Matthias Nagelschmidt

Evaluation von Annif für die maschinelle Inhaltserschließung an der Deutschen Nationalbibliothek

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

DEUTSCHE NATIONAL BIBLIOTHEK

Projekt Erschließungsmaschine

Projekttreiber:

Produktiv eingesetzte Erschließungssoftware der Fa. Averbis wird nicht weiterentwickelt,

Legacy-System soll noch für drei bis fünf Jahre eingesetzt werden.

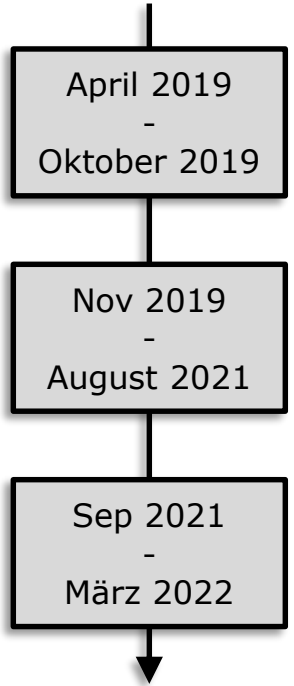
Projektiert ist die Planung und Implementierung eines neuen Erschließungssystems.

Projekt Erschließungsmaschine

Phase 1:
Anforderungsformulierung, Marktsichtung

Phase 2:
Evaluation, Anschaffung, Implementierung

Phase 3:
Produktivnahme, Ablösung des Legacy-Systems durch neues System

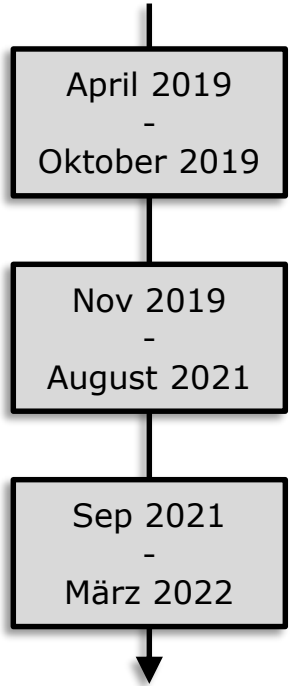


Projekt Erschließungsmaschine

Phase 1:
Anforderungsformulierung, Marktsichtung

Phase 2:
Evaluation, Anschaffung, Implementierung

Phase 3:
Produktivnahme, Ablösung des Legacy-Systems durch neues System



DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Modulares, Open Source-basiertes System zur automatischen, verbalen und klassifikatorischen Erschließung, entwickelt an der Finnischen Nationalbibliothek.

annif

Verschiedene Text Mining- und KI-basierte Verfahren als Backends implementiert. Backends werden nicht von den Annif-Entwicklern bereitgestellt, sondern dezentral entwickelt (Projektrisiko?).

1640
THE NATIONAL LIBRARY OF FINLAND

DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Einfacher, statistischer Ansatz: Abzählen der Auftretenshäufigkeiten von Termen im Dokument und ins Verhältnis setzen von Auftretenshäufigkeiten der gleichen Terme im Trainingskorpus. Implementiert durch die Python-Bibliothek „Gensim“.

- **TF-IDF**
- fastText
- Omikuji
- Maui
- Ensemble
- nn_Ensemble
- PAV

DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Geometrischer, vektorbasierter Ansatz: Aufbau eines hochdimensionalen Vektorraum-Modells aus intellektuell erschlossenen Trainingsdaten zur Ähnlichkeitsmessung von Term-Vektoren aus Dokumenten. Entwickelt durch Facebook Research.

- TF-IDF
- **fastText**
- Omikuji
- Maui
- Ensemble
- nn_Ensemble
- PAV

DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Probabilistischer, entscheidungsbaum-basierter Ansatz: Extreme Multilabel Classification ermöglicht das „Lernen" einer hohen Anzahl von „Labels" (bzw. Klassen, Schlagwörter). Verschiedene algorithmische Varianten liegen als Rust-Implementierungen vor.

- TF-IDF
- fastText
- **Omikuji**
- Maui
- Ensemble
- nn_Ensemble
- PAV

DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Linguistischer, morphologisch-semantischer Ansatz: Extrahierte N-Gramme werden verschiedenen morphologischen, quantitativen und semantischen Analysen unterzogen, woraus sich Rückschlüsse auf deren Relevanz für das zugrundeliegende Dokument ergeben. Anschließend erfolgt ein Mapping auf kontrolliertes Vokabular. Implementiert als REST-basierter Microservice.

- TF-IDF
- fastText
- Omikuji
- **Maui**
- Ensemble
- nn_Ensemble
- PAV

DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Einfache Ergebniskombination: Ergebnisse verschiedener Backends werden kombiniert, indem für jedes Einzelergebnis eine durchschnittliche Punktzahl vergeben wird und anschließend die „besten“ Ergebnisse ausgewählt werden.

- TF-IDF
- fastText
- Omikuji
- Maui
- **Ensemble**
- nn_Ensemble
- PAV

DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Dynamische Ergebniskombination: Ergebnisse verschiedener Backends werden kombiniert, indem die verschiedenen Einzelergebnisse durch ein künstliches neuronales Netz neu gewichtet werden. Im Initialzustand entspricht das Netz dem Verhalten des einfachen Ensemble-Backends, so dass zunächst ein Trainieren und ggf. ein kontinuierlich fortzuführendes „Lernen" des Modells erforderlich werden. Implementiert durch die Python-Bibliotheken „Keras" und „TensorFlow".

