Matthias Nagelschmidt

Die automatische Erschließung in der Deutschen Nationalbibliothek

DEUTSCHE NATIONAL BIBLIOTHEK

Inhalt

- Automatische Erschließung in der Entwicklung
  - Das PETRUS-Projekt und die Folgen
- Automatische Erschließung in der Routine
  - Verfahrensklassen
  - Was wird wie erschlossen?
  - Produktionsprozess
  - Produktionsvolumen
- Funktionsweisen der automatischen Erschließung
  - Was passiert bei der maschinellen Beschlagwortung?
  - Was passiert bei der maschinellen Klassifikation?
- Weiterentwicklung der automatischen Erschließung

DEUTSCHE NATIONAL BIBLIOTHEK

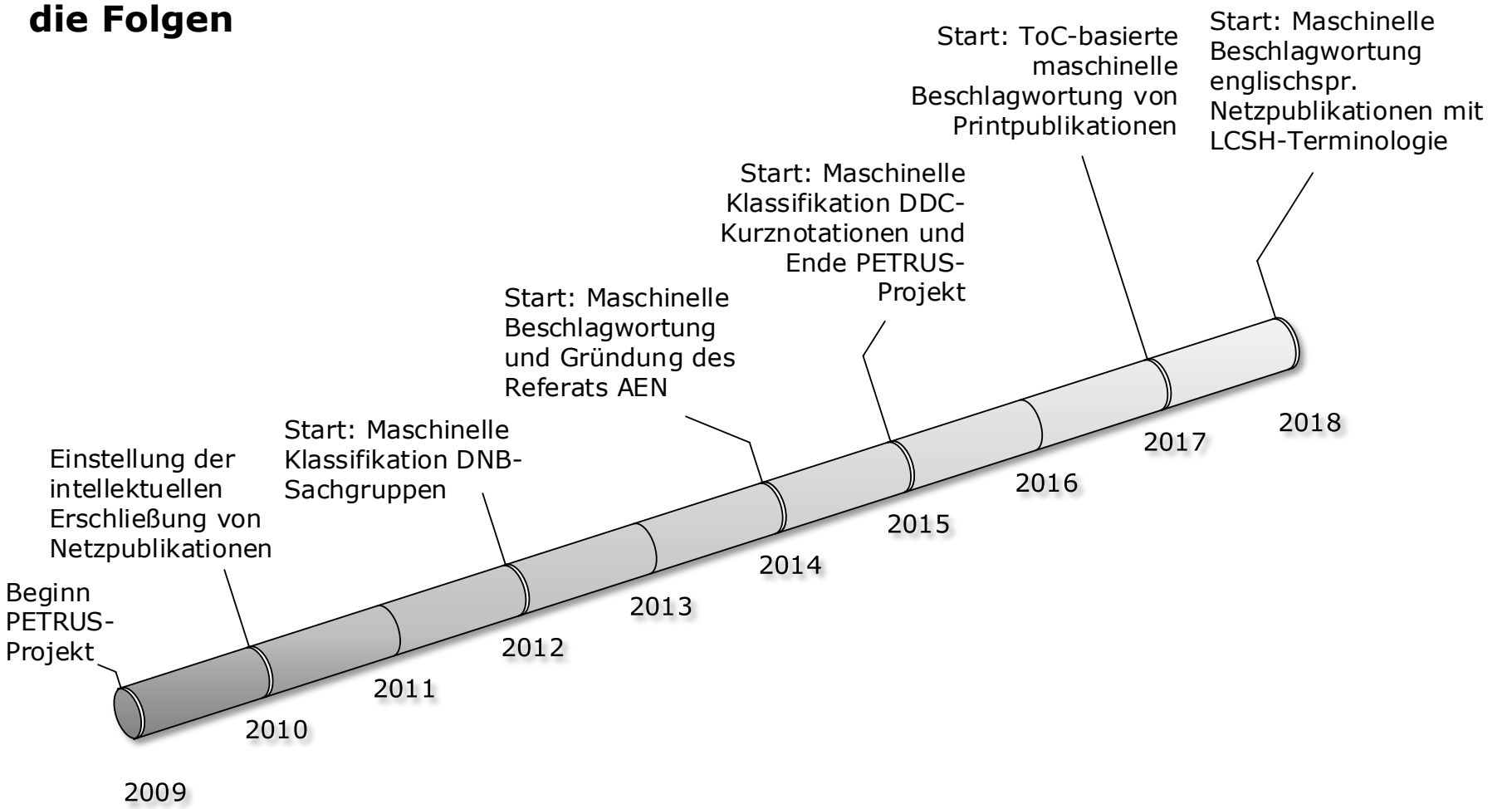
Automatische Erschließung in der Entwicklung

DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließung in der Entwicklung: Das PETRUS-Projekt und die Folgen

- „Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek".
- Projektlaufzeit von 2009 bis 2015.
- Projektinhalte: Konzeption, Entwicklung und Implementierung der technischen und prozeduralen Verfahren für die automatische Erschließung.

