Dipl.-Inf. Christoph Poley

Hardware-Ressourcen für die automatische Erschließung in der DNB

Eine Analyse mit Blick in die Zukunft

Motivation

Große Tech-Konzerne wie Alphabet und Meta haben schier unendliche Ressourcen und betreiben Datenzentren, die über die gesamte Welt verteilt sind.

Und wir haben uns – Bibliotheken.

Auch wir haben Hardware in unserem Haus, zum Beispiel für die automatische Erschließung.

Wolfgang Stille: hessian.AI, H100-Cluster „43“

Die Deutsche Nationalbibliothek

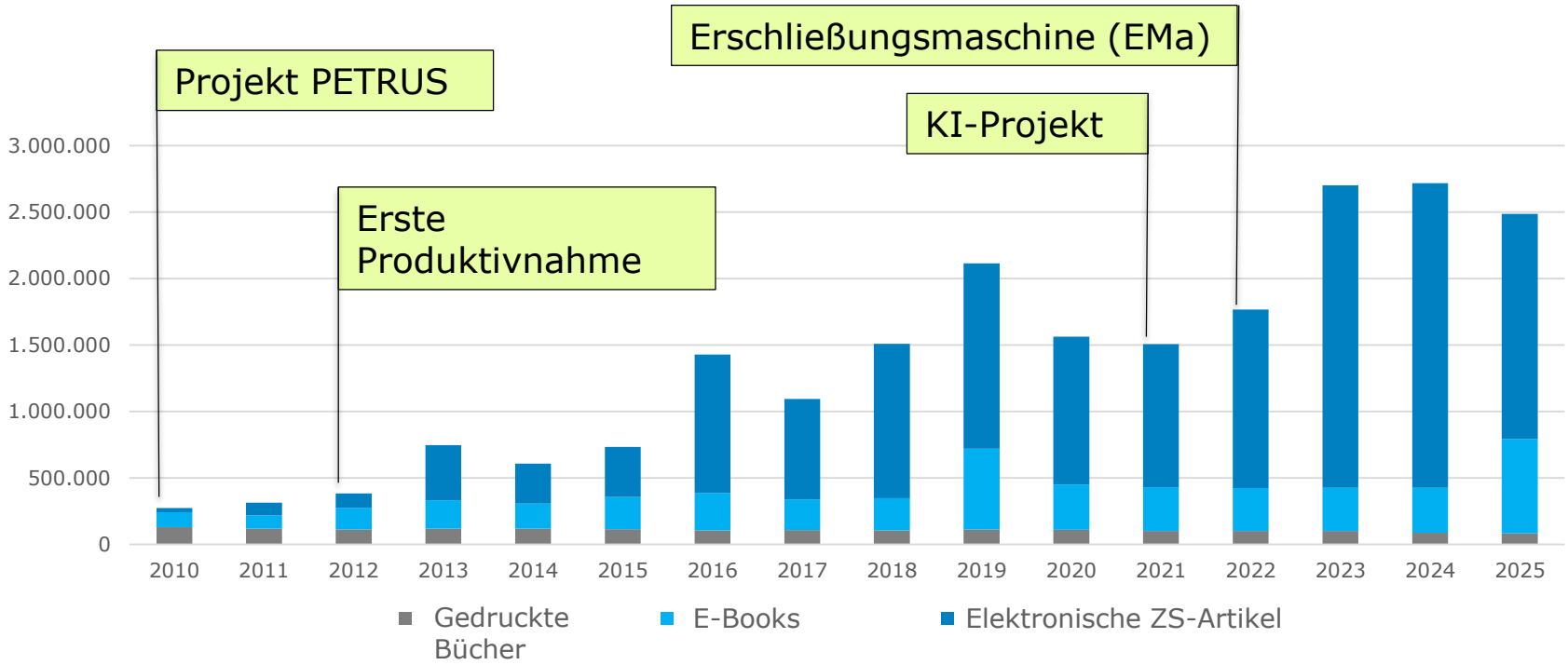
Frankfurt

Leipzig

- Zentrale Archivbibliothek und das nationalbibliografische Zentrum der Bundesrepublik Deutschland.
- 1912 in Leipzig gegründet – gegenwärtig zwei Standorte in Leipzig und Frankfurt
- Grundlage: Gesetz über die Deutsche Nationalbibliothek (DNBG)
- Sammelauftrag umfasst alle Publikationen in Schrift, Bild und Ton, die seit 1913 in Deutschland, in deutscher Sprache, als Übersetzung aus der deutschen Sprache oder über Deutschland veröffentlicht wurden.

