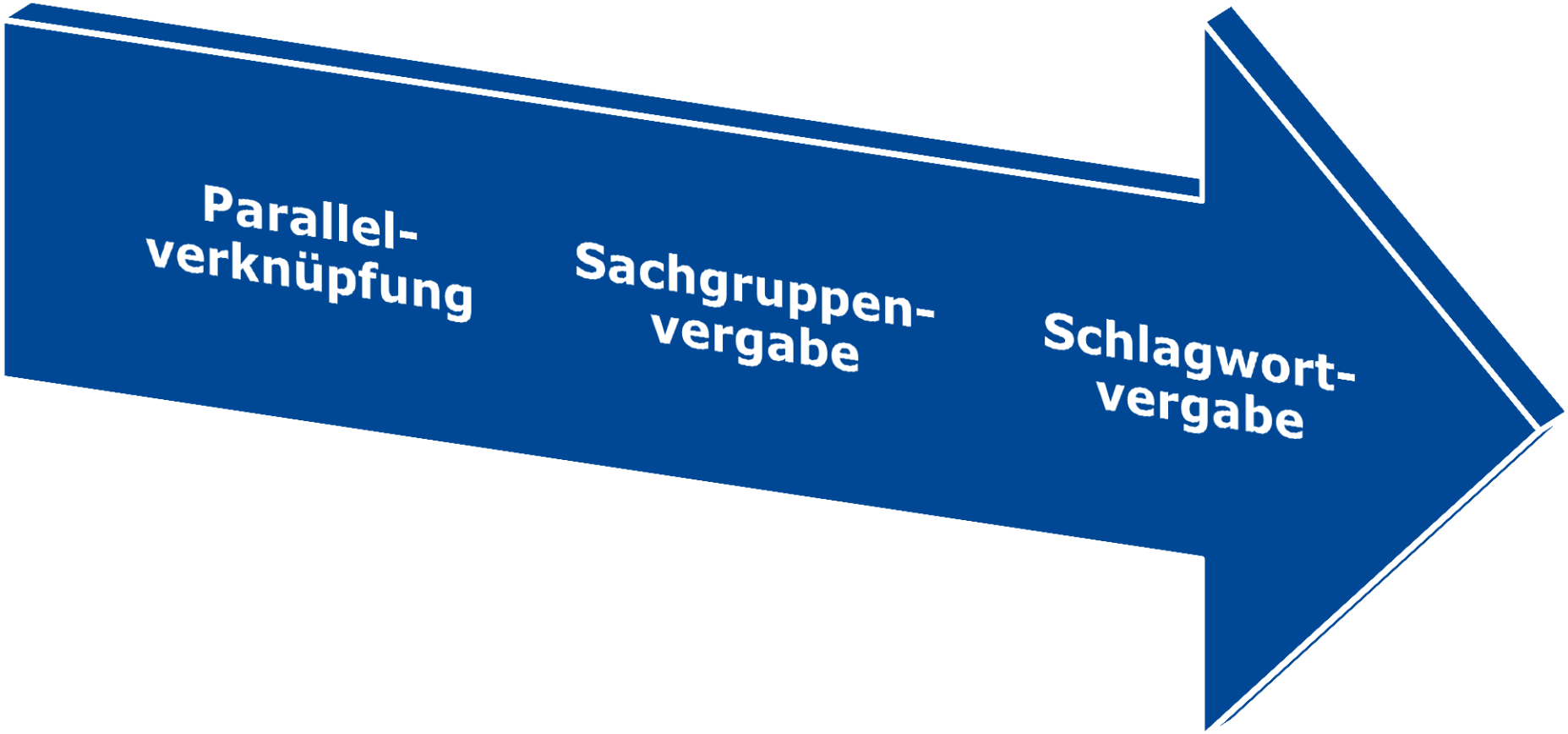
DEUTSCHE NATIONAL BIBLIOTHEK

Elisabeth Mödden

Maschinelle Klassifikation und Beschlagwortung der Reihe O in der Deutschen Nationalbibliothek Einblicke in die Praxis

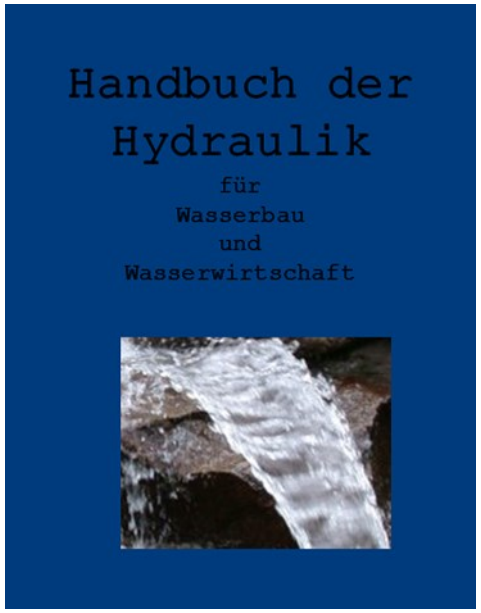
DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließungsverfahren



DEUTSCHE NATIONAL BIBLIOTHEK

Parallelverknüpfung



DEUTSCHE NATIONAL BIBLIOTHEK

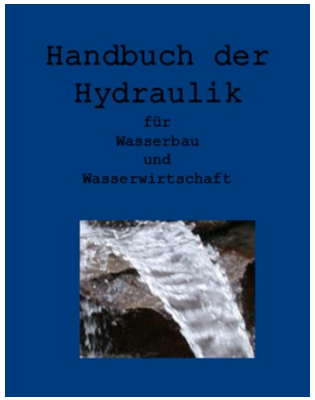
Netzpublikation

3000 Aigner, Detlef

4000 Handbuch der Hydraulik : für Wasserbau und Wasserwirtschaft / Detlef Aigner

4030 Berlin : Beuth Verlag GmbH
4060 Online-Ressource

5050 620EaHonx$D2015-07-28



DEUTSCHE NATIONAL BIBLIOTHEK

Printpublikation

3000 !1033613576!Aigner, Detlef

4000 Handbuch der Hydraulik für Wasserbau und Wasserwirtschaft / Detlef Aigner

4030 Berlin ; Wien ; Zürich : Beuth

5050 624;620
5100 !040263126!Hydromechanik
5101 !040647005!Wasserbau
5400 [DDC22ger]627
5401 627

DEUTSCHE NATIONAL BIBLIOTHEK

Netzpublikation

3000 **!1033613576!** Aigner, Detlef

4000 Handbuch der Hydraulik : für Wasserbau und Wasserwirtschaft / Detlef Aigner

4030 Berlin : Beuth Verlag GmbH
4060 Online-Ressource

4243 Druckausg.!1009964852!--Aa--Aigner, Detlef: Handbuch der Hydraulik für Wasserbau und Wasserwirtschaft
4700 |PE|*Parallelverknüpfung wurde automatisch erstellt

5050 624;620**EpD**2015-07-29
5050 620**EaH**onx**$D**2015-07-28
5100 !040263126!Hydromechanik
5101 !040647005!Wasserbau
5400 [DDC22ger]627
5401 627

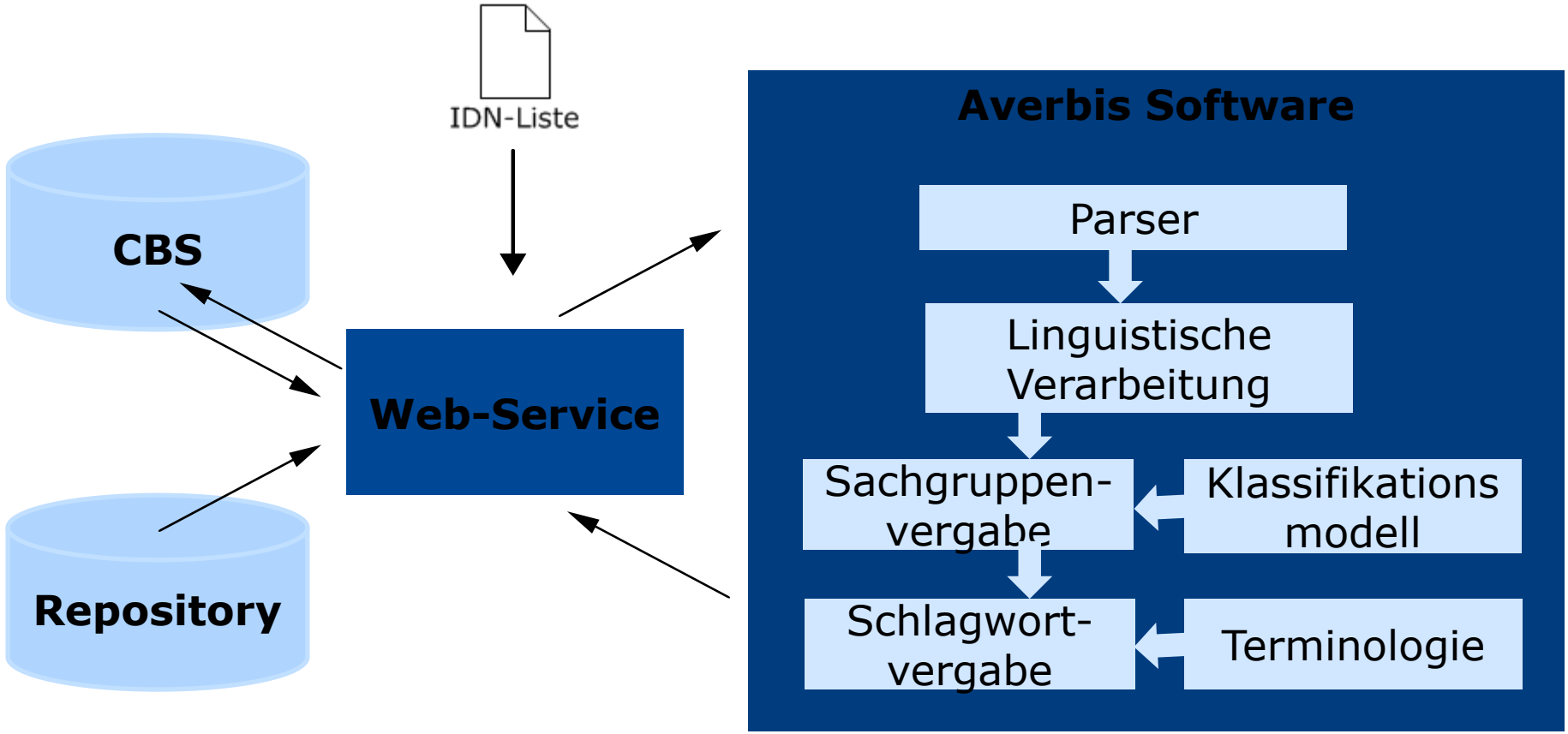
DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Sachgruppen- und Schlagwortvergabe