Automatische Erschließung in der Entwicklung: Das PETRUS-Projekt und die Folgen

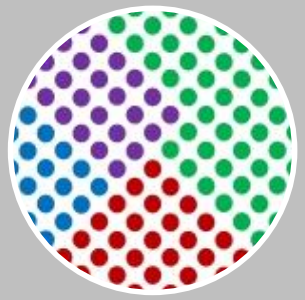


DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließung in der Routine

DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließung in der Routine: Verfahrensklassen



Maschinelles Klassifizieren von Netz- und ausgewählten Printpublikationen anhand der DNB-Sachgruppen (Clustering-Algorithmen, Support Vector Machines)

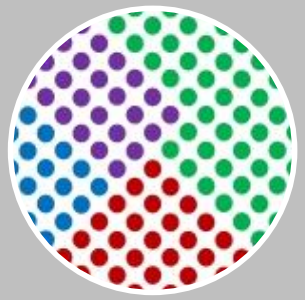
Maschinelles Beschlagworten von Netz- und ausgewählten Printpublikationen anhand der normierten Terminologien GND und LCSH (Text Mining, Textstatistik)

Maschinelles Identifizieren und Verknüpfen parallel erscheinender Print- und Netzpublikationen (Match- & Merge-Verfahren bibliografischer Metadaten)

Inhaltserschließung

Formalerschließung

Automatische Erschließung in der Routine: Verfahrensklassen



Maschinelles Klassifizieren von Netz- und ausgewählten Printpublikationen anhand der DNB-Sachgruppen (Clustering-Algorithmen, Support Vector Machines)

Maschinelles Beschlagworten von Netz- und ausgewählten Printpublikationen anhand der normierten Terminologien GND und LCSH (Text Mining, Textstatistik)

Maschinelles Identifizieren und Verknüpfen parallel erscheinender Print- und Netzpublikationen (Match- & Merge-Verfahren bibliografischer Metadaten)

Inhaltserschließung

Formalerschließung

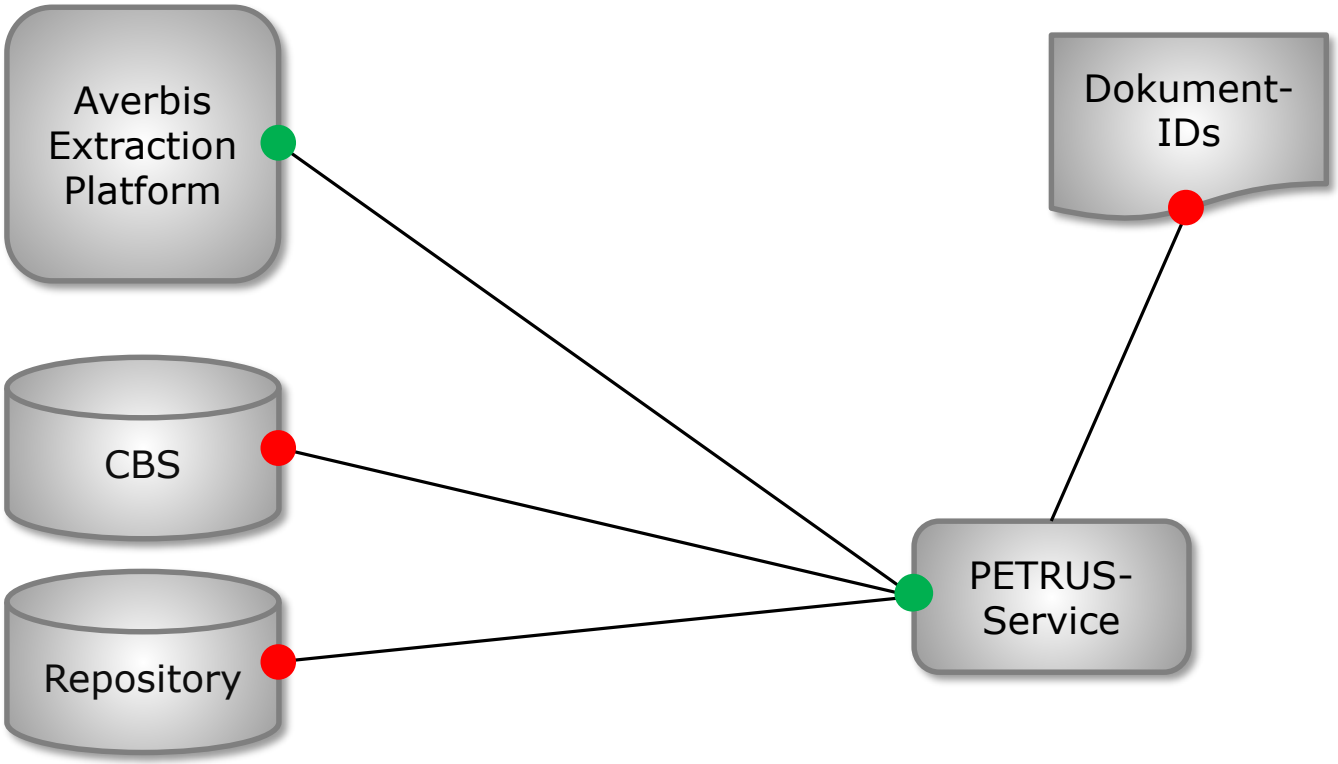
DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließung in der Routine: Was wird wie erschlossen?