*DNBG: https://www.gesetze-im-internet.de/dnbg

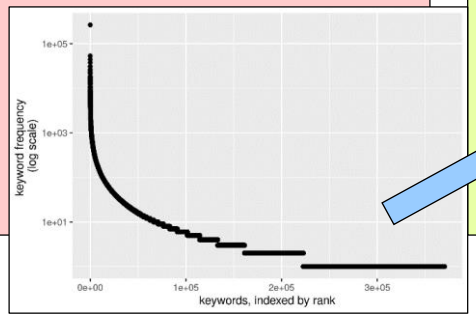
Abgelieferte Print- und Netzpublikationen an die DNB seit 2010



Fachliche Use Cases

Automatische Klassifikation

- DDC Sachgruppen (100)
- DDC Kurznotationen (~ 2500)
- Maschinelle Lernverfahren
- Seit 2012 produktiv
- 2025: ~ 1.3m Publikationen

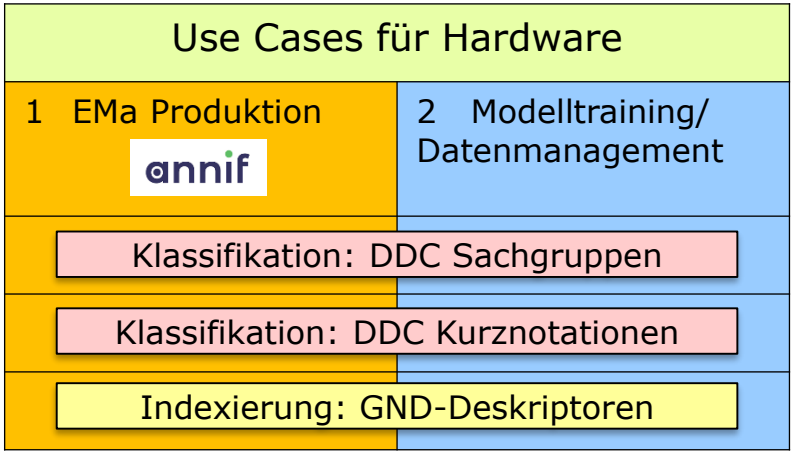


Automatische Indexierung

- Normierte GND-Terminologie (> 1,4m)
- XLMC-Problem
- Überwachte Lernverfahren
- Lexikalische Ansätze (MLLM)
- Seit 2014 produktiv
- 2025: ~ 260,000 Publikationen

Hardware-Use Cases 1 und 2

- Abhängigkeit (Dimensionierung RAM, CPU, SSD (NetApp)…) von folgenden Faktoren, z.B.:
  - Methoden: Überwachtes Lernen, lexikalische Verfahren, Kombination (Ensemble, Fusion der Methoden)
  - Use Cases der Automatischen Inhaltserschließung: Größe des Vokabulars und der Trainingsdaten
  - Anzahl der Requests für Erschließung
  - Erwartungshaltung gegenüber der Antwortzeit



- Aktuell: VMWare-Server ohne GPU
- Hardware für andere Services der EMa: auf Open-Shift-Cluster

Hardware-Use : 3 Forschung

Projekt “Automatisches Erschließungssystem” (KI-Projekt):

- Gefördert im Rahmen der KI-Strategie der BRD Deutschland*
- Dauer: ca. 4 Jahre (Oktober 2021 – Dezember 2025)
- Verbesserung der automatischen Inhaltserschließung mit innovativen Methoden des Künstlichen Intelligenz (KI) – mit Fokus auf die Indexierung deutschsprachiger wissenschaftlicher Literatur mit GND-Deskriptoren
- Bereitstellen von neuen Methoden für die automatische Indexierung (GND)
- Wissenstransfer zu den Produkten der DNB
- Finden eines Gleichgewichts zwischen Indexierungsqualität und Hardwareanforderungen /-beschränkungen