- Technologie: Averbis Extraction Platform der Firma Averbis GmbH Freiburg i.Br.
- Schlagwortvergabe durch Extraktion gezielter Informationen aus Texten und Abgleich mit der GND-Terminologie
- DDC-Sachgruppenvergabe durch Klassifikation von Texten anhand maschineller Lernverfahren

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Sachgruppen- und Schlagwortvergabe



DEUTSCHE NATIONAL BIBLIOTHEK

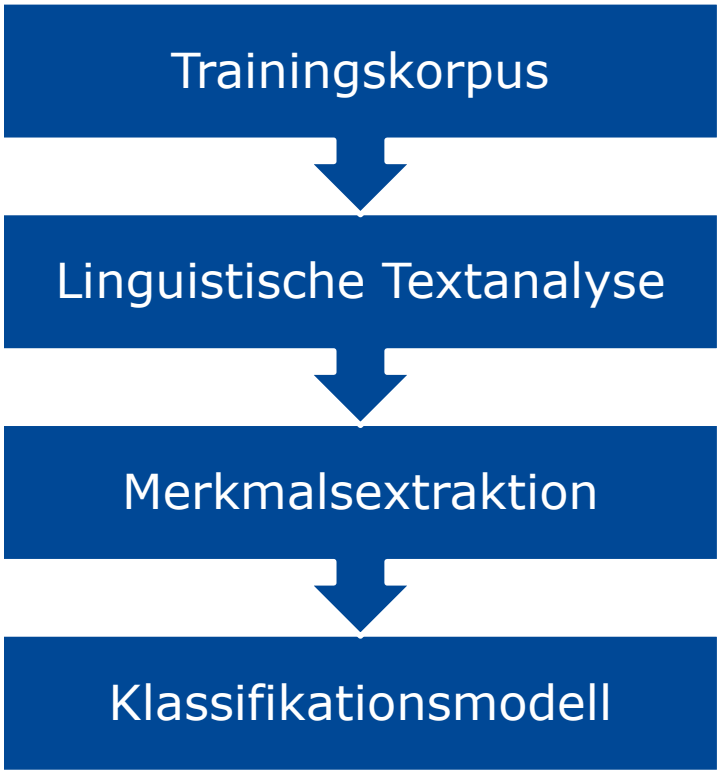
Maschinelle Sachgruppenvergabe

- Der Klassifikator ordnet Netzpublikationen in die Kategorien der DDC-Sachgruppen mit ca. 100 Klassen ein.
- Training von Klassifikationsmodellen durch Einsatz statistischer Lernverfahren z. B. Support Vector Machine (SVM)..
- Das System lernt anhand von Trainingsbeispielen und leitet Gesetzmäßigkeiten aus diesen Trainingsbeispielen ab.
- Trainingsbeispiele sind Netzpublikationen und gescannte Inhaltsverzeichnisse mit Sachgruppen, die durch die Inhaltserschließung vergeben wurden.

Produktive Anwendung für deutsch- und englischsprachige Netzpublikationen seit 2012

DEUTSCHE NATIONAL BIBLIOTHEK

Training



Trainingsdaten mit Sachgruppe

Netzpublikationen:

- ca. 111.000 in deutscher Sprache
- ca. 57.000 in englischer Sprache

Gescannte Inhaltsverzeichnisse:

- ca. 300.000 in deutscher Sprache
- ca. 48.000 in englischer Sprache

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelles Lernen

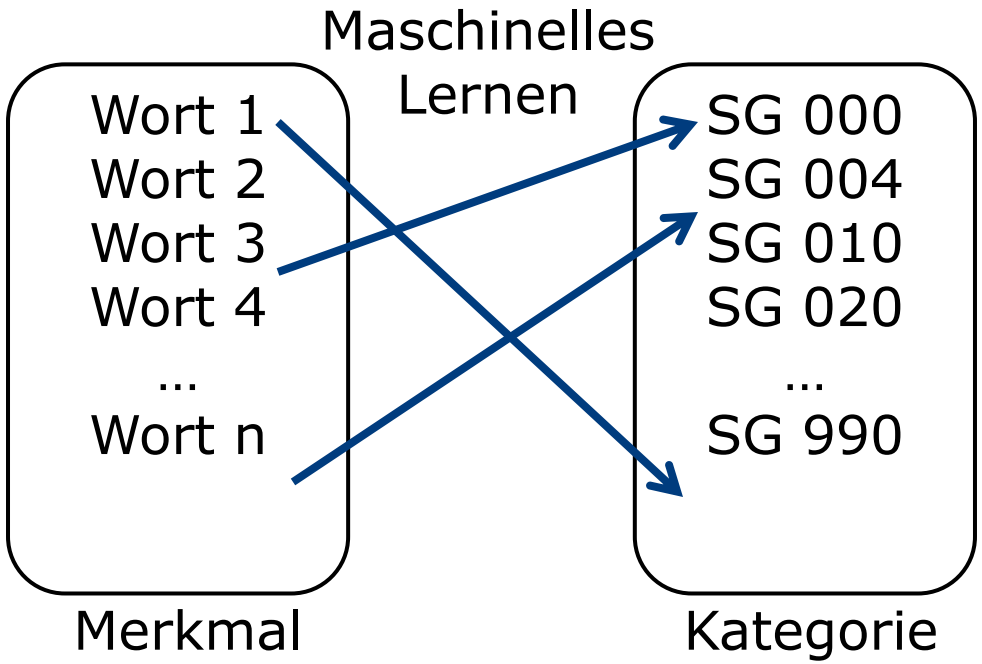
Merkmalsgewinnung

↓

Merkmalsreduktion

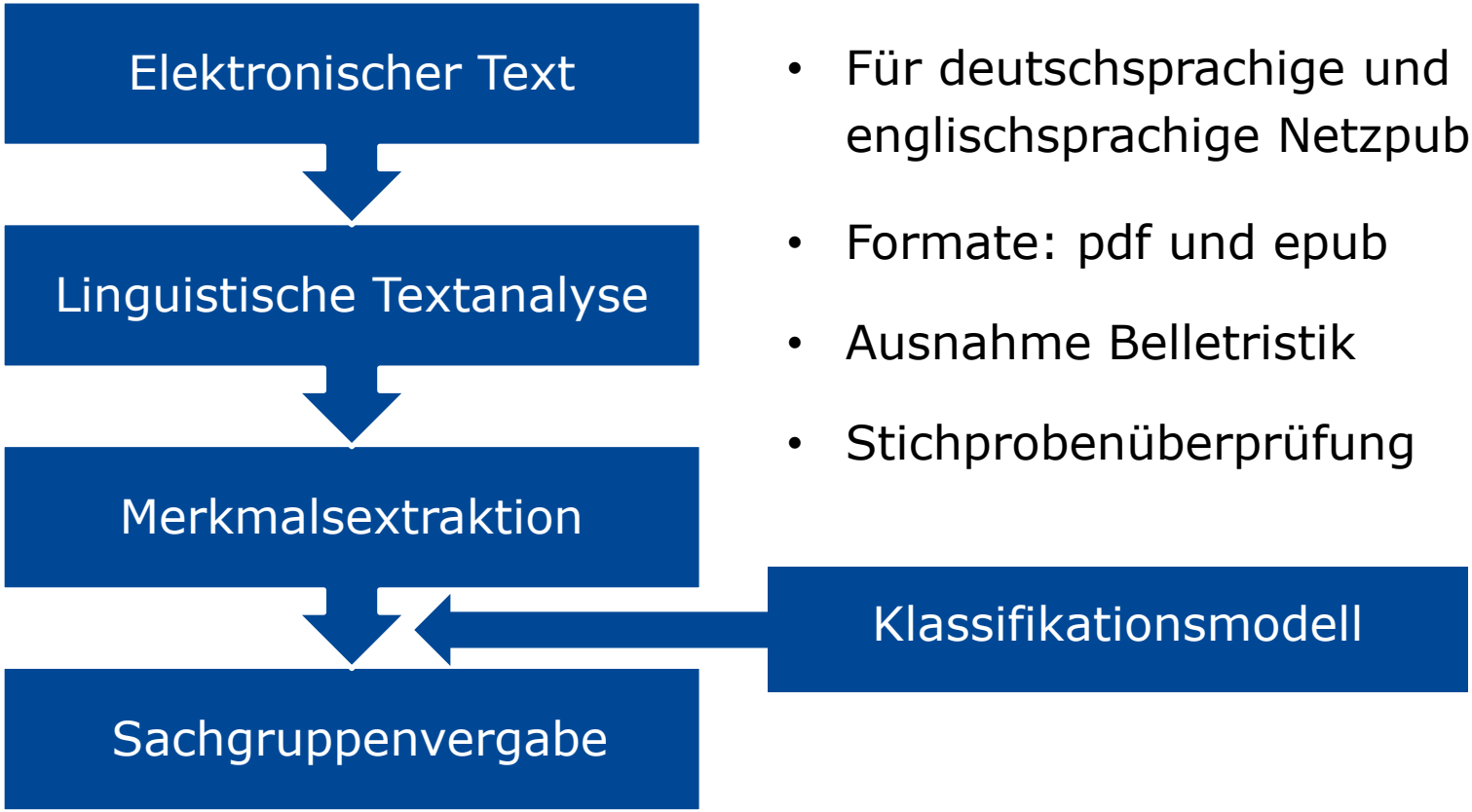
↓

Merkmalsgewichtung



DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Sachgruppenvergabe



- Für deutschsprachige und englischsprachige Netzpublikationen
- Formate: pdf und epub
- Ausnahme Belletristik
- Stichprobenüberprüfung

