DEUTSCHE NATIONAL BIBLIOTHEK

Frank Busse

Einführung in die maschinelle Erschließung

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

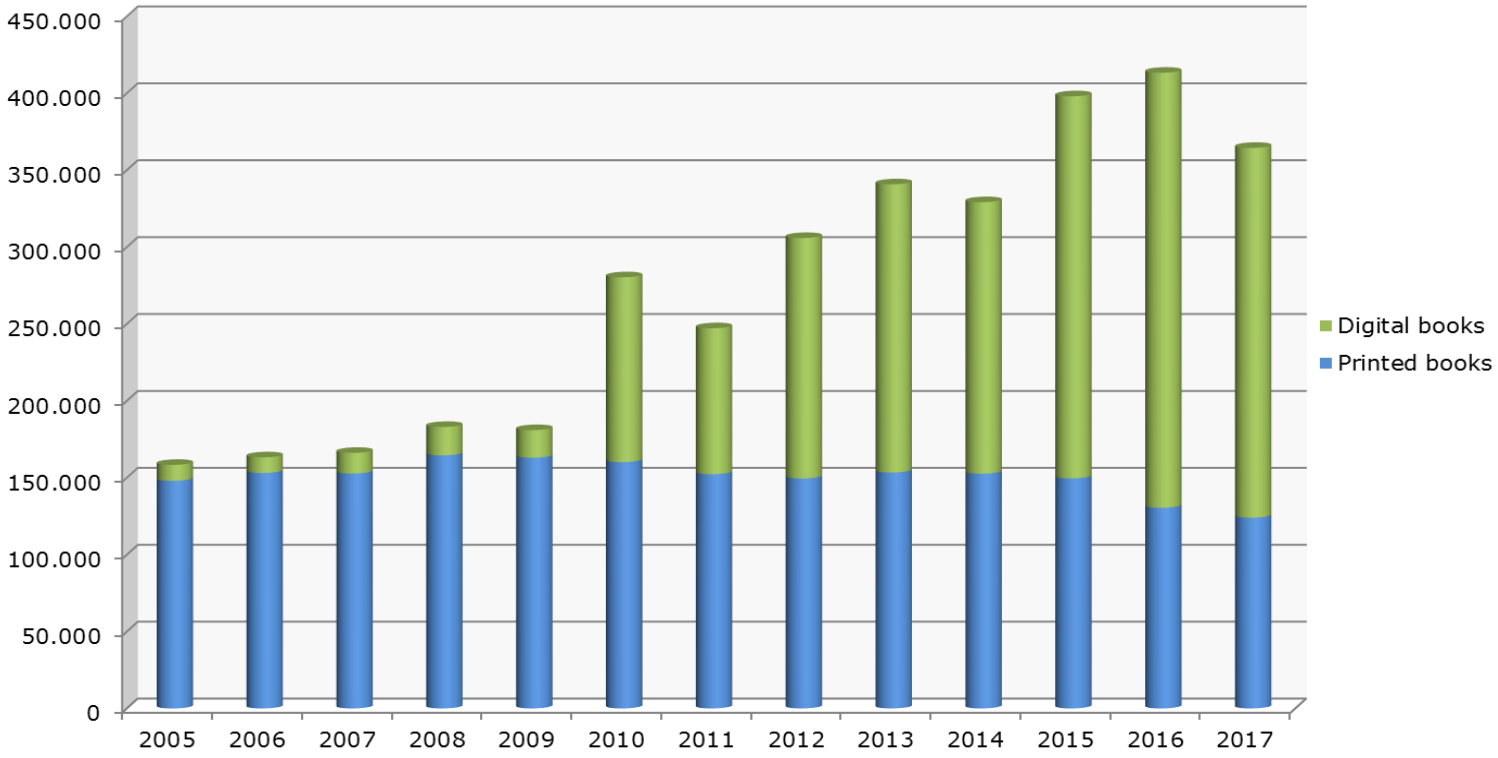
DEUTSCHE NATIONAL BIBLIOTHEK

Petrus-Projekt

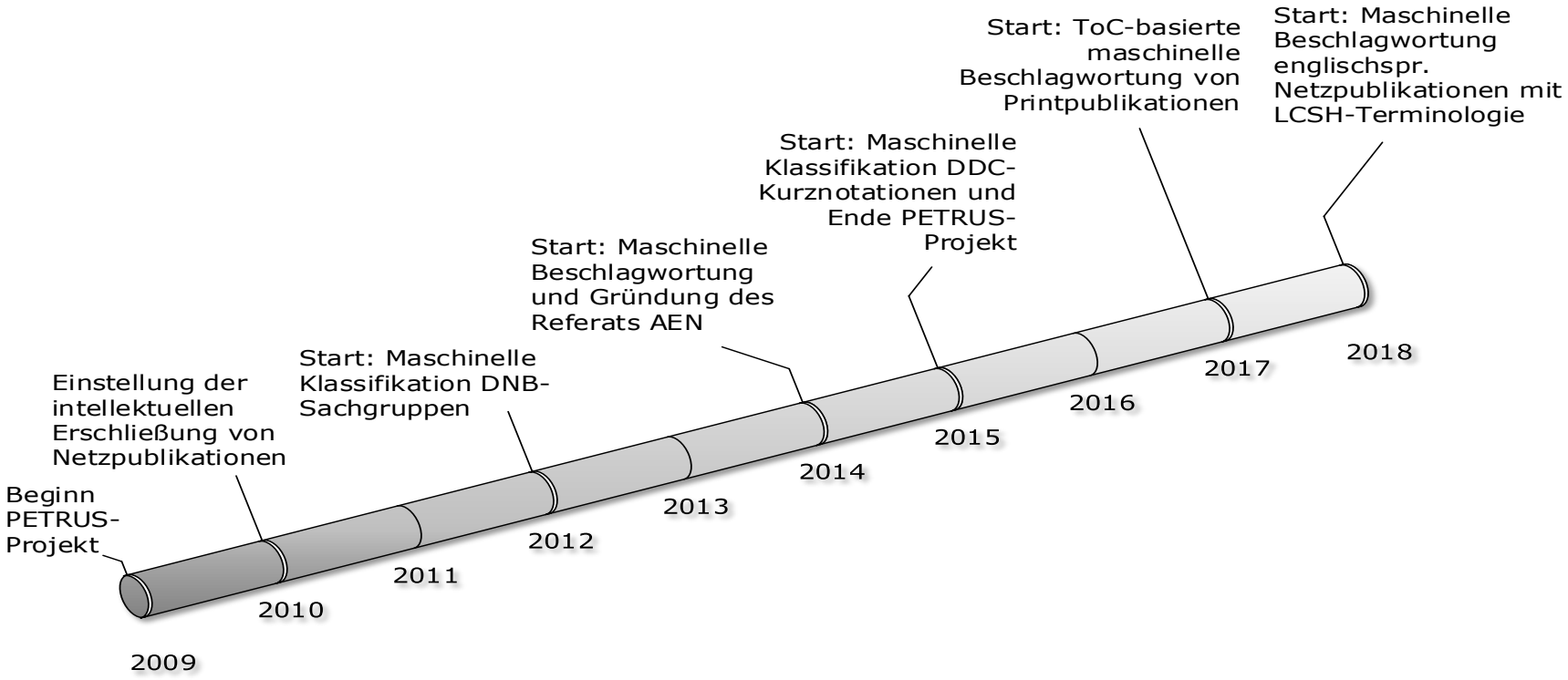
- Laufzeit: 2009 - 2015
- Prozessunterstützende Software für die Digitale Deutsche Nationalbibliothek
- Konzeption, Entwicklung und Implementierung der technischen und prozeduralen Verfahren für die automatische Erschließung
- Aufteilung des Projektes in 4 Anwendungsszenarien

