Maschinelle Indexierung der Deutschen Nationalbibliothek

Elisabeth Mödden | Helga Karg

Maschinelle Beschlagwortung

Start: 2014

- Deutschsprachige Online-Publikationen der Reihe 0
  - ohne Zeitschriftentitel
  - ohne Belletristik
- Deutschsprachige Online-Zeitschriftenartikel
- Deutschsprachige Print-Publikationen der Reihen B und H über gescannte Inhaltsverzeichnisse (TOC)

Seit Juni 2018

- Englischsprachige Online-Dissertationen der Reihe O

Umfang: 225.000 Publikationen (Stand Juni 2018)

Maschinelle Beschlagwortung

Methode: computerlinguistisches Verfahren auf der Basis eines Wörterbuches

Basis: bibliografische Titeldaten und elektronische Texte

Terminologie:

GND (nur Datensätze

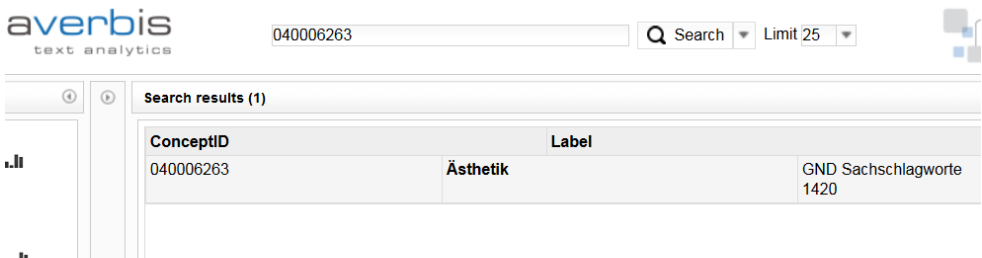
mit Qualitätslevel 1 und Teilbestand s)

NEU:

LCSH (Library of Congress Subject Headings)

in Kombination mit Nicht-Sachbegriffen der GND (Personen, Geografika, Körperschaften, Kongresse und Werke)

Tp – Person (individualisiert) 369.015 Datensätze
Ts - Sachbegriff 184.165 Datensätze
Ts1e - Hinweissatz 4715 Datensätze
Tg - Geografikum 205.244 Datensätze
Tb1 – Körperschaften 142.593 Datensätze
Tf1 - Kongresse 11.831 Datensätze
Tu1 – Werke 87.831 Datensätze



Maschinelle Beschlagwortung Beispiel Englisch

4000 Reconstructing European Copyright Law for the Digital Single Market : Between Old Paradigms and Digital Challenges / Bernd Justin Jütte