- TF-IDF
- fastText
- Omikuji
- Maui
- Ensemble
- **nn_Ensemble**
- PAV

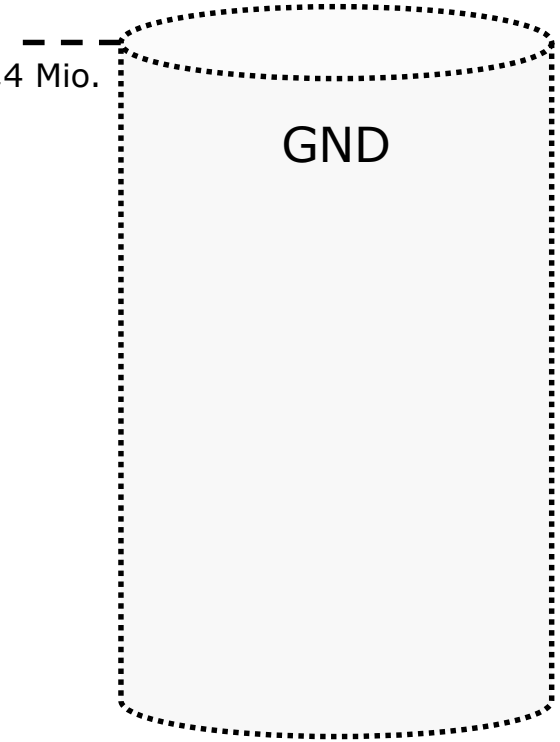
DEUTSCHE NATIONAL BIBLIOTHEK

Annif: Automated Subject Indexing Toolkit

Dynamische Ergebniskombination: Ergebnisse verschiedener Backends werden kombiniert, indem sie durch ein Regressionsverfahren neu gewichtet und die für jedes Ergebnis vergebenen Punkte durch einen probabilistischen Wert ersetzt werden. Implementiert durch die etalbierte Python-Bibliothek „scikit learn“.

Aufbau der Trainings- und Testkorpora

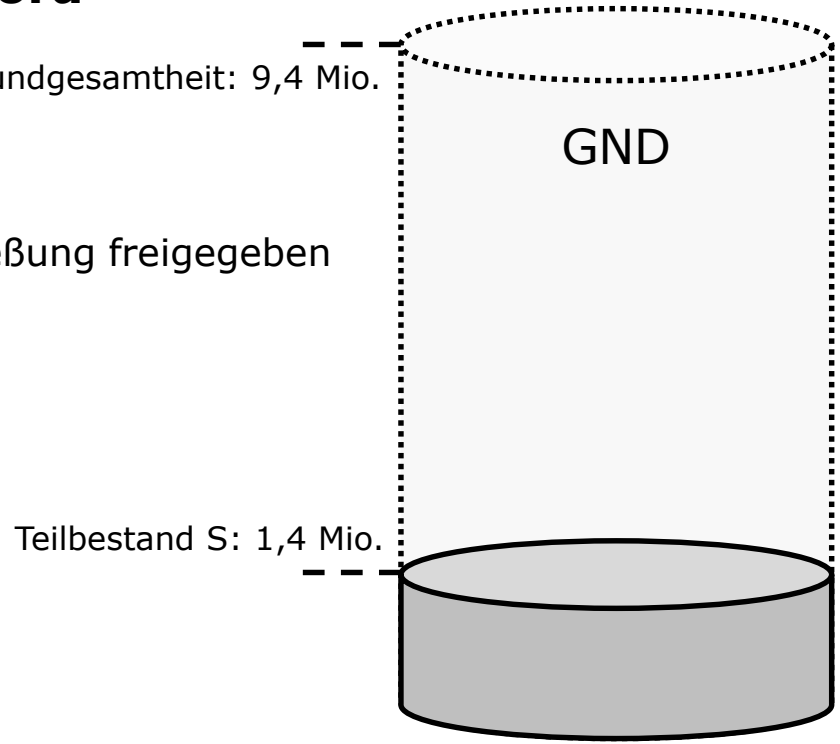
9,4 Mio. Normdatensätze in der GND



Aufbau der Trainings- und Testkorpora

9,4 Mio. Normdatensätze in der GND

1,4 Mio. Normdatensätze für die inhaltliche Erschließung freigegeben (Teilbestand S)