DEUTSCHE NATIONAL BIBLIOTHEK

Zugang monografischer Print- und Netzpublikationen 2005–2017

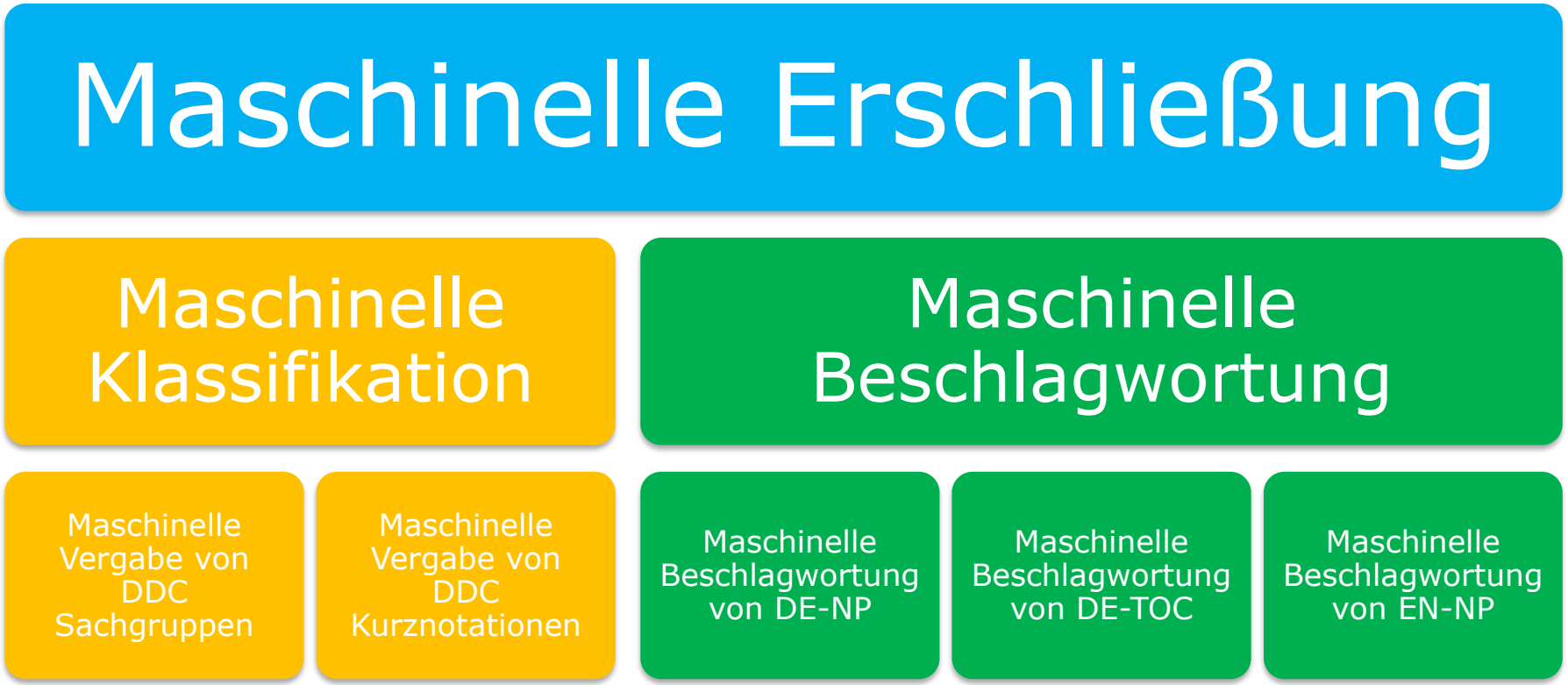


Maschinelle Erschließung in der Entwicklung



DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Erschließung: Anwendungsfälle



DEUTSCHE NATIONAL BIBLIOTHEK

Exkurs: DDC-Kurznotionen

- Begrenztes Set von DDC-Notationen für eine bestimmte Sachgruppe
- Ursprünglich 2005/2006 für die Sachgruppe Medizin zur Erschließung gedruckter medizinischer Dissertationen entwickelt
- 2015 maschinelle Vergabe von medizinischen Kurznotationen für Netzpublikationen
- 2017 Weiterentwicklung und Ausweitung auf weitere Sachgruppen

Exkurs: DDC-Kurznotationen II

Thema:
Studie
Übergewicht bei Kindern
Kiel
2000-2009

DNB-SG	610
DDC	618.92398009435123090511
Kurznotation	618.92~~398009435123090511~~