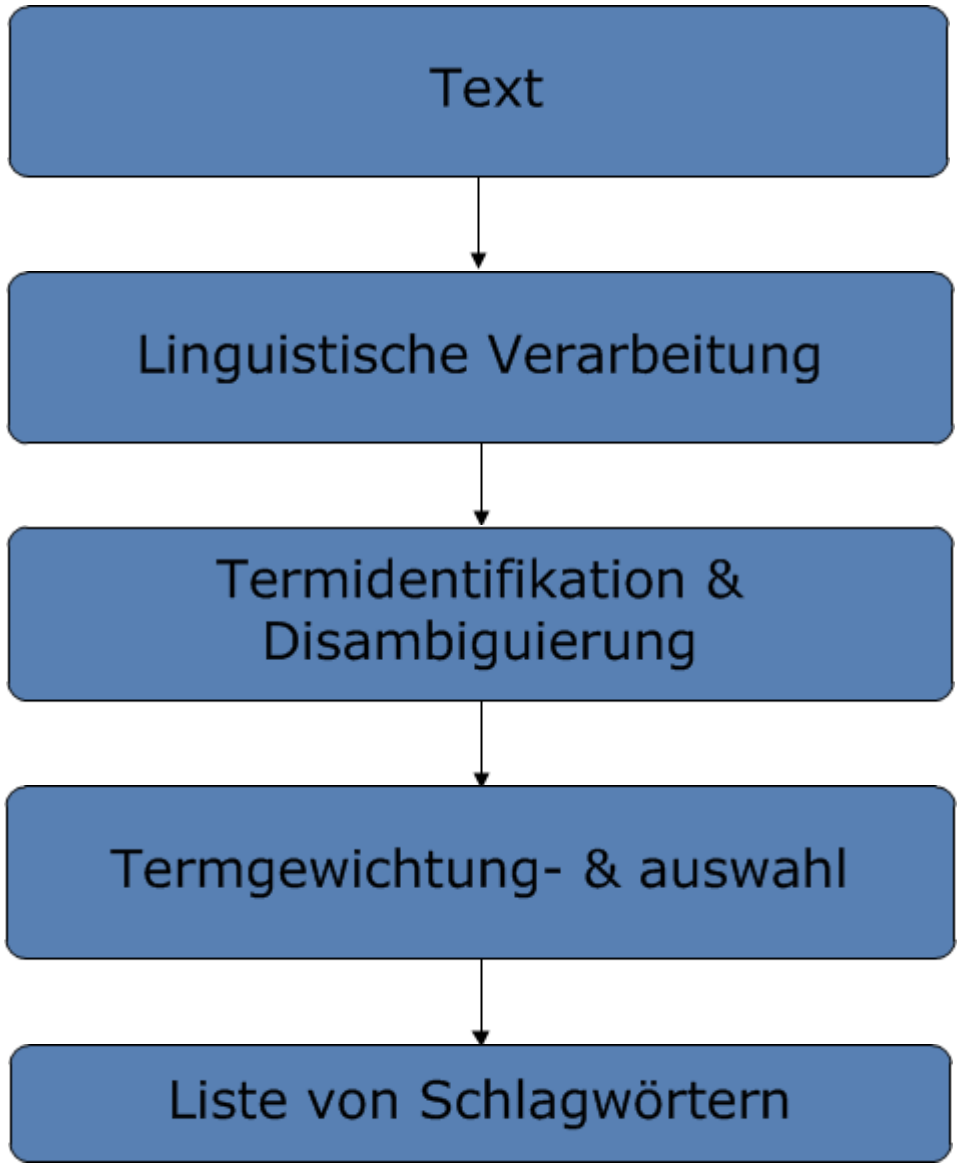
4204 $dDissertation$eUni. Luxembourg$f2016