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Sachgruppenvergabe

0500 Oaf

0600 ro;ra;pb

3000 |m|!1031805168!Colverson, Michael

4000 Bist Du schon wach oder schläfst Du noch? : In Geiselhaft der Großbanken und Großkonzerne werden wir entweder geschoren oder geschlachtet! / Michael Colverson

5050 330**EiD**2013-03-05

5050 330**EpD**2013-03-08

5050 330**EmH**dnb**$K**0,8**$D**2013-03-0

5050 000**EaH**onx**$D**2013-03-01

Die dargestellte Reihenfolge entspricht der Rangfolge der Sachgruppen:

$Ei – intellektuell erstellt
$Ep – aus paralleler Ausgabe
$Em – maschinell gewonnen
$Ea – abgeliefert (Fremddaten)

DEUTSCHE NATIONAL BIBLIOTHEK

Herausforderungen

- Ungleichmäßige Verteilung der Netzpublikationen auf die Sachgruppen!
- Kein Trainingsmaterial für alle Sachgruppen!
- 2004 bis 2006 Vergabe der Sachgruppen, ohne Kenntnis der DDC. Fehler im Trainingsmaterial. (16.000 Netzpublikationen müssten überprüft werden)!
- Neue Sachgruppe seit 2010:
  - 333.7 Natürliche Ressourcen, Energie und Umwelt
  - 620 Ingenieurwissenschaften und Maschinenbau
  - 621.3 Elektrotechnik, Elektronik
  - 624 Ingenieurbau und Umwelttechnik
  - 491.8 Slawische Sprachen
  - 891.8 Slawische Literatur

DEUTSCHE NATIONAL BIBLIOTHEK

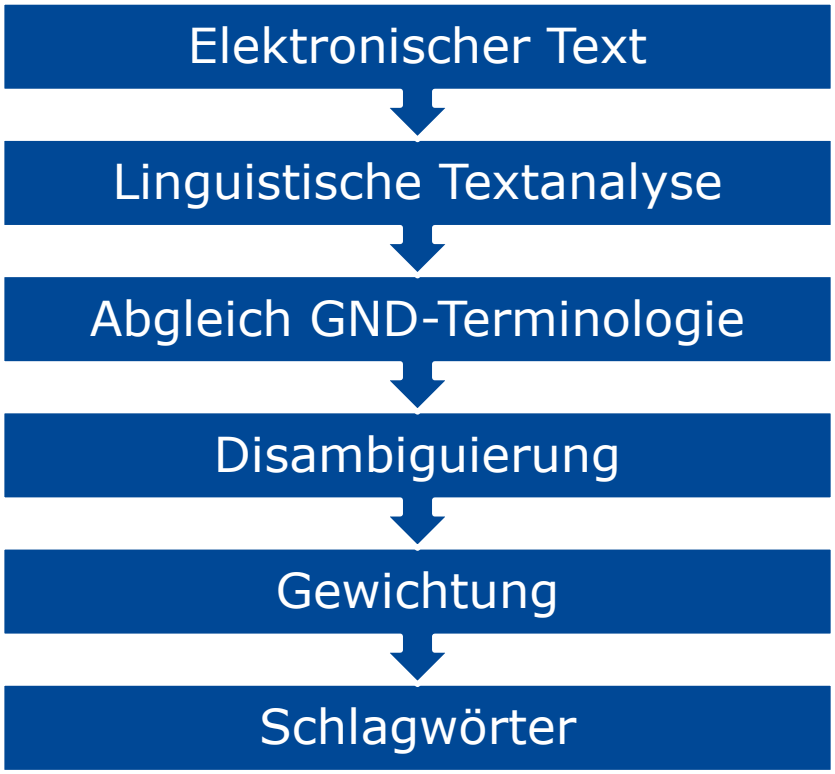
Maschinelle Schlagwortvergabe

- maschinelle Beschlagwortung von deutschsprachigen Online-Hochschulschriften mit dem Vokabular der Gemeinsamen Normdatei (GND)
- GND-Vokabular
  - Qualitätslevel 1
  - Teilbestand s

Tp - Person (individualisiert) / 369.015 Datensätze
Ts - Sachbegriff / 184.165 Datensätze
Ts1e – Hinweissatz / 4.715 Datensätze
Tg – Geografikum / 205.244 Datensätze
Tb1 – Körperschaft / 142.593 Datensätze
Tf1 – Kongress/ 11.831 Datensätze
Tu1 – Werk / 87.831 Datensätze

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Schlagwortvergabe



- pro Online-Hochschulschriften bis zu 10 GND-Schlagwörter, gesteuert über einen Konfidenzwert
- Workflow für Stichprobenprüfung
- Wörterbuchpflege

Produktive Anwendung für deutschsprachige Online-Hochschulschriften seit April 2014

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistische Analyse

Original Text

…zur Bewertung der Tragfähigkeit von Asphaltstraßen-befestigungen aufgrund von FWD-Messungen…