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Klassifikation

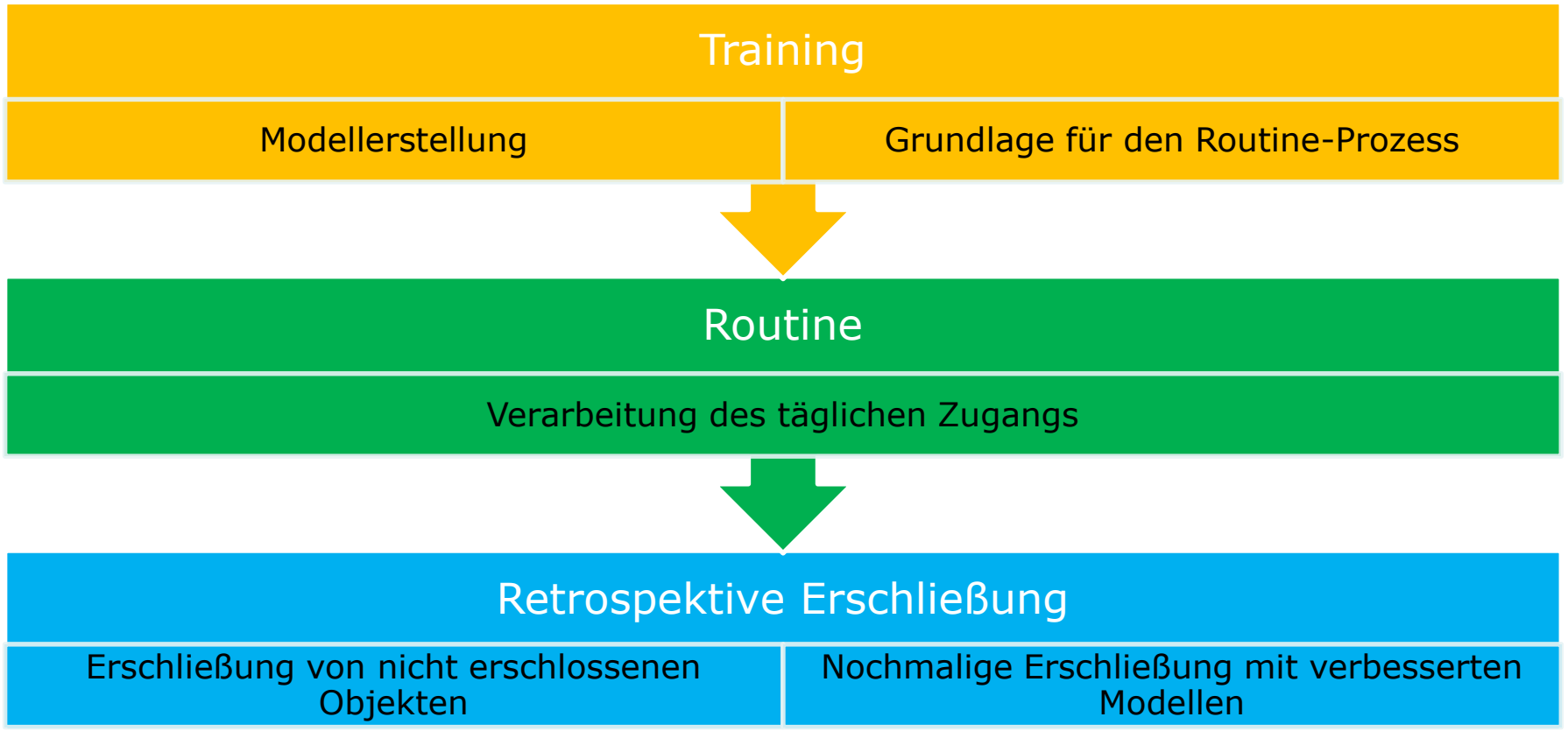
- Start: 2012 Sachgruppe / 2015 Kurznotationen
- Methode: Maschinelles Lernen / SVM
- Averbis-Extraction Plattform (AEP)
- Dokumentarten:
  - Alle NPs ohne Belletristik
  - Formate PDF (2012) & Epub (2015)
  - Sprache Ger/Eng
- Umfang: 1.651.965 Publikationen (09/2018)