		Publikationen			
		Print			Digital
		Reihe A	Reihe B	Reihe H	Reihe O
 Masch. Klassifizieren	Sachgruppen				
	DDC-Kurz				
	DDC				
 Masch. Beschlagworten	Schlagwörter				

DEUTSCHE NATIONAL BIBLIOTHEK

Automatische Erschließung in der Routine: Produktionsprozess



Automatische Erschließung in der Routine: Produktionsvolumen

Maschinelle Sachgruppen

Reihe O	Deutsch- und englischsprachige Netzpublikationen (Monografien, Zeitschriftenartikel)*
Reihe B Reihe H	Printpublikationen mit digitalisierten ToCs
Methode	Geometrischer Ansatz, Support Vector Machines
Volumen (08/2018)	ca. 1,5 Mio. Datensätze

*Von der Verarbeitung ausgeschlossen sind Zeitschriftentiteldatensätze und Publikationen der Sachgruppe B (Belletristik).

Maschinelle DDC-Kurznotationen

Reihe O	Deutsch- und englischsprachige Netzpublikationen (Monografien, Zeitschriftenartikel)* mit den Sachgruppen 004 (Informatik), 300 (Sozialwissenschaften), 540 (Chemie) und 610 (Medizin)
Reihe B Reihe H	Printpublikationen mit digitalisierten ToCs nur mit Sachgruppe 610 (Medizin)
Methode	Geometrischer Ansatz, Support Vector Machines
Volumen (08/2018)	ca. 376 Tsd. Datensätze

Automatische Erschließung in der Routine: Produktionsvolumen

Maschinelle Schlagwörter für das Deutsche

Reihe O	Deutschsprachige Netzpublikationen (Monografien, Zeitschriftenartikel)*
Reihe B Reihe H	Printpublikationen mit digitalisierten ToCs
Methode	Text Mining, Kombination etablierter statistischer, algorithmischer und lexikonbasierter Verfahren
Volumen (08/2018)	ca. 242 Tsd. Datensätze

Maschinelle Schlagwörter für das Englische

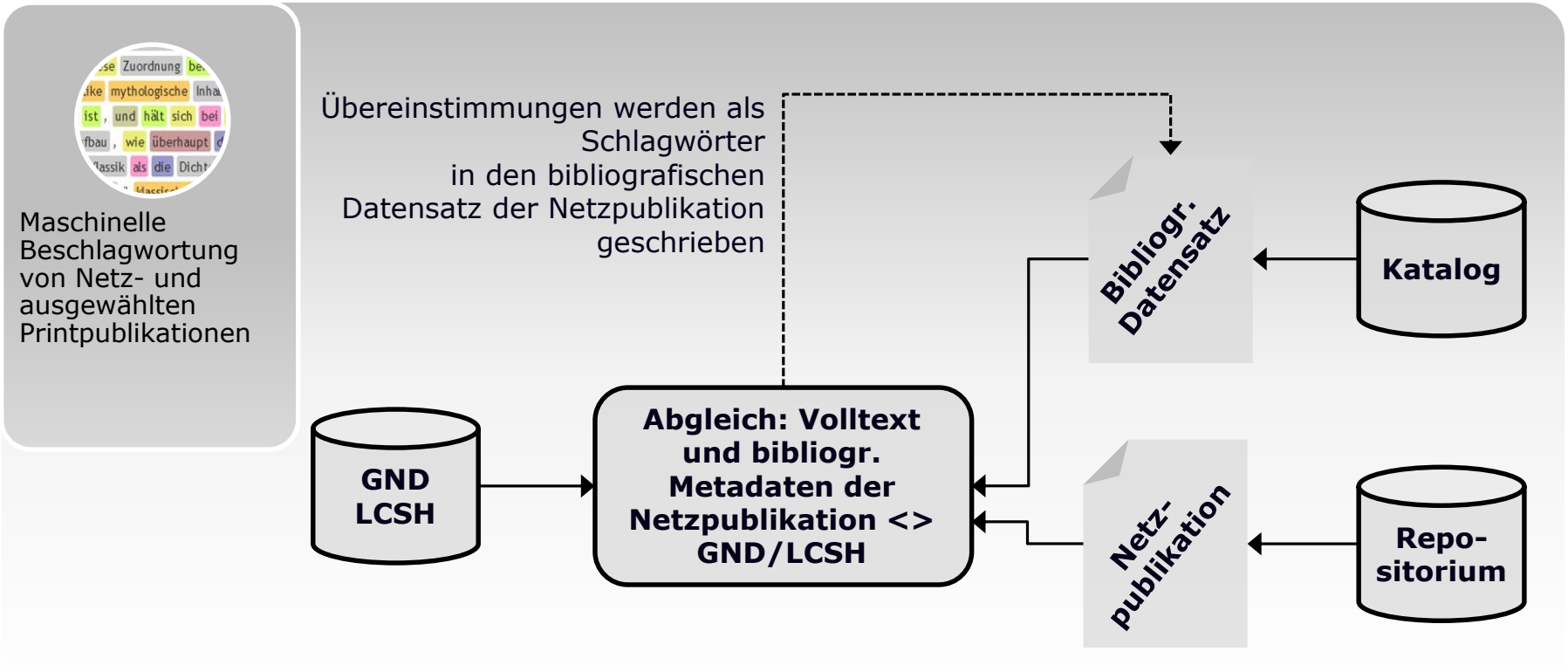
Reihe O	Englischsprachige Netzpublikationen (Monografien, Zeitschriftenartikel)*
Methode	Text Mining, Kombination etablierter statistischer, algorithmischer und lexikonbasierter Verfahren
Volumen (08/2018)	2.220 Datensätze