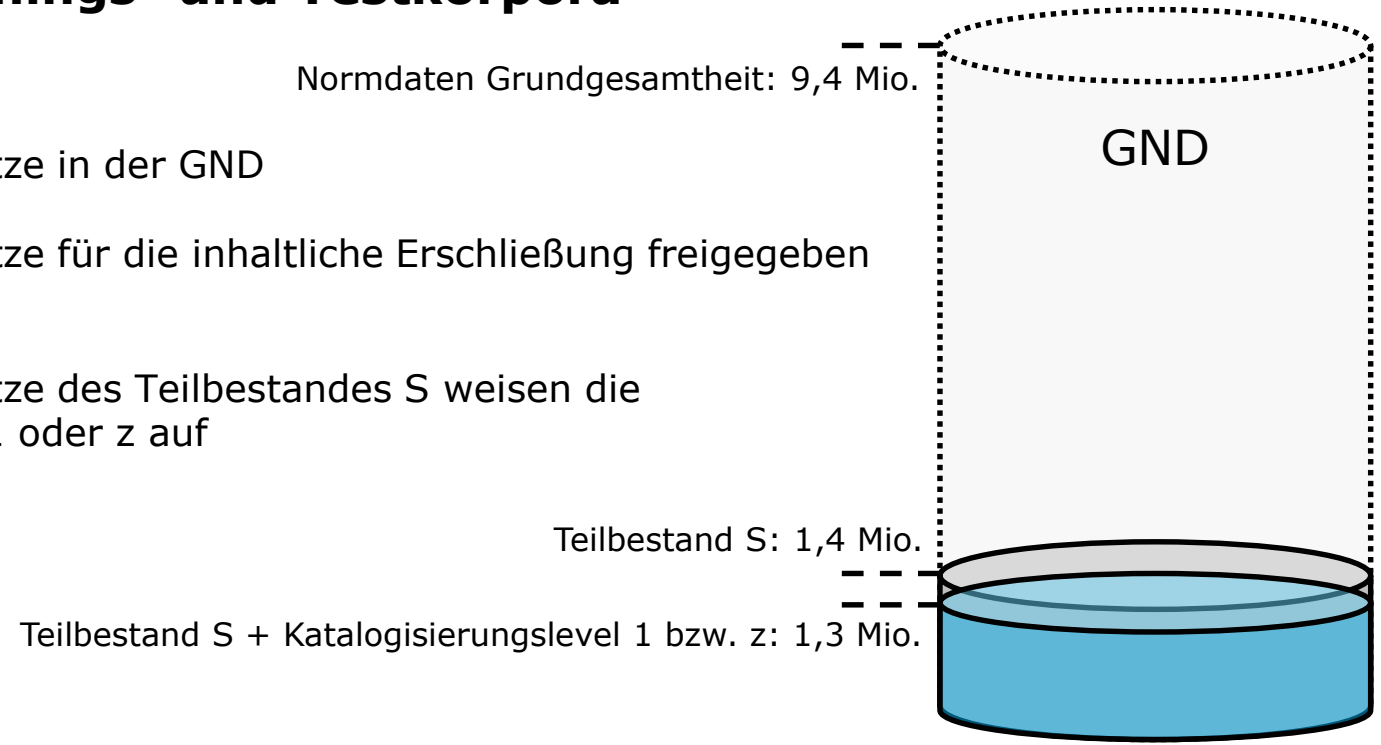


Aufbau der Trainings- und Testkorpora

9,4 Mio. Normdatensätze in der GND

1,4 Mio. Normdatensätze für die inhaltliche Erschließung freigegeben (Teilbestand S)

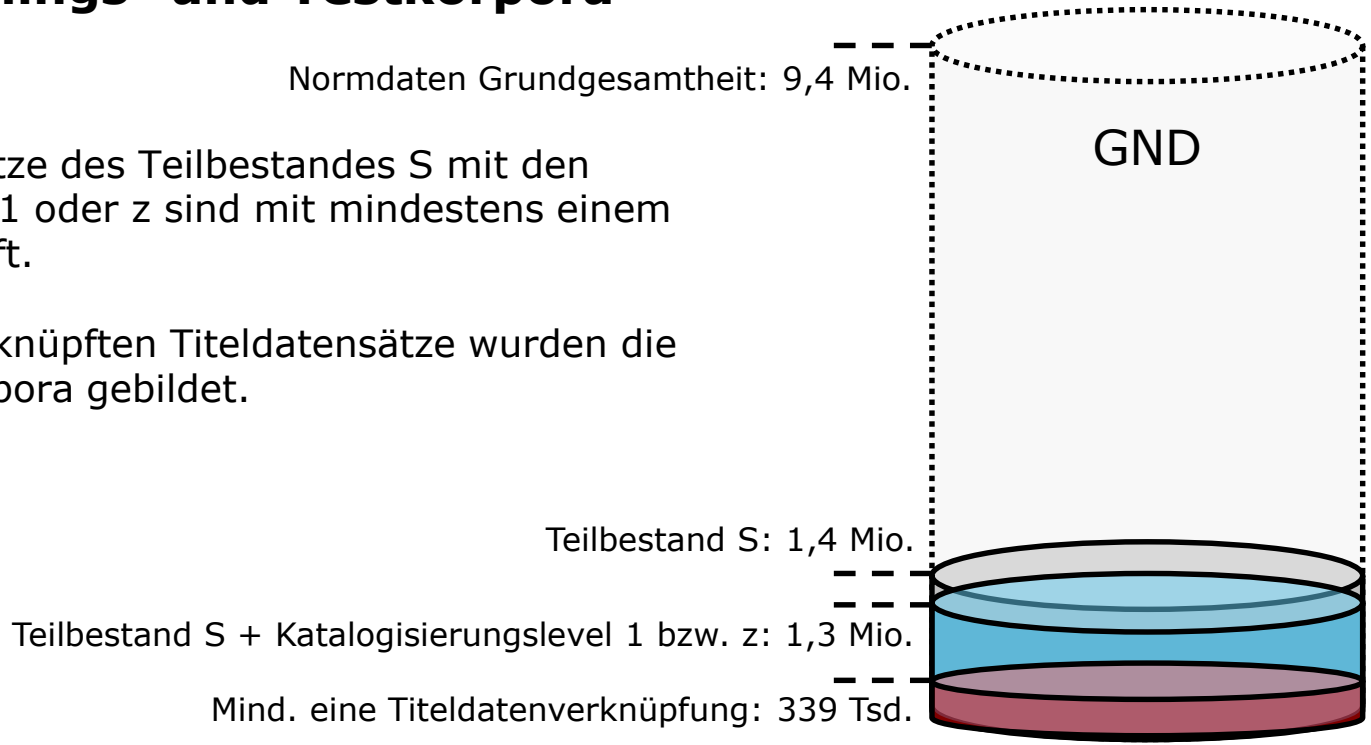
1,3 Mio. Normdatensätze des Teilbestandes S weisen die Katalogisierungslevel 1 oder z auf



Aufbau der Trainings- und Testkorpora

339 Tsd. Normdatensätze des Teilbestandes S mit den Katalogisierungsleveln 1 oder z sind mit mindestens einem Titeldatensatz verknüpft.