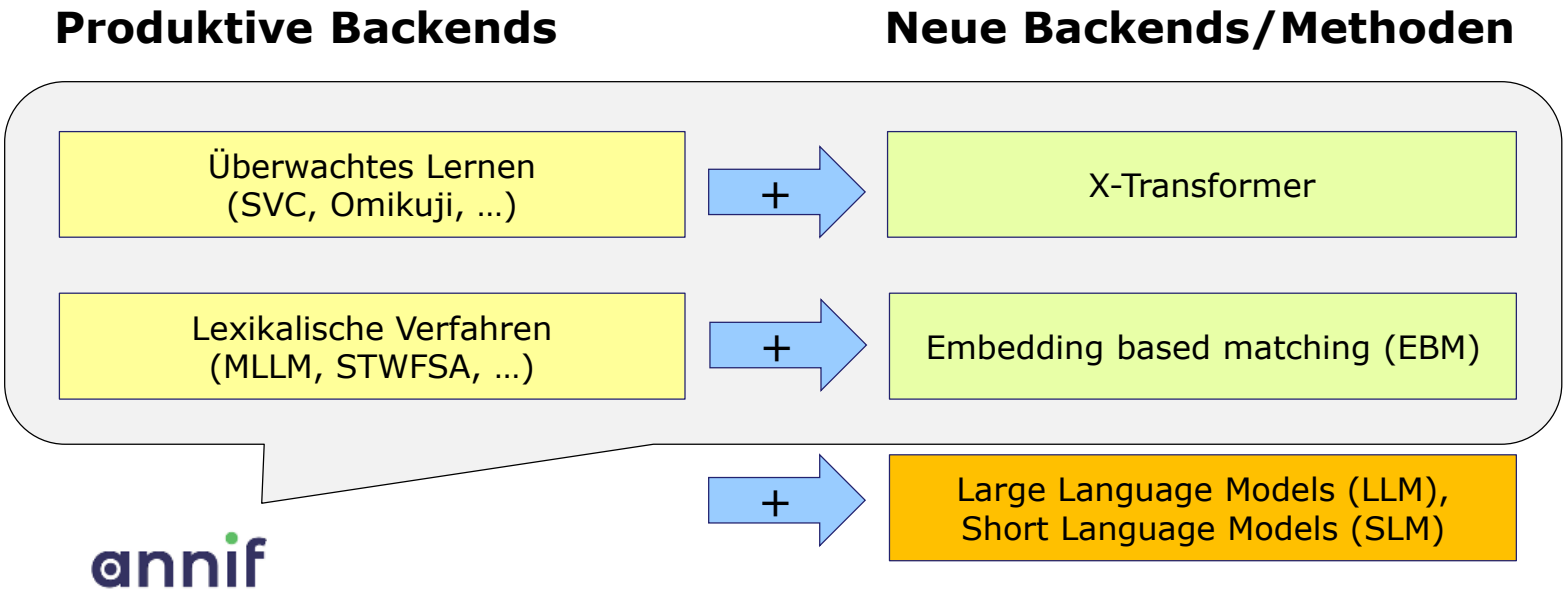
Der Beauftragte der Bundesregierung für Kultur und Medien

Themen mit Bezug zur Hardware:

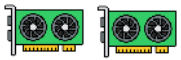
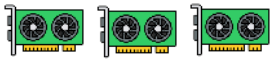
- Wie muss die Hardware in Zukunft dimensioniert sein? Wo sind die Grenzen des Leistbaren? Wo sind sinnvolle Grenzen?
- Sammeln von Erfahrungen im Umgang mit GPU
- Nutzung externer HPC-Infrastrukturen (ZIH Dresden, Job-Verarbeitung)

*BKM: Der Beauftragte der Bundesregierung für Kultur und Medien
https://www.kulturstaatsminister.de/

Neue Verfahren erfordern leistungsfähigere Hardware



Zusammenfassung: Hardware-Ressourcen

Hardware-Use cases	1. Produktion		2. Modelltraining/ Datenmanagement	
	3. Forschung			
Methoden	Kosten	GPU?	Kosten	GPU?
Überwachtes Lernen (SVM, Omikuji, …)	🟢🟢		🔴	
Lexical Matching (MLLM)	🟢		🟢	
Embedding based Matching (EBM)	🟡		🟡	
X(R)-Transformer	🟡		🔴🔴	
LLM/SLM	🔴🔴		🔴🔴🔴	

GPU Bild: Flaticon.com

DEUTSCHE NATIONAL BIBLIOTHEK

Effektivität

Nicht benötigte Hardware ist die beste Hardware.

Grundlegend: “Spontane” Hardwarekäufe sind unüberlegt.

Neue Hardware benötigt viel Energie (während Herstellung / Betrieb) – CO_2-Fußabdruck (A100: bis zu 400W, H100/H200: bis zu 700W)

Gezielte Optimierung von Methoden und Daten trägt zu einer verantwortungsvollen Nutzung von Hardware-Ressourcen bei, z.B.:

- Optimierung der Leistungsfähigkeit der Verfahren (Verarbeitungszeit MLLM)
- Reduzierung der Größe der Modelle (Optimierung des Vokabulars, ...)
- Flaschenhälse in den Workflows beseitigen (z.B. durch die getrennte Bereitstellung von Texten/Metadaten vs. Verarbeitung)
- …

Exkurs: Optimierung anhand zweier Beispiele (GND NP)