*Von der Verarbeitung ausgeschlossen sind Zeitschriftentiteldatensätze und Publikationen der Sachgruppe B (Belletristik).

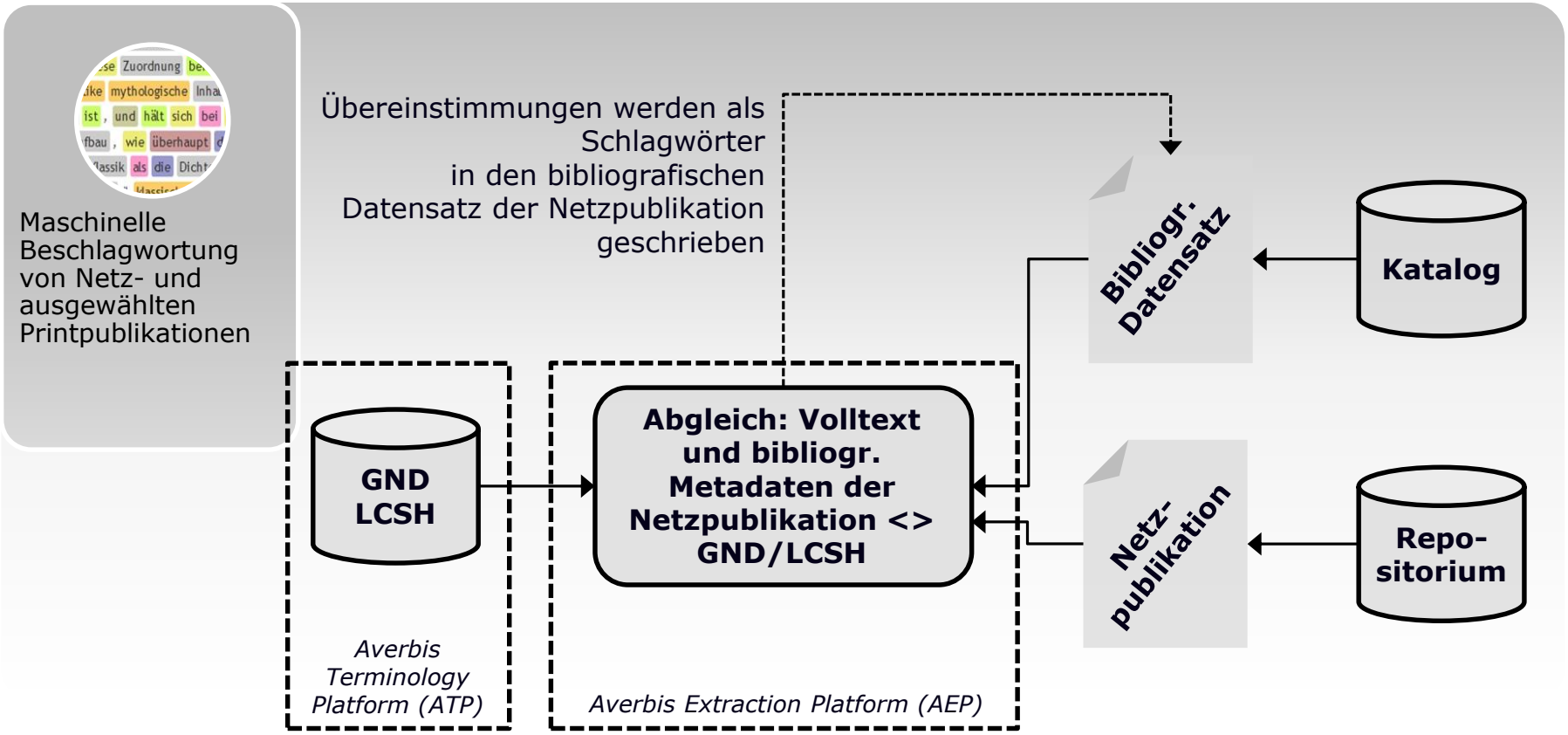
DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