Aus der Menge der verknüpften Titeldatensätze wurden die Trainings- und Testkorpora gebildet.



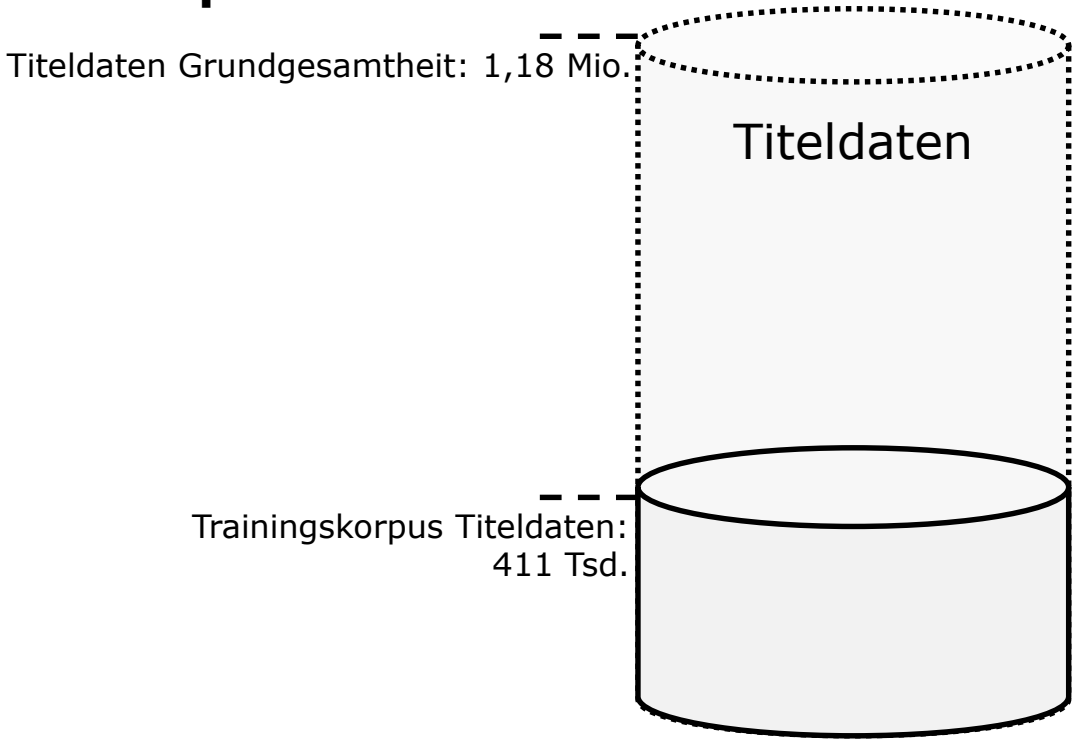
Aufbau der Trainings- und Testkorpora

1,18 Mio. Titeldatensätze
als Grundgesamtheit

Aufbau der Trainings- und Testkorpora

1,18 Mio. Titeldatensätze
als Grundgesamtheit

411 Tsd. Titeldatensätze
als Trainingskorpus

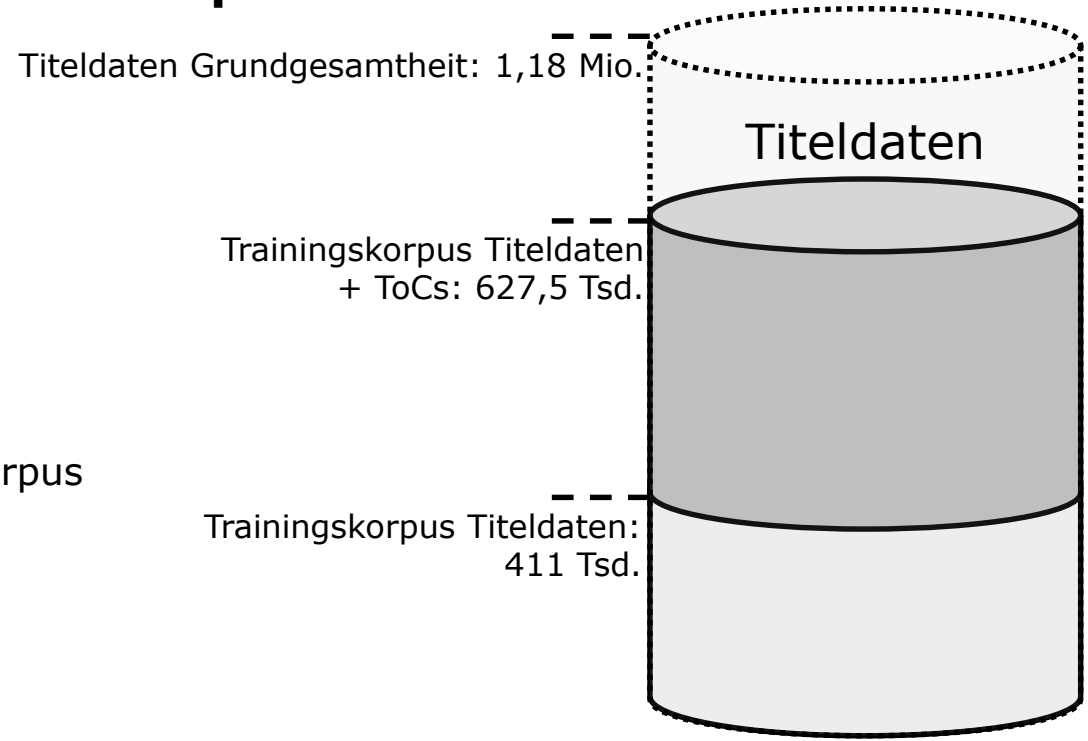


Aufbau der Trainings- und Testkorpora

1,18 Mio. Titeldatensätze
als Grundgesamtheit

411 Tsd. Titeldatensätze