Satzerkennung

Worterkennung

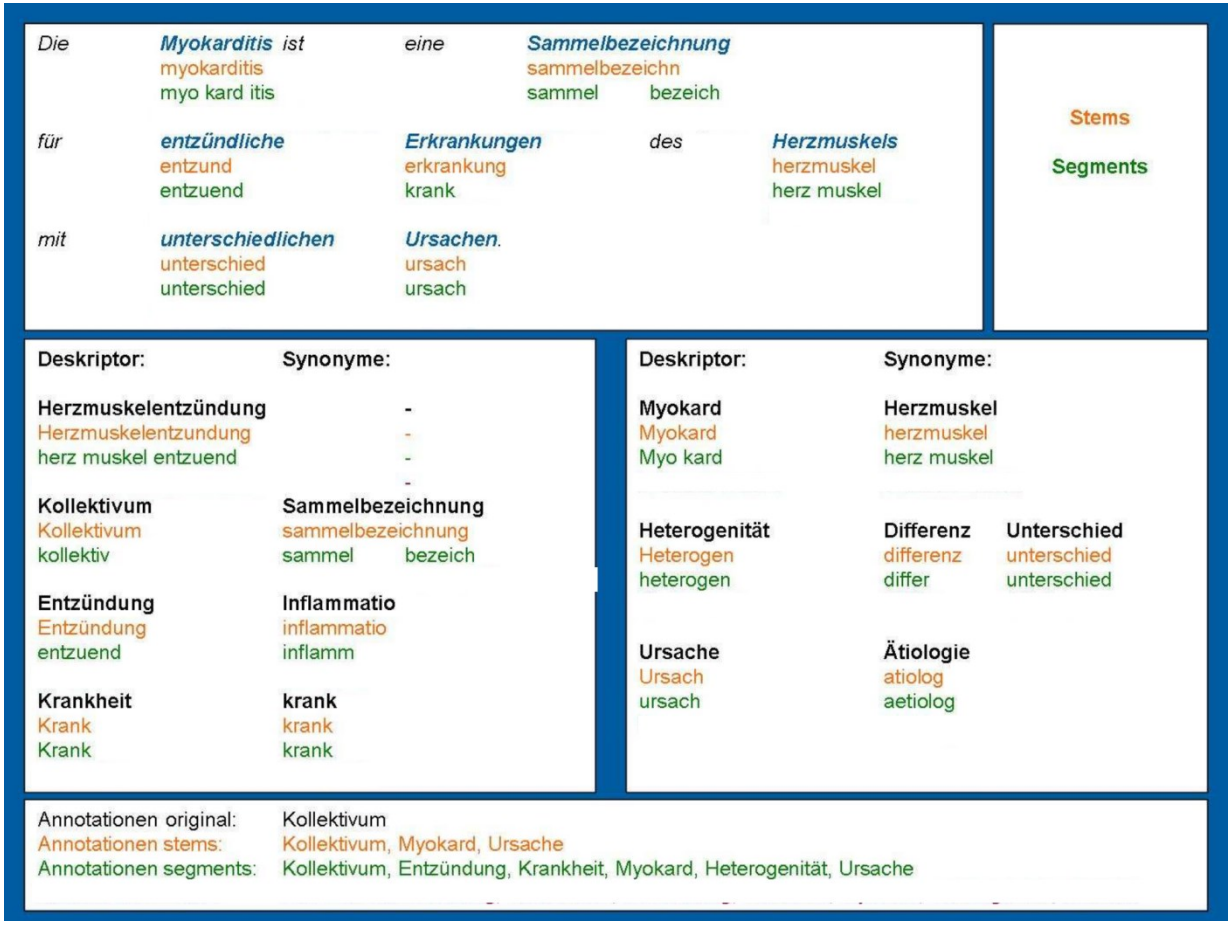
Wortartenerkennung

Phrasenerkennung (Nominalphrasen)

Stammformbildung

Morpho-Semantische Analyse

Abgleich GND-Terminologie



auf unterschiedlichen Ebenen:

- Original
- Stem
- Segment

Bildquelle: Averbis GmbH

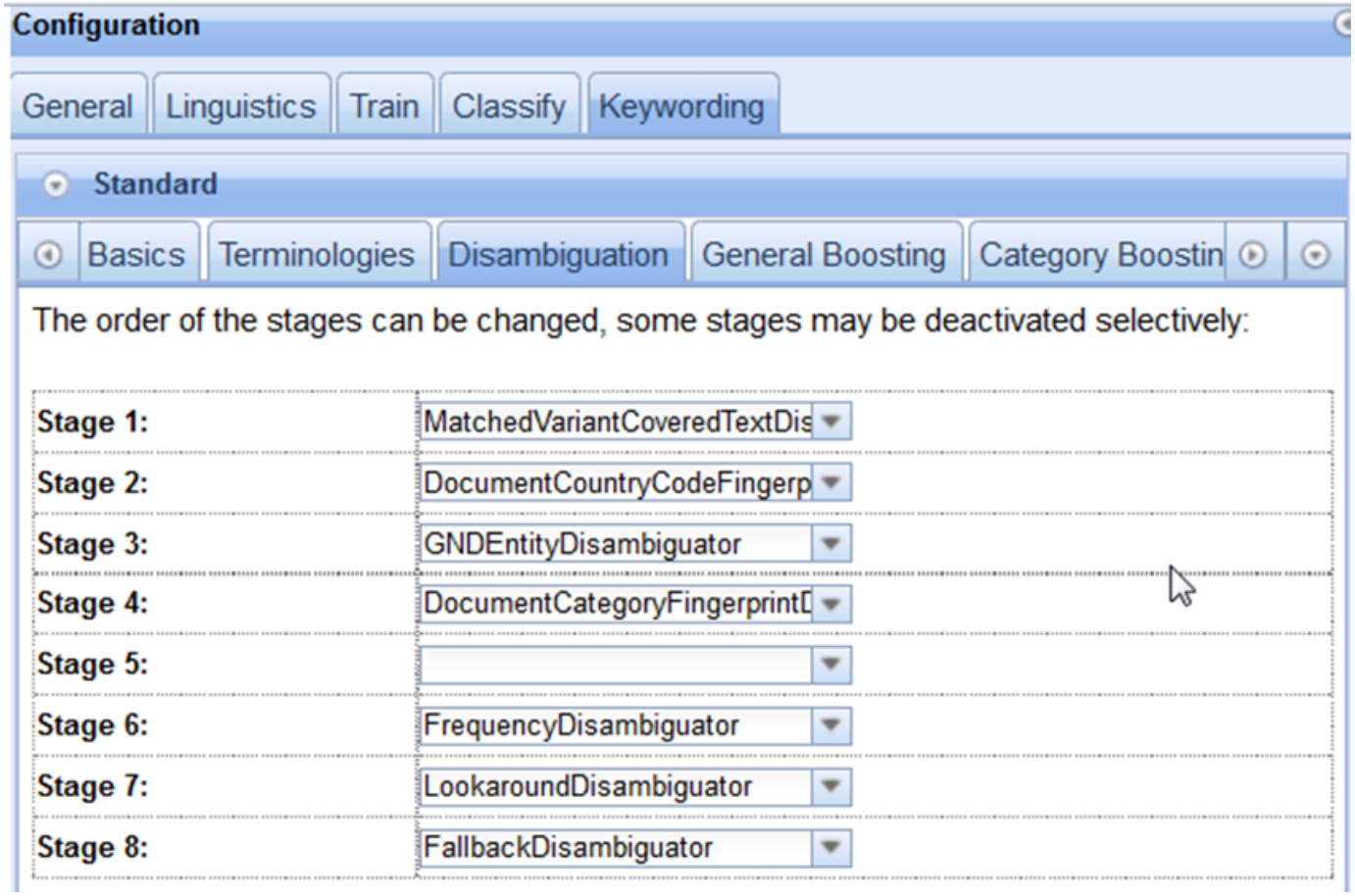
DEUTSCHE NATIONAL BIBLIOTHEK

Ambiguität

005 Ts1 011 s 065 10.9a;10.2ea 083 332.456 150 Wechselkurs 450 Devisenkurs 450 **Kurs**$gDevisen	005 Ts1 011 s 065 10.9c 083 332.63222 150 Aktienkurs 450 **Kurs**$gAktie 450 Aktienpreis	005 Ts1 011 s 065 10.9c 083 332.63222 150 Wertpapierkurs 450 Effektenkurs 450 **Kurs**$gWertpapier
005 Ts1 011 s 065 10.9c 083 332.63222 150 Börsenkurs 450 **Kurs**$gBörse 450 Börsenpreis	005 Ts1 011 s 065 6.4 083 T1--071 150 **Kurs** 450 Lehrgang 450 Seminar$gKurs 450 Seminar$gLehrgang 450 Workshop 450 Kurse	005 Ts1 011 s 065 10.6a 083 387.52 150 **Kurs**$gNavigation 550 Navigation$4obal

DEUTSCHE NATIONAL BIBLIOTHEK

Disambiguierung



Das Verfahren durchläuft mehrere Stufen.

DEUTSCHE NATIONAL BIBLIOTHEK

Document-Fingerprint basierend auf der GND-Systematik

Link zu diesem Datensatz: http://d-nb.info/968571956
Titel: **Wissen im Fluß** [Elektronische Ressource] : **Prozeßorientierung im Wissensmanagement unter Verwendung grafischer Modelle** / Katja Franziska Pook
Schlagwörter: Wissensmanagement ; Kognitive Psychologie ; Online-Publikation; Mitarbeiter ; Einarbeitung ; Wissensvermittlung ; Informationssystem ; Prozesskette ; Graphische Darstellung
Sachgruppe(n): 150 Psychologie ; 650 Management

Top 5 GND-Systematik Topics gemäß Document Fingerprint

(GND-Systematik Nummer=Anzahl, GND-Systematik Notationsbenennung)

4.3=106, Erkenntnistheorie, Logik
1=82, Allgemeines, Interdisziplinäre Allgemeinwörter
30=64, Informatik, Datenverarbeitung
9.3c=61, Gruppe, Organisationssoziologie, Interaktion
10.11a=61, Betriebswirtschaftslehre (Allgemeines), Unternehmen, Management
6.5=45, Wissenschaft