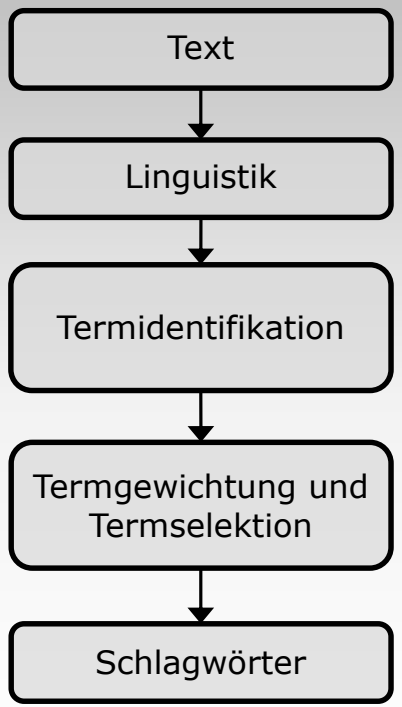


Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?



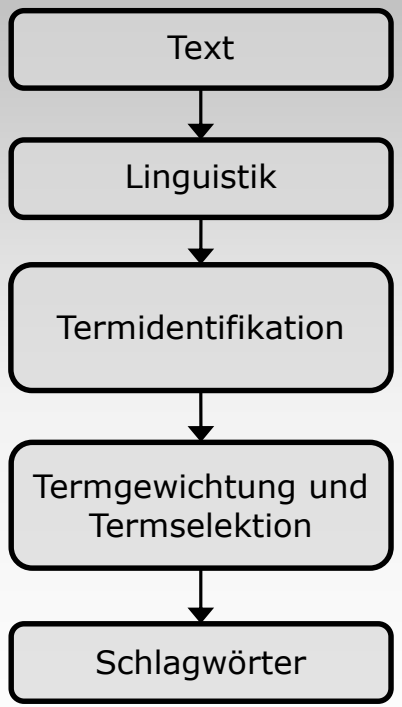
Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen



Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

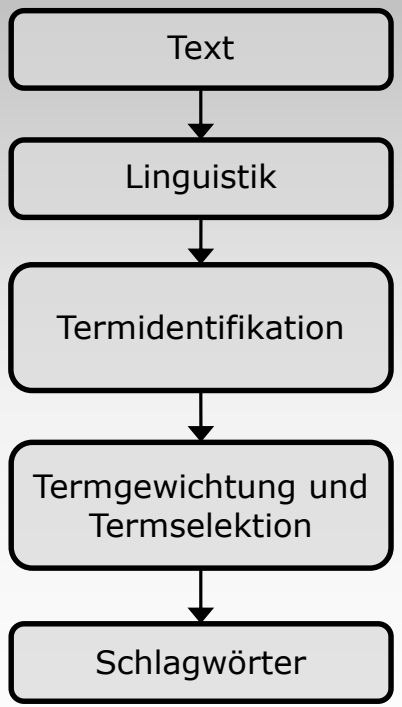
Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen



Elektronischen Volltext der Netzpublikation einlesen.

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

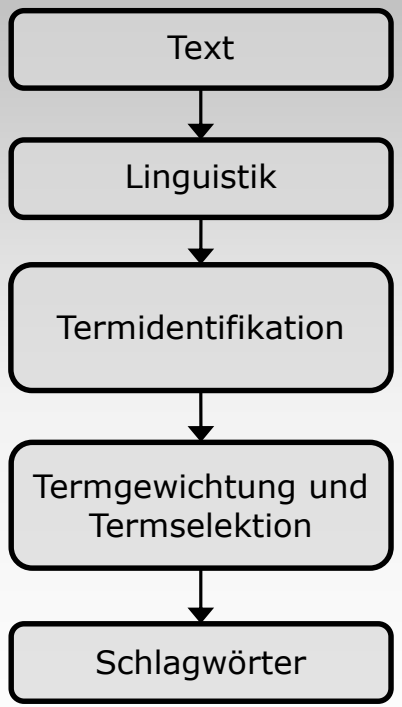


Elektronischen Volltext der Netzpublikation einlesen.