DEUTSCHE NATIONAL BIBLIOTHEK

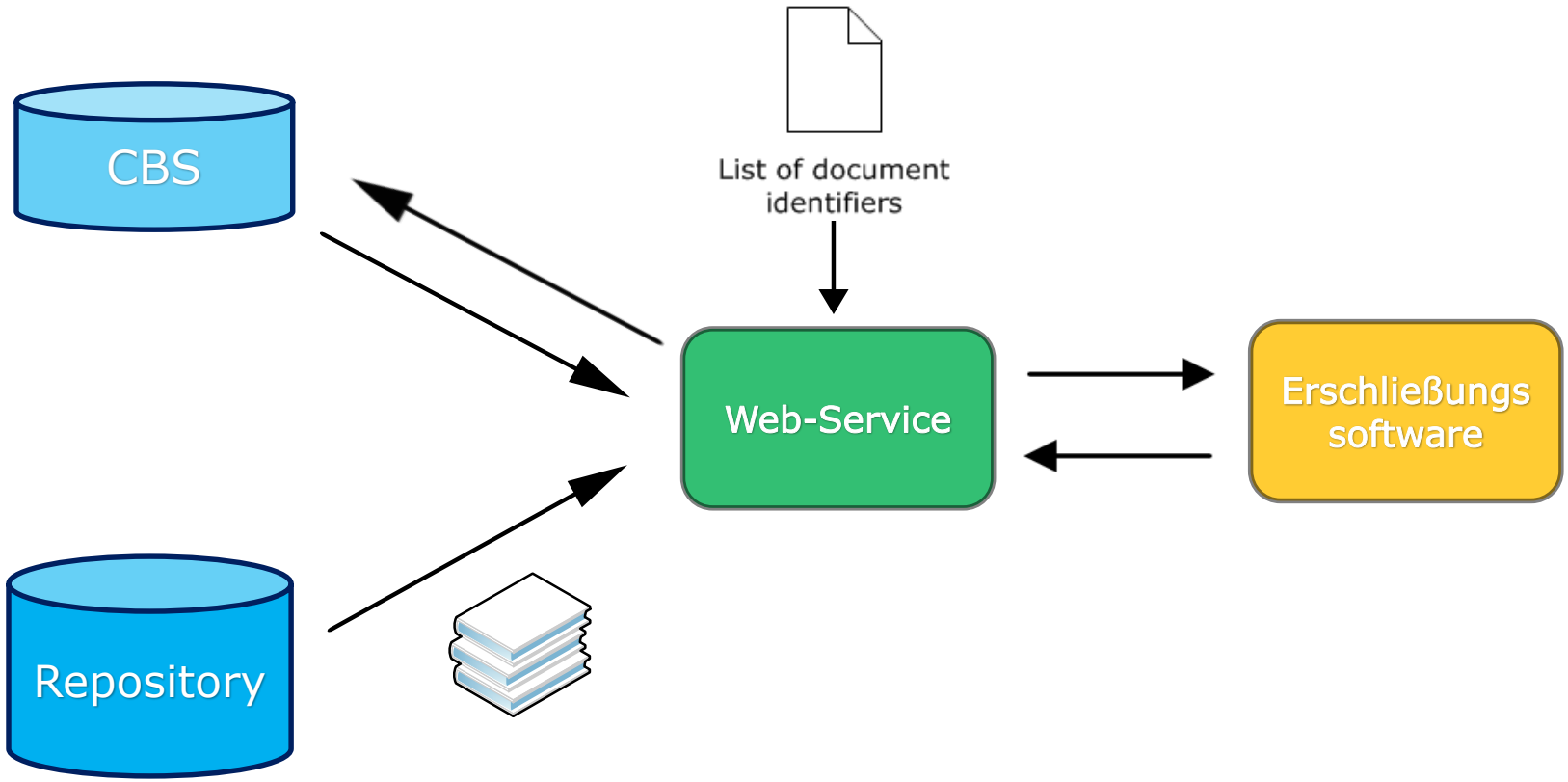
Maschinelles Lernen

- Lernen aus Beispielen
- Erkennen von Mustern
- Verallgemeinerung der Muster
- Unbekannte Objekte können klassifiziert werden