als Trainingskorpus

627,5 Tsd. Titeldatensätze
plus digitalisierte ToCs als Trainingskorpus



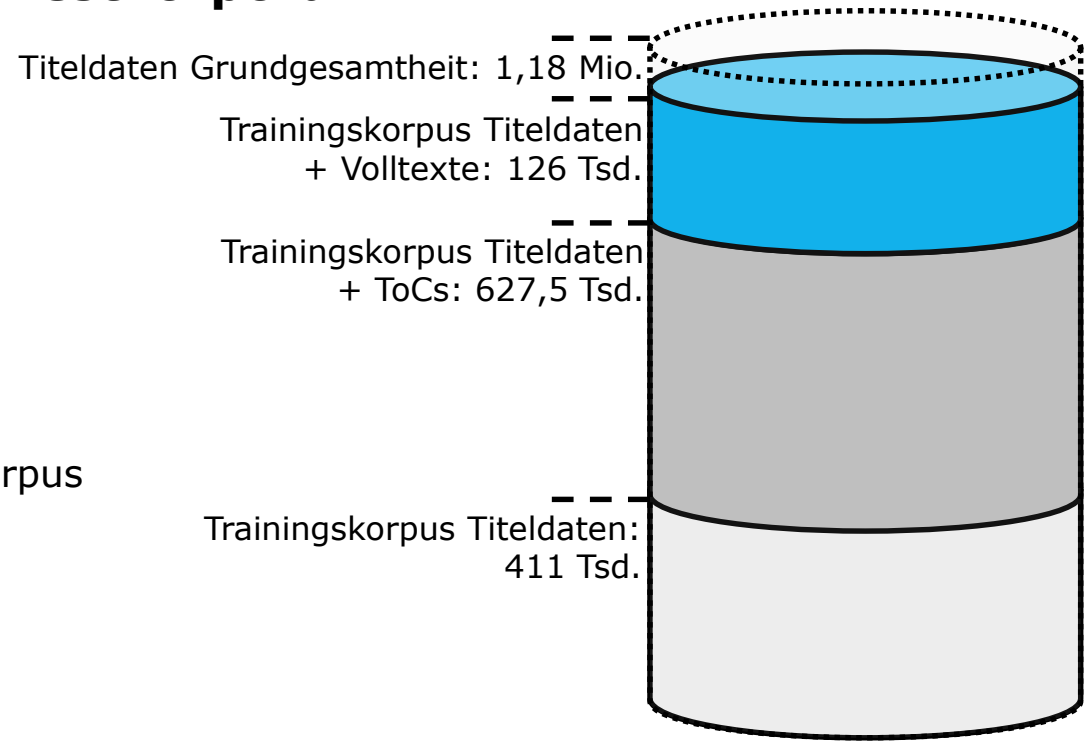
Aufbau der Trainings- und Testkorpora

1,18 Mio. Titeldatensätze
als Grundgesamtheit

411 Tsd. Titeldatensätze
als Trainingskorpus

627,5 Tsd. Titeldatensätze
plus digitalisierte ToCs als Trainingskorpus

126 Tsd. Titeldatensätze
plus Volltexte als Trainingskorpus



Aufbau der Trainings- und Testkorpora

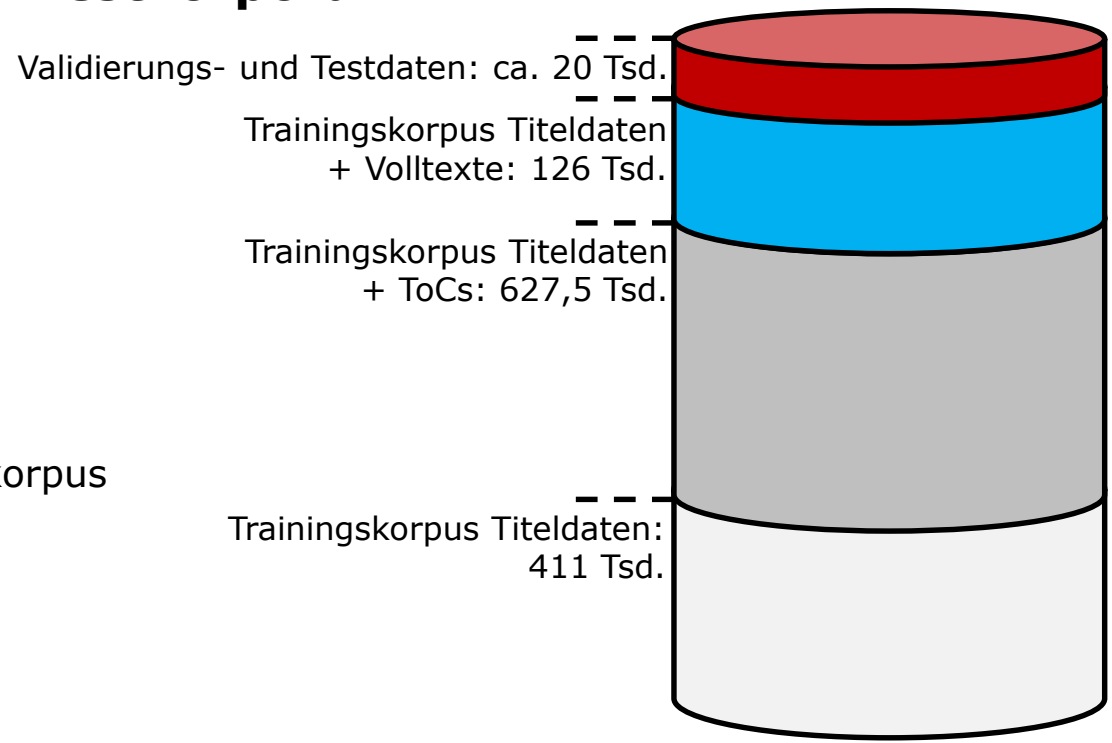