Spracherkennung, Satzerkennung, Wort- und Wortartenerkennung. Bereitstellung potenzieller Schlagwortkandidaten als Nominalphrasen.

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen



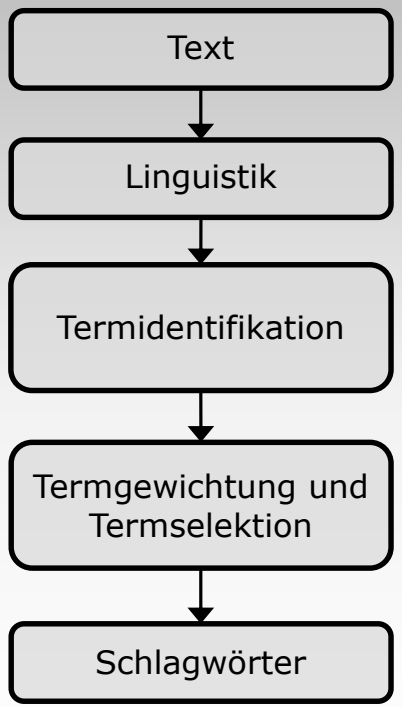
Elektronischen Volltext der Netzpublikation einlesen.

Spracherkennung, Satzerkennung, Wort- und Wortartenerkennung. Bereitstellung potenzieller Schlagwortkandidaten als Nominalphrasen.

Abgleich der Nominalphrasen mit GND-Vokabular.

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen



Elektronischen Volltext der Netzpublikation einlesen.

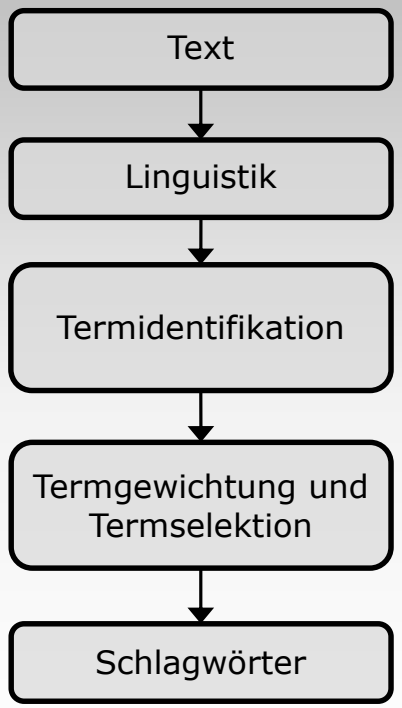
Spracherkennung, Satzerkennung, Wort- und Wortartenerkennung. Bereitstellung potenzieller Schlagwortkandidaten als Nominalphrasen.

Abgleich der Nominalphrasen mit GND-Vokabular.

Gewichtung der abgeglichenen Nominalphrasen nach versch. Kriterien (Position im Text, Auftretenshäufigkeit etc.).Berechnung eines Konfidenzwertes und Definition eines Konfidenz-Schwellenwertes.

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen



Elektronischen Volltext der Netzpublikation einlesen.

Spracherkennung, Satzerkennung, Wort- und Wortartenerkennung. Bereitstellung potenzieller Schlagwortkandidaten als Nominalphrasen.

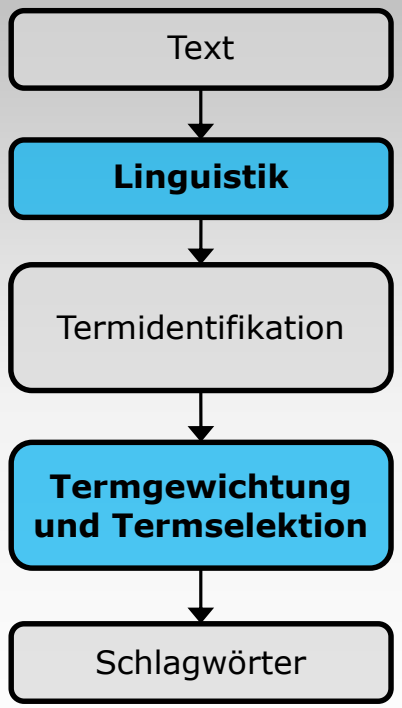
Abgleich der Nominalphrasen mit GND-Vokabular.

Gewichtung der abgeglichenen Nominalphrasen nach versch. Kriterien (Position im Text, Auftretenshäufigkeit etc.).Berechnung eines Konfidenzwertes und Definition eines Konfidenz-Schwellenwertes.