Workflows

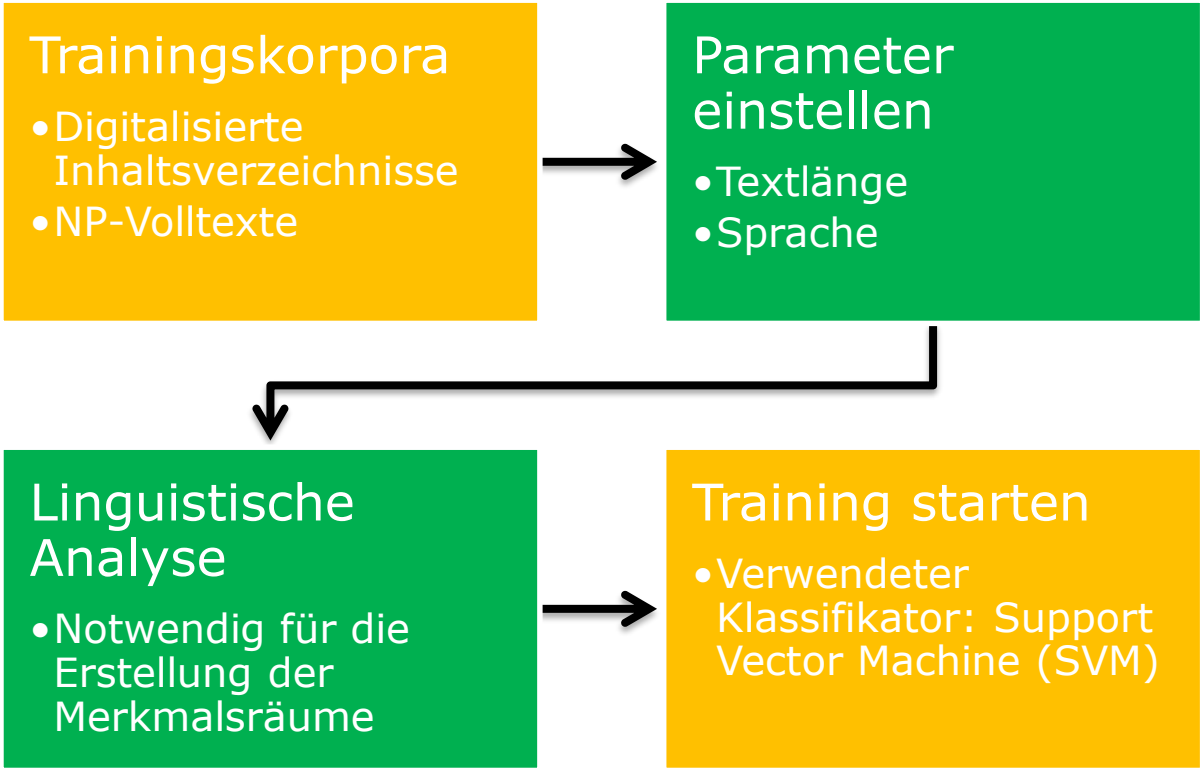

Training
Modellerstellung
Grundlage für den Routine-Prozess
Routine
Verarbeitung des täglichen Zugangs
Retrospektive Erschließung
Erschließung von nicht erschlossenen Objekten
Nochmalige Erschließung mit verbesserten Modellen

Routine Workflow



DEUTSCHE NATIONAL BIBLIOTHEK

Modellerstellung



DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Beschlagwortung

- Ziel: Automatische Beschlagwortung von NPs mit kontrolliertem GND-Vokabular (GND)
- Methode: Wörterbuchbasiert, Computerlinguistik
- Start: 2014
- Dokumentarten:
  - NPs mit maschineller Sachgruppe
  - Sprache Ger/Eng
  - Formate PDF & Epub
- Ausgabe: max. 12 Schlagwörter pro Dokument
- Umfang: 261.115 Publikationen (09/2018)

DEUTSCHE NATIONAL BIBLIOTHEK

Was passiert bei der maschinellen Beschlagwortung?