5050 340EmHaep-sg$K0,99414$D2018-06-05

5540[**GND**]!940431181!**Europäische Union** [Tb1]$K0,00119$D2018-06-05

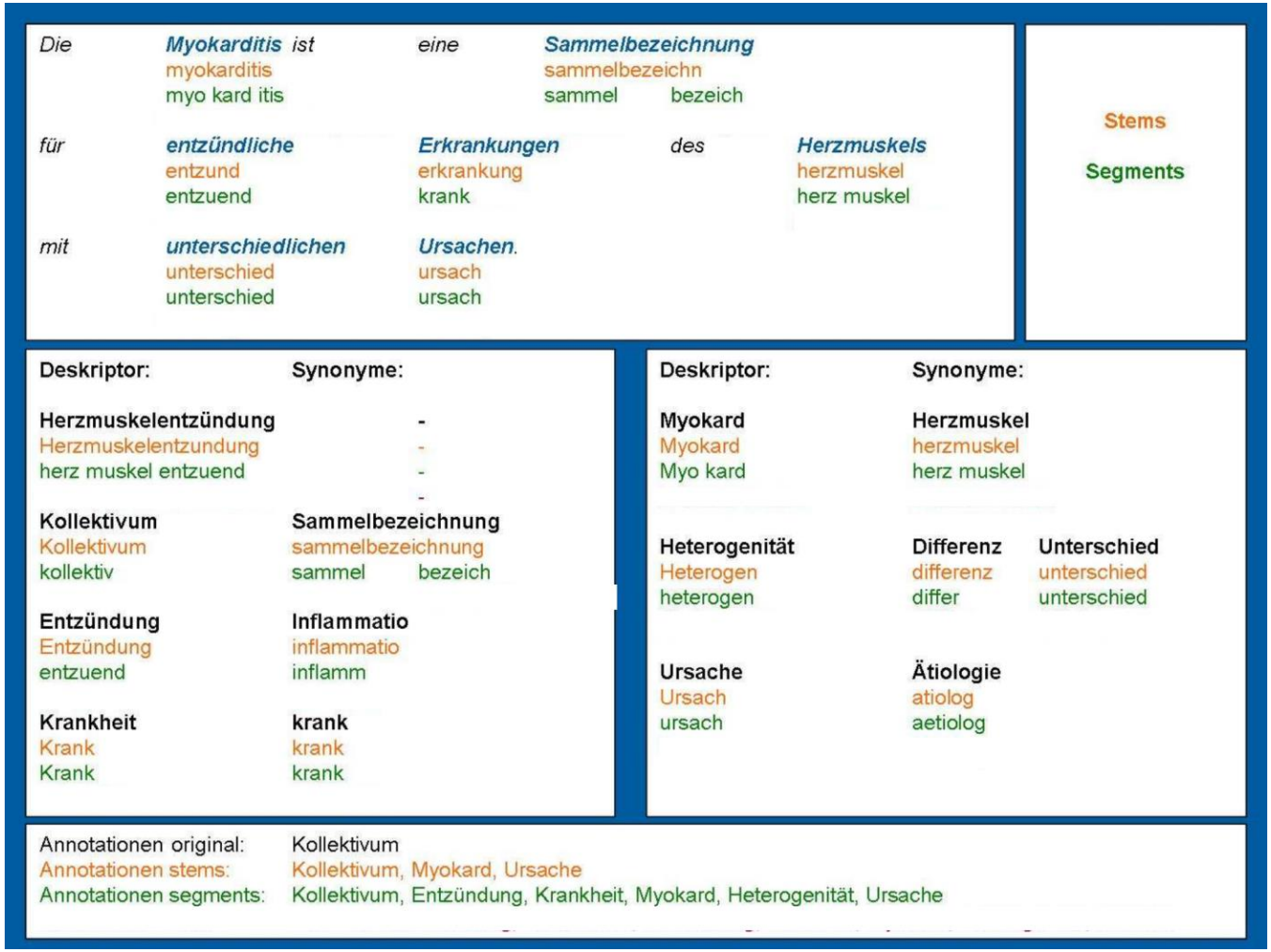
5540[**LCSH**]**Copyright**$Lsh85032446$uhttp://id.loc.gov/authorities/subjects/sh85032446$K0,20713$D2018-06-05