Die *n* Terme mit dem höchsten Konfidenzwert und oberhalb des Konfidenz-Schwellenwertes werden als Schlagwörter ausgegeben.

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen



Elektronischen Volltext der Netzpublikation einlesen.

Spracherkennung, Satzerkennung, Wort- und Wortartenerkennung. Bereitstellung potenzieller Schlagwortkandidaten als Nominalphrasen.

Abgleich der Nominalphrasen mit GND-Vokabular.

Gewichtung der abgeglichenen Nominalphrasen nach versch. Kriterien (Position im Text, Auftretenshäufigkeit etc.).Berechnung eines Konfidenzwertes und Definition eines Konfidenz-Schwellenwertes.

Die *n* Terme mit dem höchsten Konfidenzwert und oberhalb des Konfidenz-Schwellenwertes werden als Schlagwörter ausgegeben.

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen. Obwohl eine Vielzahl der Myokarditen …

Eingabe

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen. Obwohl eine Vielzahl der Myokarditen …

Eingabe

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen.

Sentence Detector

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen. Obwohl eine Vielzahl der Myokarditen …

Eingabe

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen.

Sentence Detector

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels

Tokenizer

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen. Obwohl eine Vielzahl der Myokarditen …	*Eingabe*
Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen.	*Sentence Detector*
Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels	*Tokenizer*
Die {**ART**} Myokarditis {**NN**} ist {**VAFIN**} eine {**ART**} Sammelbezeichnung {**NN**} für {**APPR**} entzündliche {**ADJA**} Erkrankungen {**NN**} des {**ART**} Herzmuskels {**NN**}	*POS-Tagger*

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen. Obwohl eine Vielzahl der Myokarditen …	*Eingabe*
Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen.	*Sentence Detector*
Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels	*Tokenizer*
Die {**ART**} Myokarditis {**NN**} ist {**VAFIN**} eine {**ART**} Sammelbezeichnung {**NN**} für {**APPR**} entzündliche {**ADJA**} Erkrankungen {**NN**} des {**ART**} Herzmuskels {**NN**}	*POS-Tagger*
Die Myokarditis {***myo kard itis***} ist eine Sammelbezeichnung {***sammel bezeich***}	*Segments*

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Ergebnisse im bibliografischen Datensatz

IDN	1052530672
…	
4000	Vergleich der Single-Port-Laparoskopie mit der konventionellen Multiport Laparoskopie bei ausgewählten urologischen Operationen [[Elektronische Ressource]] / Seven Johannes Sam Aghdassi
…	
5051	$KK_A4_02_20140123_de$LS_WA3_WB38_20160510_de
5540	[GND]!041401786!Laparoskopie
5540	[GND]!040756645!Operation
5540	[GND]!041718895!Nierenzyste
5540	[GND]!040012409!Alleinstehender
5540	[GND]!041235479!Komplikation
5540	[GND]!042193397!Spätkomplikation

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Ergebnisse im bibliografischen Datensatz

IDN	1052530672
…	
4000	Vergleich der Single-Port-Laparoskopie mit der konventionellen Multiport Laparoskopie bei ausgewählten urologischen Operationen [[Elektronische Ressource]] / Seven Johannes Sam Aghdassi
…	
5051	$KK_A4_02_20140123_de$LS_WA3_WB38_20160510_de
5540	[GND]!041401786!Laparoskopie
5540	[GND]!040756645!Operation
5540	[GND]!041718895!Nierenzyste
5540	[GND]!040012409!**Alleinstehender**
5540	[GND]!041235479!Komplikation
5540	[GND]!042193397!Spätkomplikation

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Ergebnisse im bibliografischen Datensatz

IDN	1052530672
...	
4000	Vergleich der **Single**-Port-Laparoskopie mit der konventionellen Multiport Laparoskopie bei ausgewählten urologischen Operationen [[Elektronische Ressource]] / Seven Johannes Sam Aghdassi
...	
5051	$KK_A4_02_20140123_de$LS_WA3_WB38_20160510_de
5540	[GND]!041401786!Laparoskopie
5540	[GND]!040756645!Operation
5540	[GND]!041718895!Nierenzyste
5540	[GND]!040012409!**Alleinstehender**
5540	[GND]!041235479!Komplikation
5540	[GND]!042193397!Spätkomplikation

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Ergebnisse im bibliografischen Datensatz

IDN	1052530672
…	
4000	Vergleich der Single-Port-Laparoskopie mit der konventionellen Multiport Laparoskopie bei ausgewählten urologischen Operationen [[Elektronische Ressource]] / Seven Johannes Sam Aghdassi
…	
5051	$KK_A4_02_20140123_de$LS_WA3_WB38_20160510_de
5540	[GND]!041401786!Laparoskopie
5540	[GND]!040756645!Operation
5540	[GND]!041718895!Nierenzyste
5540	[GND]!040012409!Alleinstehender
5540	[GND]!041235479!**Komplikation**
5540	[GND]!042193397!**Spätkomplikation**

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Beschlagwortung?