1,18 Mio. Titeldatensätze
als Grundgesamtheit

411 Tsd. Titeldatensätze
als Trainingskorpus

627,5 Tsd. Titeldatensätze
plus digitalisierte ToCs als Trainingskorpus

126 Tsd. Titeldatensätze
plus Volltexte als Trainingskorpus

ca. 20 Tsd. Titeldatensätze
als Validierungs- und Testkorpus



DEUTSCHE NATIONAL BIBLIOTHEK

Indexierungstests: Testgrößen und Testdesign

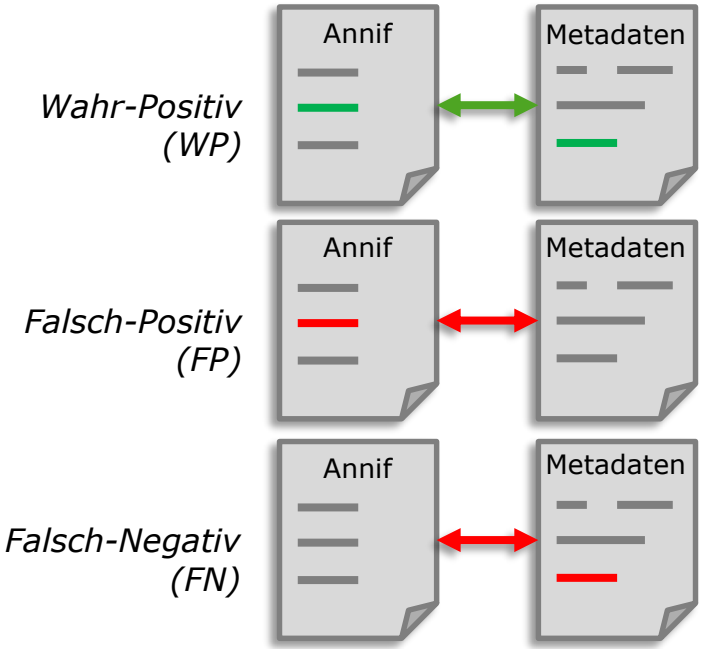
Zentrale Testgröße: F1-Score

Harmonisches Mittel aus Präzisions- (P) und Vollständigkeitsquote (V)