Auslieferung der LCSH ab Januar 2019 geplant



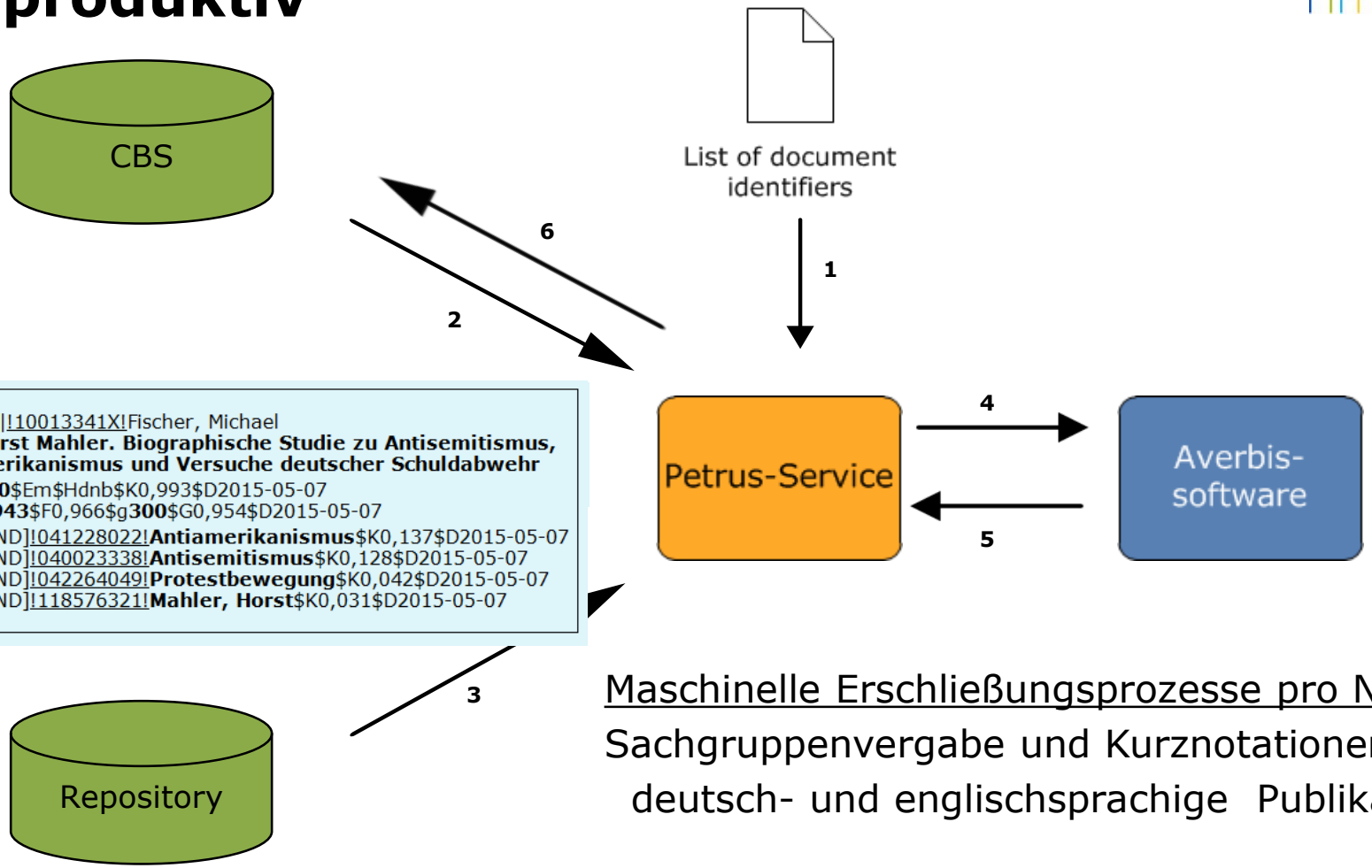
DEUTSCHE NATIONAL BIBLIOTHEK

Abgleich GND-Terminologie



- Original
- Stem
- Segment

Maschinelle Erschließung produktiv

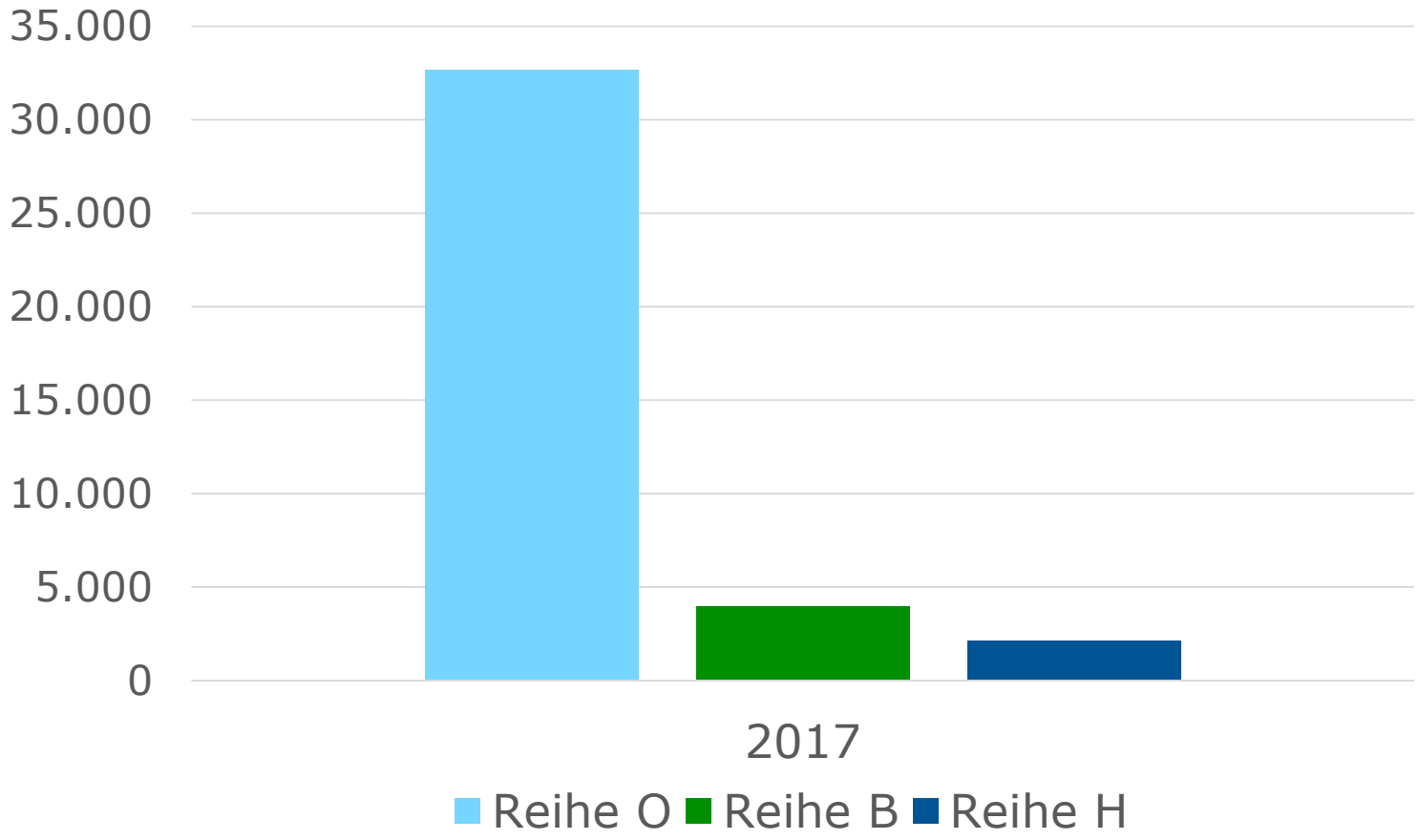


Maschinelle Erschließungsprozesse pro Nacht

Sachgruppenvergabe und Kurznotationen:
- deutsch- und englischsprachige Publikationen

Beschlagwortung:
- deutschsprachige Publikationen
- englischsprachige Dissertationen

Bibliografiejahrgang 2017 maschinell beschlagwortet



Maschinelle Beschlagwortung produktiv

Konfiguration *Wissenschaftliche Arbeiten:*

Reihe O: Hochschulschriften, Monographien von Universitätsverlagen, Wissenschaftliche Literatur verschiedener Verlage

Maschinelle Beschlagwortung produktiv

Konfiguration *BookonDemand* :

Publikationen von BOD-Verlagen

Maschinelle Beschlagwortung produktiv

Konfiguration *TableOfContents* :

Verarbeitung von gedruckten Publikationen mittels TOCs: Reihe B und Reihe H (Monografien)

produktiv seit 20. April 2017

Maschinelle Beschlagwortung produktiv

Konfiguration *Artikel:*

Online-Artikel von Universitätsverlagen,
Wissenschaftliche Papers verschiedener Verlage inkl.
Springer de

Maschinelle Beschlagwortung produktiv

Konfiguration *Englisc*h:

englischsprachige Dissertationen der Reihe O

Fünf Konfigurationen – Fünf Ergebnismengen

- Die eingesetzten Konfigurationen führen täglich zu fünf Ergebnismengen der maschinellen Beschlagwortung.
- Der Anzahl an Publikationen pro Ergebnismenge kann täglich stark schwanken: Zwischen 1 und ca. 1.000 Publikationen können innerhalb eines Tages beschlagwortet worden sein.
- Analyse der Beschlagwortung von zufällig ausgewählten Publikationen (Stichproben) über alle Ergebnismengen.
- Produktionsanalyse in der Tiefe, statt in der Breite.