Text

- Einlesen des elektronischen Volltexts der Netzpublikation

Linguistik

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

Termidentifikation

- Abgleich der Nominalphrasen mit GND-Vokabular

Termgewichtung und Termselektion

- Gewichtung der abgeglichenen Nominalphrasen nach versch. Kriterien

Schlagwörter

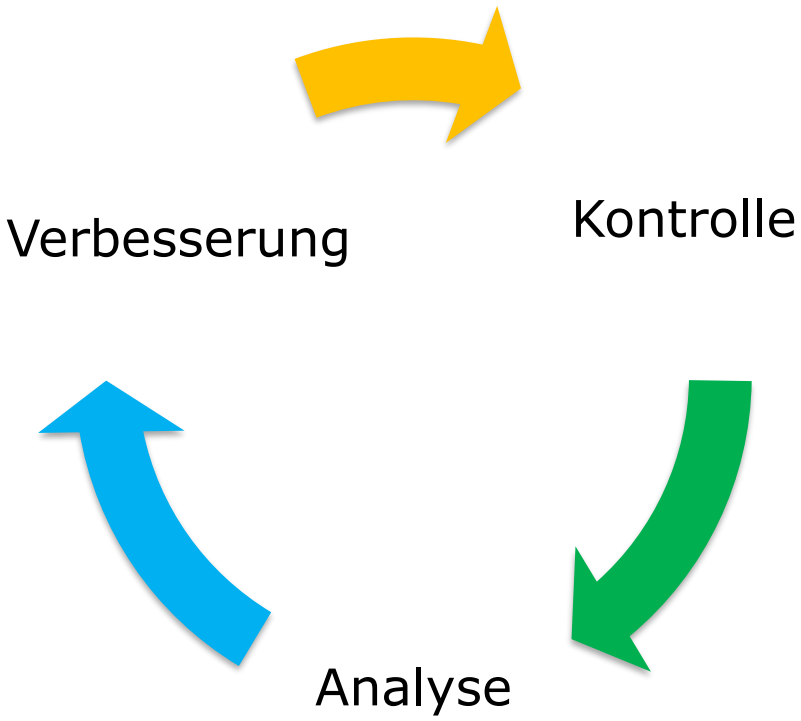
- Ausgabe der Terme mit dem höchsten Konfidenzwert und oberhalb des Konfidenz-Schwellenwertes

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistische Verarbeitung

DEUTSCHE NATIONAL BIBLIOTHEK

Qualitätsmanagement



DEUTSCHE NATIONAL BIBLIOTHEK

Ergebnisse Klassifikation

- Bewertungsgrundlage
  Übereinstimmung zwischen maschineller und intellektuell vergebener Notation
- Stichprobenerhebung
  - Intellektuelle Überprüfung
  - Parallelausgaben

DEUTSCHE NATIONAL BIBLIOTHEK

Ergebnisse 2012 - 2016

- Objekte klassifiziert: 596.773
- Stichprobengröße: 105.912 (18%)
- Ergebnis: 76% Übereinstimmung

DEUTSCHE NATIONAL BIBLIOTHEK

Erste Ergebnisse Kurznotationen

004 Informatik	80% Übereinstimmung
650 Management	72% Übereinstimmung
610 Medizin	68% Übereinstimmung
540 Chemie	67% Übereinstimmung
720 Architektur	67% Übereinstimmung
300 Sozialwissenschaften	66% Übereinstimmung
330 Wirtschaft	62% Übereinstimmung
020 Bibliotheks- u. Informationwiss.	58% Übereinstimmung

Ergebnis Beschlagwortung I

Link	http://d-nb.info/1145338054
Titel	Wohnungsmodernisierung des Vermieters in Deutschland und Frankreich
Person(en)	Schepers, Verena
Sprache(n)	Deutsch (ger)
Schlagwörter	**Vermieter* ; Frankreich* ; Deutschland* ; Wohnungsmodernisierung* ; Mieterhöhung* ; Mietrecht* ; Miete*** (*maschinell ermittelt)
Sachgruppe(n)	**340** Recht ; **330** Wirtschaft

Ergebnis Beschlagwortung II

Link	http://d-nb.info/1140596047
Titel	Heiraten in Würzburg
Organisation(en)	Stadt Würzburg
Sprache(n)	Deutsch (ger)
Schlagwörter	**Würzburg* ; Eheschließung* ; Rathaus* ; Adressbuch* ; Unternehmen*** (*maschinell ermittelt)
Sachgruppe(n)	**390** Bräuche, Etikette, Folklore ; **640** Hauswirtschaft und Familienleben