$$P = \frac{WP}{WP + FP}$$

$$V = \frac{WP}{WP + FN}$$

$$F1Score = 2 \times \frac{P \times R}{P + R}$$



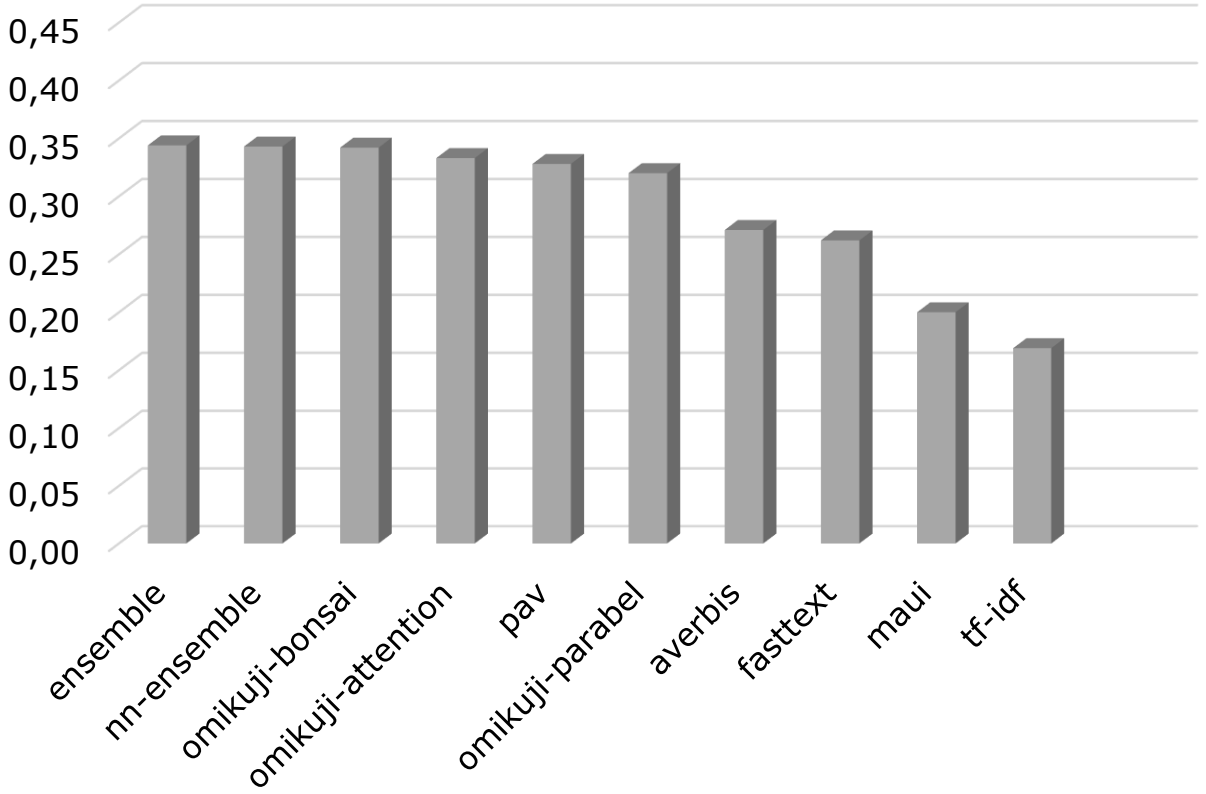
DEUTSCHE NATIONAL BIBLIOTHEK

Indexierungstests: Ergebnisse

Automatische Indexierung anhand verschiedener Verfahrensklassen mit GND-Vokabular.

Testkorpus:
Titeldaten plus Volltexte (n=2.450), intellektuell erschlossen,

F1-Score über die jeweils ersten fünf vergebenen GND-Deskriptoren.



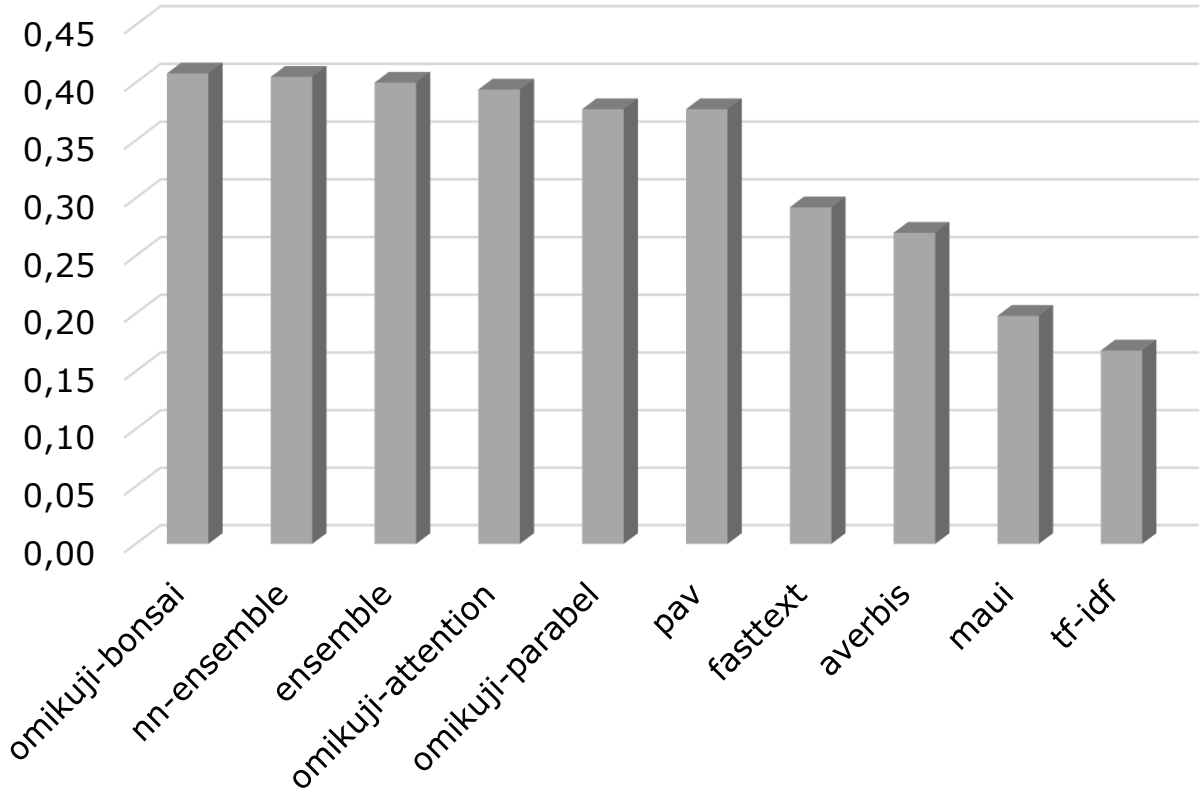
DEUTSCHE NATIONAL BIBLIOTHEK

Indexierungstests: Ergebnisse

Automatische Indexierung anhand verschiedener Verfahrensklassen mit GND-Vokabular.

Testkorpus:
Titeldaten plus ToCs (n=937),
intellektuell erschlossen.

F1-Score über die jeweils ersten fünf vergebenen GND-Deskriptoren.



