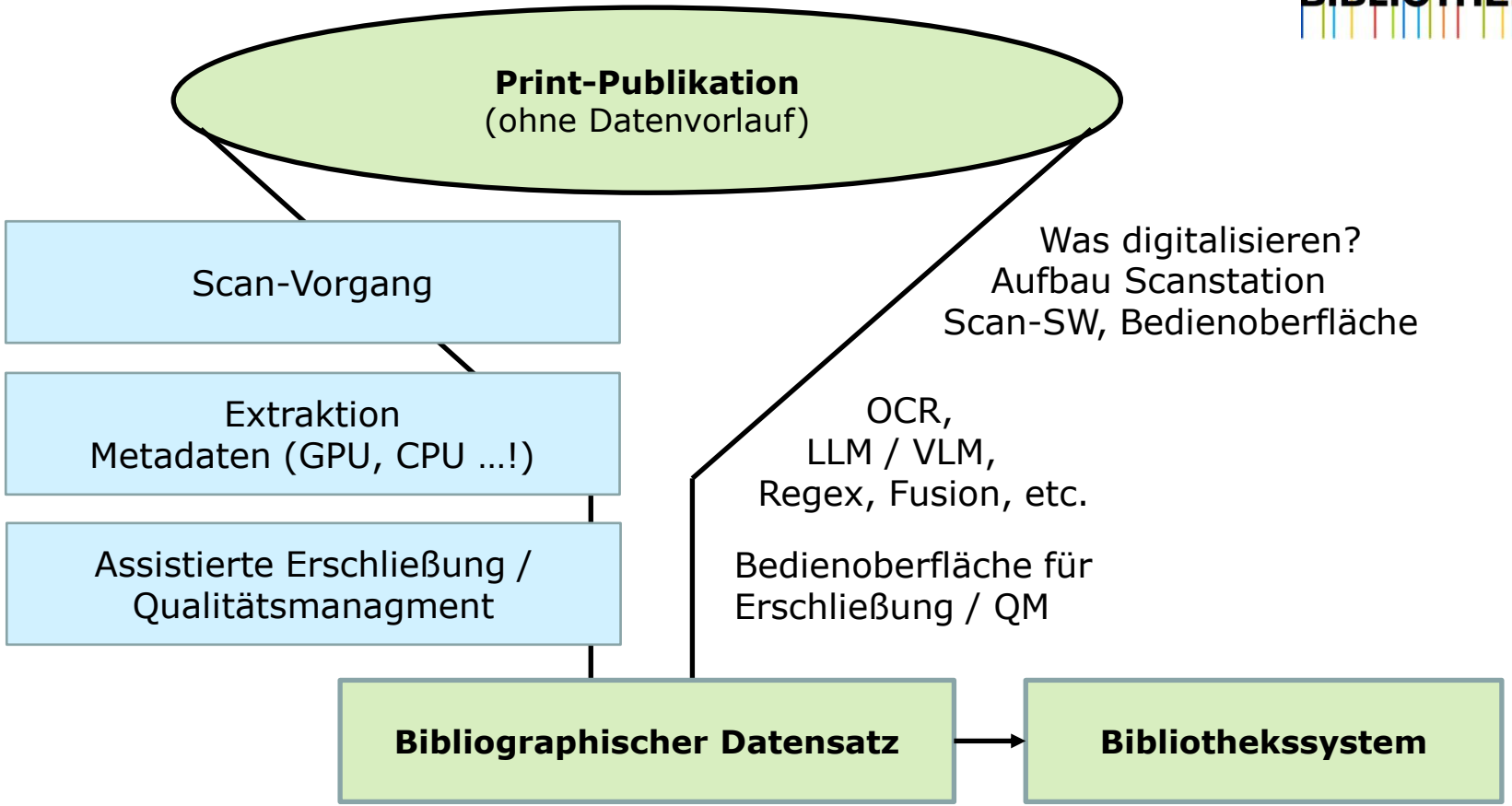
DEUTSCHE NATIONAL BIBLIOTHEK

Assistenz für Printpublikationen ohne Datenvorlauf (durch MVB, Verbünde, parallele Publikationen, etc):

- Unterstützung der
  - Erstellung von Rumpfdatensätzen im Medieneingang
  - Formalerschließung durch automatische Erschließungsvorschläge zur effizienteren Nachbearbeitung
- Frühzeitige Überführung von Büchern ins Magazin (nach Scanvorgang)

- Zunächst: Entwicklung eines prototypischen Systems, Proof of Concept

Prozesssicht Metadatenextraktion



Framework BIBRA

- BIBRA ist ein Open-Source-Werkzeugkasten zur automatischen Extraktion bibliografischer Metadaten aus Dokumenten
- Entwickelt von der Finnischen Nationalbibliothek
- Web-Oberfläche, Client-Interface, REST API



Qualitätsbewertung Metadatenextraktion

DEUTSCHE NATIONAL BIBLIOTHEK

Feld	Typ	Hauptmetrik	Was bedeutet das?	Besonders wichtig weil…
Haupttitel	**Text**	Zeichenfehlerrate (CER)	Wie viele Zeichen sind richtig? 100 % = perfekt	Grundlage aller Prozesse
Titelzusatz	**Text**	Zeichenfehlerrate (CER)	Wie viele Zeichen sind richtig? 100 % = perfekt	Inhaltserschließung, Discovery
Geistiger Schöpfer	**Text**	Zeichenfehlerrate (CER)	Wie viele Zeichen sind richtig? 100 % = perfekt	GND-Verknüpfung, VG Wort
Ausgabebezeichnung	**Exakt**	Exact Match (normalisiert)	„2. Auflage" = „Zweite Auflage"? → Ja/Nein	Deduplizierung, Ausgabenunterscheidung
Verlagsname	**Exakt**	Exact Match (normalisiert)	Stimmt der Verlag? → Ja/Nein	Zitation, Fremddatenübernahme
Erscheinungsdatum	**Exakt**	Exact Match (Jahreszahl)	Ist die Jahreszahl korrekt? → Ja/Nein	Chronologie, Nationalbibliografie
Umfang (Seiten)	**Exakt**	Exact Match (Zahl)	Stimmt die Seitenzahl? → Ja/Nein	Zitation, Vollständigkeit
Reihe + Zählung	**Exakt**	Exact Match (Nr.)	Wurde Reihe erkannt? Ist die Bandnummer korrekt? → Ja/Nein	Werkbündelung
Halluzinationen (alle Felder)	**Risiko**	Halluzinationsrate (%)	Anteil erfundener Felder an allen extrahierten Feldern	LLM-spez. Risiko: erfundene Werte schlimmer als leere Felder

≈ Ähnlichkeitsvergleich (wie ähnlich ist der Text?) ✓ Exakte Übereinstimmung (Richtig / Falsch) ⚠ Halluzination (Feld erfunden, obwohl nicht vorhanden)

Vielen Dank für Ihre Aufmerksamkeit!

Dipl.-Inf. Christoph Poley

Deutsche Nationalbibliothek
Teamleiter Automatische Erschließung
Deutscher Platz 1
04103 Leipzig

c.poley@dnb.de