Qualitätskontrolle - Vorgehensweise

- Intellektuell
- Stichprobenartig
- Publikationen der Reihe O – auf der Basis von Volltexten
- Reihen H und B – auf der Basis von Inhaltsverzeichnissen
- In möglichst vielen Fachgebieten
- Regel der RSWK zur Schlagwortfolgenbildung bleiben unberücksichtigt
- Bewertung der einzelnen Schlagwörter nach festgelegten Kriterien
- Bewertung des Gesamtindexats

Bewertung der maschinellen Beschlagwortung

DEUTSCHE NATIONAL BIBLIOTHEK

Qualitätsanforderungen:

- **Informationsqualität** der maschinell erstellten Metadaten
- **Rechercheanforderungen** der digitalen Wissens- und Informationsgesellschaft
- Verbesserte Auffindbarkeit orientiert an den **Suchfragen der Nutzer**

Qualitätsanforderungen

Die einzelnen GND-Terme sollen das Thema der Publikation richtig und sinnvoll wiedergeben durch:

- Nützliche, gebräuchliche Schlagwörter
- Angemessene Anzahl der relevanter Schlagwörter
- Aktuelle / Gültige Schlagwörter
- Spezifität / Genauigkeit der Schlagwörter im Hinblick auf das vorhandene Vokabular
- Konsistente Vergabe von Schlagwörtern

Bewertungskriterien - Schlagwort

Das einzelne Schlagwort beschreibt einen wesentlichen Aspekt des Textes ausreichend und trifft absolut zu	Sehr nützlich 3
Das einzelne Schlagwort beschreibt einen wichtigen Aspekt des Textes aus einer etwas weiteren/engeren Perspektive zutreffend	Nützlich 2
Das einzelne Schlagwort beschreibt einen wichtigen Aspekt des Textes nicht ausreichend, ist aber auch nicht völlig unzutreffend oder falsch	Wenig nützlich 1
Das einzelne Schlagwort beschreibt keinen wichtigen Aspekt des Textes und ist falsch oder nicht nützlich	Falsch 0

Stichprobenprüfung Beispiel

Titel: Numerische Analyse des Nachstroms und Propellereffektivität am manövrierenden Schiff

Schlagwörter:

5540 [GND]!…!Schiff**$K**0,476…	wenig nützlich
5540 [GND]!…!Manöver$gSchifffahrt**$K**0,356…	sehr nützlich
5540 [GND]!…!Wirbelschleppe**$K**0,312…	falsch
5540 [GND]!…!Schiffsantrieb**$K**0,041…	nützlich
5540 [GND]!…!Nachstrom**$K**0,003…	falsch

GND: Nachstrom
OB Elektrischer Strom

Ergänzung Fehlender Aspekte Disambiguierung

4000 Numerische Analyse des Nachstroms und Propellereffektivität am manövrierenden Schiff

5540	[GND]…Schiff**$K**0,476…	wenig nützlich
5540	[GND]…Manöver$gSchifffahrt**$K**0,356…	sehr nützlich
5540	[GND]…Wirbelschleppe**$K**0,312…	falsch
5540	[GND]…Schiffsantrieb**$K**0,041…	nützlich
5540	[GND]…Nachstrom**$K**0,003…	falsch