DEUTSCHE NATIONAL BIBLIOTHEK

Indexierungsevaluationen: Bewertungsschema

Evaluation der Annif-vergebenen GND-Deskriptoren durch die Inhaltserschließung der DNB anhand ordinal skaliertem, vierstufigen Bewertungsschema:

sehr nützlich (Deskriptor beschreibt wichtigen Aspekt ausreichend und trifft absolut zu),

nützlich (Deskriptor beschreibt wichtigen Aspekt aus einer weiteren oder engeren Perspektive und trifft zu),

wenig nützlich (Deskriptor beschreibt einen wichtigen Aspekt nicht ausreichend, ist aber auch nicht völlig unzutreffend),

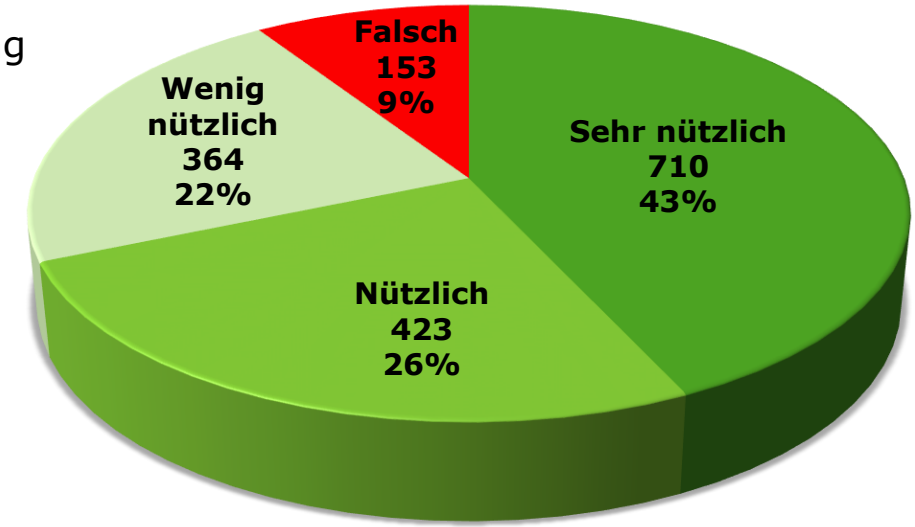
falsch (Deskriptor keinen wichtigen Aspekt oder ist falsch).

Indexierungsevaluationen: Ergebnisse

Evaluation der Annif-vergebenen GND-Deskriptoren durch die Inhaltserschließung der DNB anhand ordinalskalierter, vierstufiger Bewertungsskala.

Testkorpus:
Titeldaten plus Volltext (n=434),
1.650 GND-Deskriptoren mit Annif,
davon maximal fünf
pro Dokument.

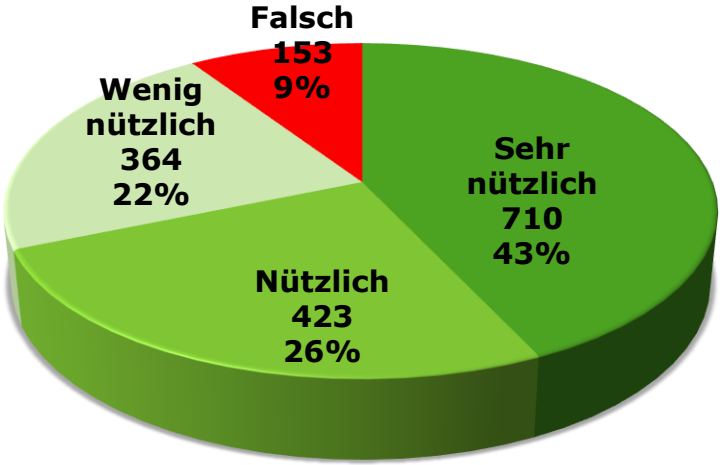
735 fehlende Aspekte
(1,7 pro Dokument)



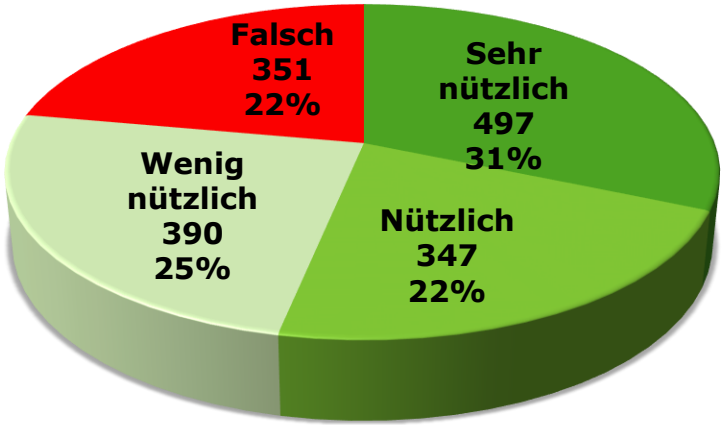
DEUTSCHE NATIONAL BIBLIOTHEK

Indexierungsevaluationen: Ergebnisse

Testkorpus Titeldaten plus Volltext (n=434)
1.650 GND-Deskriptoren mit Annif
735 fehlende Aspekte (1,7 pro Dokument)



Testkorpus Titeldaten plus Volltext (n=434)
1.585 GND-Deskriptoren mit Averbis
859 fehlende Aspekte (2 pro Dokument)



Vielen Dank für die Aufmerksamkeit

Matthias Nagelschmidt
Deutsche Nationalbibliothek
Automatische Erschließungsverfahren, Netzpublikationen
Deutscher Platz 1
04103 Leipzig
m.nagelschmidt@dnb.de