Ergebnis Beschlagwortung III

Link	http://d-nb.info/1140135228
Titel	Ursachen des Studienabbruchs bei Studierenden mit Migrationshintergrund : eine vergleichende Untersuchung der Ursachen und Motive des Studienabbruchs bei Studierenden mit und ohne Migrationshintergrund auf Basis der Befragung der Exmatrikulierten des Sommersemesters 2014
Person(en)	Ebert, Julia ; Heublein, Ulrich
Sprache(n)	Deutsch (ger)
Schlagwörter	**Studienabbruch* ; Student* ; Interview* ; Motiv <Musik>* ; Studienbedingungen* ; Migrationshintergrund* ; Studienverhalten*** (*maschinell ermittelt)
Sachgruppe(n)	370 Erziehung, Schul- und Bildungswesen

DEUTSCHE NATIONAL BIBLIOTHEK

Mögliche Fehlerursachen

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

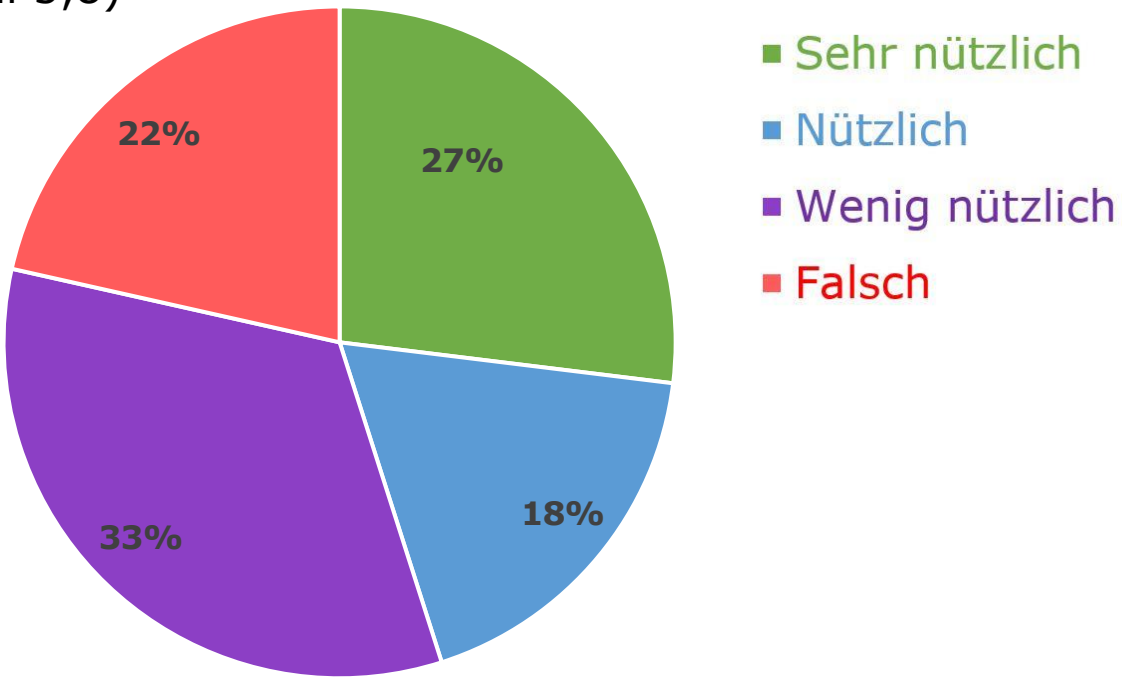
Beurteilung der maschinellen Beschlagwortung

- Intellektuelle Kontrolle
- Das Thema der Publikation muss von der maschinellen Beschlagwortung getroffenen werden
- Bewertung:
  - Sehr Nützlich
  - Nützlich
  - Weniger Nützlich
  - Falsch

Stichprobenergebnisse 2016

1.729 Stichproben **Reihe O 2016**

(10.012 mSW | Ø Anzahl 5,8)



Reihe O (2016) Verteilung der Bewertungen

Vielen Dank für Ihre Aufmerksamkeit!

Frank Busse
Deutsche Nationalbibliothek
Automatische Erschließungsverfahren, Netzpublikationen
Telefon: +49-69-1525-1550
mailto:f.busse@dnb.de