5540 [FA]…Nachstrom$gStrömungsmechanik ← Neuansetzung

5540 [FA]…Schiffspropeller

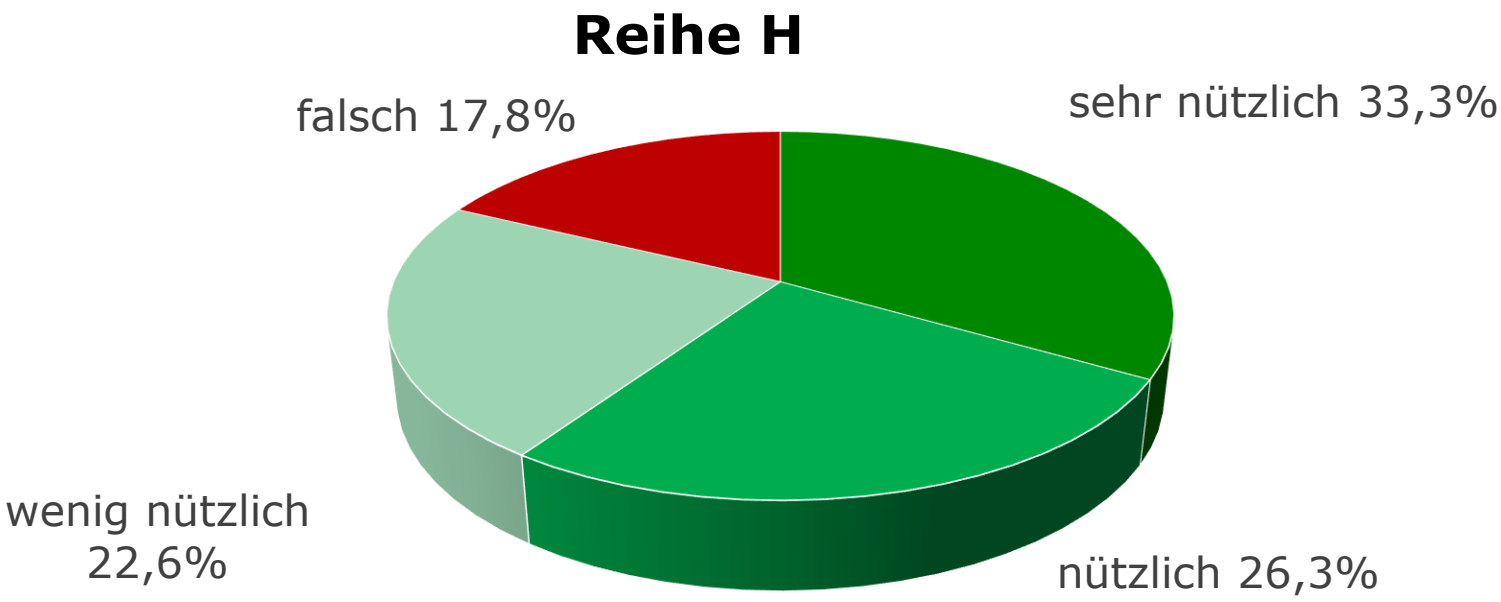
5540 [FA]…Numerische Strömungssimulation

Bewertungskriterien - Gesamtindexat

Das Indexat ist vollständig und bildet den Inhalt / das Thema der Publikation vollständig ab	Sehr gut	3
Das Indexat ist nicht vollständig, enthält aber das / die wesentlichen Schlagwörter	Gut	2
Das Indexat ist unvollständig und die Schlagwörter bilden den Inhalt / das Thema zu weit, zu eng oder nur zu einem Teil ab	Mäßig	1
Das Indexat bildet den Inhalt / das Thema nicht ab oder es wurden zu viele falsche Schlagwörter vergeben	Unbrauchbar	0