10.9a=1, 10.9c=1, 6.1a=1, 12.1b=1, 25.2a=1, 7.15a=1, 8.4=1, 10.8b=1, 9.3a=1, 19.1b=1,
19.1a=1, 7.5c=1, 10.4=1, 31.14=1, 18=1, 10.7b=1, 28p=1, 21.1=1, 27.4=1, 16.1p=1, 13.1c=1,
12.1p=1, 30m=1, 8.2a=1, 10.6a=1, 13.1cy=1, 10.11g=1, 6.4p=1, 15.4=1, 15.3=1, 15.1=1,
12.4y=1, 10.12a=1, 7.6a=1, 13.5=1, 14.1=1, 27.20=1, 12.4=1, 24.2b=1, 2.3p=1, 13.6p=1,

DEUTSCHE NATIONAL BIBLIOTHEK

Stichprobenprüfung

- Titel-Stichproben

  Vierstufige Bewertungsskala:
  sehr nützlich / nützlich / wenig nützlich / falsch

- Schlagwort-Stichproben

  Überprüfung der Erschließungskonsistenz

- Gezielte Suche nach systematischen Fehlern
- Abgleich mit parallel erschlossenen Publikationen

- Analyse der Stichproben
- Einordnung in Fehlerklassen

DEUTSCHE NATIONAL BIBLIOTHEK

Stichprobenprüfung Bewertung

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

Vierstufige Bewertungsskala:
sehr nützlich / nützlich / wenig nützlich / falsch

005 Ts1
006 http://d-nb.info/gnd/4776222-6
011 s
035 gnd/4776222-6
039 swd/4776222-6$vzg
065 31.9a
083 621.31$d2$t2009-09-29
150 **Nachstrom**
550 !040707458!*Elektrischer Strom*$4obal
670 Lex. Elektrotechn.

DEUTSCHE NATIONAL BIBLIOTHEK

Stichprobenprüfung Ergänzung FA

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	

DEUTSCHE NATIONAL BIBLIOTHEK

Wörterbuch
Aufbau und Pflege

- regelmäßiges Einspielen der GND-Terminologie, wahlweise als Voll-Update oder als inkrementelles Update
- Pflege der Wörterbuch-Profile (Terminologie-Plattform): bevorzugte Benennungen, Synonyme, komplette Datensätze oder auch ganze Teilbäume der GND können hinsichtlich der gewünschten Verarbeitungsform markiert werden

DEFAULT – Term wird segmentiert

EXACT – Term wird nicht segmentiert (also in der exakten Schreibweise prozessiert)

IGNORE – Term wird „stillgelegt" (also nicht in das Wörterbuch eingelesen)

DEUTSCHE NATIONAL BIBLIOTHEK

Ausblick maschinelle Schlagwortvergabe

- Auslieferung der maschinell ermittelten Schlagwörter
- Verbesserung des Verfahrens
- Ausweitung auf andere Netzpublikationstypen wie E-Books, Zeitschriftenartikel etc.
- Weiterentwicklung für englischsprachige Netzpublikationen z.B. über Crosskonkordanzen GND – LCSH

DEUTSCHE NATIONAL BIBLIOTHEK

LCSH – GND

Monetary policy

URI(s)...
Instance Of....
Scheme Membership(s)
Library of Congress Subject Headings
Collection Membership(s)...
Variants
Monetary management
Broader Terms
Economic policy
Narrower Terms
Credit control
Devaluation of currency
Dollarization
Inflation targeting
Open market operations
Quantitative easing (Monetary policy)
Transmission mechanism (Monetary policy)
Unemployment--Effect of monetary policy on
Related Terms
Currency boards
Money supply

Exact Matching Concepts from Other Schemes
monetary policy
Closely Matching Concepts from Other Schemes
Geldmengenpolitik
Geldmengensteuerung
Geldpolitik
Geldverfassung
Notenbankpolitik
Politica monetaria
Politique monétaire
Währungspolitik
Währungssystem
LC Classification
HG230.3
Change Notes
1986-02-11: new
1998-01-09: revised
Alternate Formats ...

DEUTSCHE NATIONAL BIBLIOTHEK

Ausblick maschinelle Schlagwortvergabe

- Auslieferung der maschinell ermittelten Schlagwörter
- Verbesserung des Verfahrens
- Ausweitung auf andere Netzpublikationstypen wie E-Books, Zeitschriftenartikel etc.
- Weiterentwicklung für englischsprachige Netzpublikationen z.B. über Crosskonkordanzen GND – LCSH
- Weiterentwicklung für gescannte Inhaltsverzeichnisse
- Vorschlagstool zur GND-Pflege

DEUTSCHE NATIONAL BIBLIOTHEK

Vielen Dank

für Ihre Aufmerksamkeit

Elisabeth Moedden

Automatische Erschließungsverfahren, Netzpublikationen

Deutsche Nationalbibliothek

Telefon: +49-69-1525-1533

E-Mail: e.moedden@dnb.de