(reine Verarbeitungszeit, Textlänge jeweils ca. 30.000 Zeichen)

Produktion Ensemble mit Annif 1.0 Dauer: 1h, 0 min, 48s 3,64s pro IDN ca. 16 IDN/min	Opt. MLLM	Produktion Ensemble mit Annif 1.3 Dauer: 0h, 27 min, 40s 1,66s pro IDN ca. 36 IDN/min	kleinere Modelle	Produktion Ensemble mit Annif 1.3 Dauer: 0h, 14 min, 46s 0,87s pro IDN ca. 69 IDN/min
Anzahl x pro s - 0: - - 1: 3 - 2: 129 - 3: 683 - 4:93 - 5: 9 - 6-10: 65 - 11-107: 15		Anzahl pro s - 0: - - 1: 412 - 2: 568 - 3: 11 - 4: 4 - 5: - - 6-10: - - 11-14: 5		Anzahl pro s - 0: 226 - 1: 740 - 2: 21 - 3: 4 - 4: 8 - 5: 1 - >= 6: -

Ausblick: Auf dem Weg zu einem Hardware-Konzept für die Automatische Erschließung an der DNB

Die Hardware für die Automatische Erschließung ist Teil der DNB-Hardware und kein separates Silo.

Hardware-Anforderungen und Kosten steigen stark. Sie zu beherrschen, benötigt spezielle Fachkenntnisse.

Welche Dimension an Infrastruktur kann ich mir in Zukunft leisten (Beschaffung/ Wartung)?

Das Erstellen eines Hardwarekonzeptes kann komplex werden. Das Sammeln von Parametern (Erfahrungen aus Produktion und Forschung) hilft dabei.

Erstes Ziel: Erfüllen der Anforderungen für die nächsten 2-3 Jahre.

Beschaffung von zwei Servern (+GPU):
(1) für die Produktion
(2) Für Training/Datenmanagement

Mittelfristiges Hardware-Konzept

- Hardware- und Cloudstrategie als ein Teil der IT- und Bibliotheksstrategie denken
- Produkte, die entwickelt bzw. produktiv betrieben werden sollen, sowie Ressourcen müssen kommuniziert und priorisiert sein
- Ausgewogene Vorgehensweise um Erreichen von Zielen mit den vorhandenen Ressourcen
- Kontinuierlicher und iterativer Prozess
- Parallel dazu: Reduzieren der Hardware-Anforderungen und Hinterfragen der Notwendigkeit teurer Hardware

Beispiel: Betrieb auf On-Premises-Hardware:

- Wo ist On-Premises obligatorisch, wo sinnvoll? Leitgedanke @DNB: EMa-Produktion obligatorisch On-Premises
- Alternativen: Externe Server, Nutzung HPC, Cloud-Services (SaaS)
  - Sind gewöhnlich (auch) nicht kostenfrei.
  - Klärung des rechtsicheren Umgangs mit lizenzierter Literatur und Datensouveränität.

Literatur (Auswahl)

- German National Library, Annual Report 2024. (2025) https://jahresbericht.dnb.de/Webs/jahresbericht/EN/2024/Home/home_node.html
- Kasprzik, A. (2024): Die KI(irche) im Dorf lassen, O-BIB., DOI: https://doi.org/10.5282/o-bib/6201
- Kähler, M., Rietdorf, C. (2025): Embedding based matching for Automated Subject Indexing. https://github.com/deutsche-nationalbibliothek/ebm4subjects
- Poley, C. et al. (2025): Automatic Subject Cataloguing at the German National Library. LIBER Quarterly: The Journal of the Association of European Research Libraries, 35(1), 1-29. https://doi.org/10.53377/lq.19422
- Suominen, O. (2019): Annif - DIY automated subject indexing using multiple algorithms. LIBER Quarterly, 29(1), 1-25. https://doi.org/https://doi.org/10.18352/lq.10285
- Suominen, O. (2024): Building Civilized AI, https://docs.google.com/presentation/d/e/2PACX-1vRA1o11pODoJ0FmFc8dRj-xNZRUs7lsxzDACkiYt6d-Bdfql1ujw3gGpSedTQnXDG0MrRg3_WAl1GQS/pub
- Fu, Yuankun (2025): LLM Inference Sizing and Performance Guide: https://blogs.vmware.com/cloud-foundation/2024/09/25/llm-inference-sizing-and-performance-guidance/

Vielen Dank!

Dipl.-Inf. Christoph Poley

Deutsche Nationalbibliothek
Teamleiter Automatische Erschließung
Deutscher Platz 1
04103 Leipzig

mailto:c.poley@dnb.de