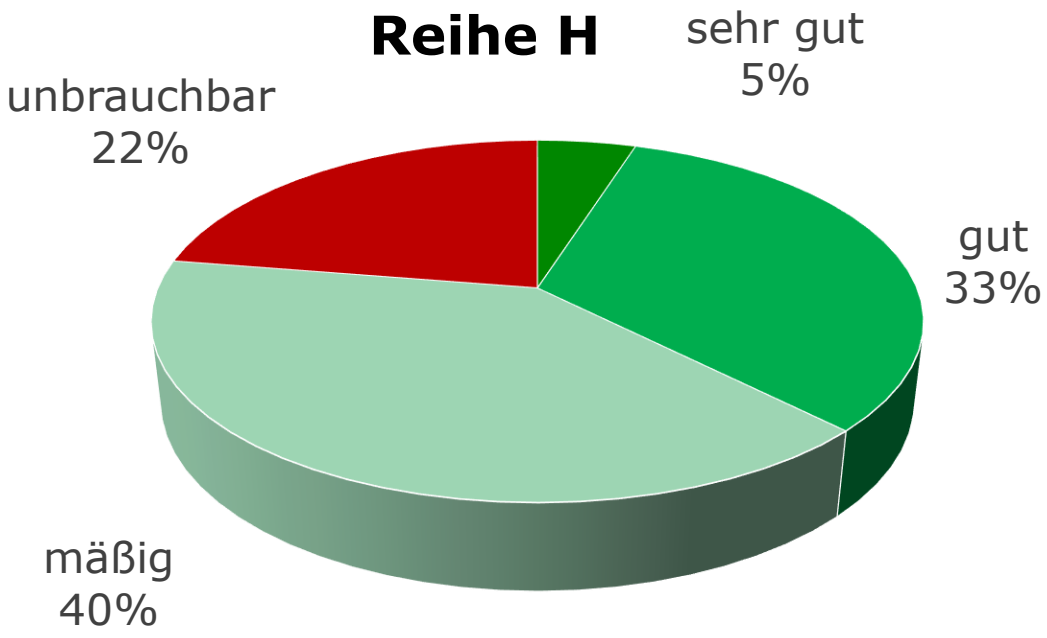
Evaluationsgebnisse der intellektuellen Bewertung maschinell vergebenen Schlagwörter

Betrachtungszeitraum: 20.04-14.11.2017



Evaluationsgebnisse der intellektuellen Bewertung des Gesamtindexats

Betrachtungszeitraum: 20.04-14.11.2017



Beobachtungen der Fachreferenten

- Höhere **Präzision der Titelstichwörter** bezogen auf den Inhalt des Werks führt zu besseren Ergebnissen
- **Werbende Titel, Wortspiele, Metaphern** oder andere rhetorische Figuren (oder auch oftmals auch Zitate aus Primärquellen) führen zu falschen Ergebnissen
- Besonders **abstrakte Fachsprache** bzw. ein metaphorischer Sprachgebrauch in bestimmten Wissenschaftsfächern ist problematisch, da sich ihre Sprache von der Alltagssprache stark unterscheidet.
- **Viele verschiedene Themen**, die sich im Inhaltsverzeichnis angemessen widerspiegeln sind oft besser erschlossen als stringent auf nur ein bestimmtes inhaltliches Merkmal fokussierte Dokumente

Beobachtungen der Fachreferenten

- Erkennen der richtigen **Körperschaft oder Person** ist über alle Fächer hinweg problematisch.
  Bsp.: Akzidenzschriften, Sammelbände, bei denen immer wieder die in Inhaltsverzeichnissen aufgeführten Namen der Verfasser der Einzelbeiträge als Schlagwörter zur Beschreibung des Inhalts verwendet werden.

- Erkenne von **Geographika**

- Es konnten bisher keine filterbaren **Dokumentgruppen** innerhalb der Reihe B erkannt werden, die für bes. gute oder schlechte Qualität steht

- Verbesserung der ToC basierten Erschließung der Reihen B /H durch Erweiterung der **Textbasis** auf Abstracts und Klappentexte

Fehleranalyse - Hauptursachen

Falsche Beschlagwortung durch:

- GND-Vokabular:
  - Falsche Auflösung von Mehrdeutigkeiten
    (Bank -> Sitzgelegenheit oder Bank -> Kreditinstitut)
  - Fehlende Schlagwörter, Synonyme
- Kein passendes Schlagwort in der GND -> falsches Schlagwort wird vergeben
- Linguistische Analyse der Software führt zu falscher Beschlagwortung
- Schlagwort wird korrekt vergeben, ist für die Abbildung des Publikationsinhaltes aber irrelevant

Optimierungsstrategien

DEUTSCHE NATIONAL BIBLIOTHEK

- Pflege des Vokabulars, der GND mit Hilfe von Tools zur Visualisierung, Vorschlagssystem für neue Deskriptoren etc.
- Wörterbuchpflege
- Suchmaschinenoptimierung
- Verbesserung der maschinellen Verfahren durch Qualitätsmanagement
- Entwicklung neuer Komponenten
- Verbesserung der Scans sowie der Qualität der OCR

Vielen Dank für Ihre Aufmerksamkeit