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

Zwei mögliche Fehlerklassen:

- Maschinelles Schlagwort fehlt (inhaltlicher Aspekt wird nicht repräsentiert
- Maschinelles Schlagwort ist falsch (inhaltlicher Aspekt wird falsch repräsentiert)

Viele mögliche Ursachen:

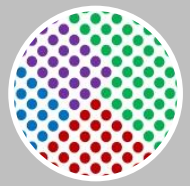
- Kein geeignetes GND-/LCSH-Schlagwort vorhanden?
- Geeignetes GND/LCSH-Schlagwort ist vorhanden, wird aber nicht berücksichtigt?
- Falsches GND/LCSH-Schlagwort durch Disambiguierungsfehler?
- Falsches GND/LCSH-Schlagwort durch Parametrisierung der Beschlagwortung?

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Klassifikation?

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Klassifikation?

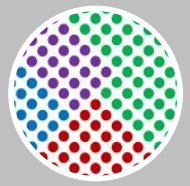


Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

- Erzeugung von Dokument-Vektoren durch Merkmalsextraktion („Bag-of-Words"-Verfahren),
- Abbildung der Dokument-Vektoren im *n*-dimensionalen Vektorraum,
- Training: Beschreibung idealtypischer Dokumentenmerkmale für einzelne DNB-Sachgruppen,
- maschinelles Klassifizieren: Anwendung des Trainingsmodells auf neue Dokumente.

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Klassifikation?

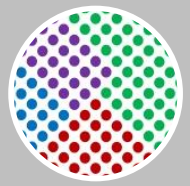


Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

- Klassifiziert wird anhand von Support-Vector-Machines (SVMs), die auf den Grenzverlauf zwischen zwei Merkmalsräumen abstellen.
- Es werden solche Dokumentvektoren identifiziert, die am äußeren Rand ihres jeweiligen Merkmalsraumes liegen („Support Vectors“).
- Anschließend wird der Abstand zwischen diesen Support-Vektoren maximiert (Prinzip des „Large-Margin-Classifiers“).
- Es entstehen homogene, trennscharfe Merkmalsräume, die als „Klassen“ interpretiert werden können.

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Klassifikation?

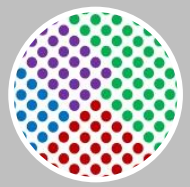


Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

- DDC-Kurznotationen: Über 2.400 DDC-Notationen.
- Gebildet anhand der DDC-Abridged oder nach eigenem Schema.
- Produktive Tests für die Informatik (Sachgruppe 004), die Sozialwissenschaften (Sachgruppe 300), die Chemie (Sachgruppe 540) und die Medizin (Sachgruppe 610).
- Anzahl der DDC-Kurznotationen pro Sachgruppe variiert zwischen 15 (für die Chemie) und 120 (für die Medizin).

DEUTSCHE NATIONAL BIBLIOTHEK

Funktionsweisen der automatischen Erschließung: Was passiert bei der maschinellen Klassifikation?



Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

DDC-Kurznotation im bibliografischen Datensatz

IDN	1131599683
…	
4000	Die Sarkoidose als Systemerkrankung / von R. Bergner, P. Korsten
…	
5051	$KK_A8_03_20160930_de$LS_ART2_WB41_20170313_de $MMK_610_A2_06_20161020_de
…	
5470	[MKN]**616.7**$K0,894$D2017-05-09 *Krankheiten des Bewegungsapparats

Weiterentwicklung der automatischen Erschließung

Weiterentwicklung der automatischen Erschließung

Maschinelle Beschlagwortung von Netz- und ausgewählten Printpublikationen

- **ToC-Experimente** (läuft seit 04/2017): Maschinelle Beschlagwortung von Printpublikationen der Reihen B und H – Indexierung mit reduzierter Datenbasis (bibliografische Metadaten und gescannte Inhaltsverzeichnisse anstatt Volltexte),
- **Projekt MAEN** (abgeschlossen in 06/2018): Maschinelle Beschlagwortung englischsprachiger Netzpublikationen – Indexierung englischsprachiger Netzpublikationen mit LCSH-Terminologie,
- **Projekt TeMa** (läuft seit 04/2017): Terminologiemanagement zur Unterstützung der intellektuellen und automatischen Inhaltserschließung – Harmonisierung der Erschließungsarbeit.

Vielen Dank für die Aufmerksamkeit!

Matthias Nagelschmidt
Deutsche Nationalbibliothek
Automatische Erschließungsverfahren, Netzpublikationen
Deutscher Platz 1
04103 Leipzig
+49 341 2271-541
m.nagelschmidt@dnb.de