DEUTSCHE NATIONAL BIBLIOTHEK

Frank Busse

Maschinelle Klassifikation in der Deutschen Nationalbibliothek

DEUTSCHE NATIONAL BIBLIOTHEK

Übersicht

- Allgemeines
- Maschinelle Klassifikation
- Workflow
- Kennzeichnung
- Qualitätsmanagement
- Ausblick

Kurz vorgestellt:

Die Deutsche Nationalbibliothek

1912 Gründung der Deutschen Bücherei in Leipzig

1946 Gründung der Deutschen Bibliothek in Frankfurt am Main

1970 Eingliederung des Deutschen Musikarchivs

seit 1990 eine Gesamtinstitution

2006 Gesetz über die Deutsche Nationalbibliothek

- neuer Name
- erweiterter Sammelauftrag

DEUTSCHE NATIONAL BIBLIOTHEK

Sammelauftrag

Alles, was seit 1913 veröffentlicht wurde:

- in Deutschland
- in deutscher Sprache
- Übersetzungen aus dem
- Deutschen
- Werke über Deutschland

Sammelauftrag seit 2006

Körperliche Medienwerke

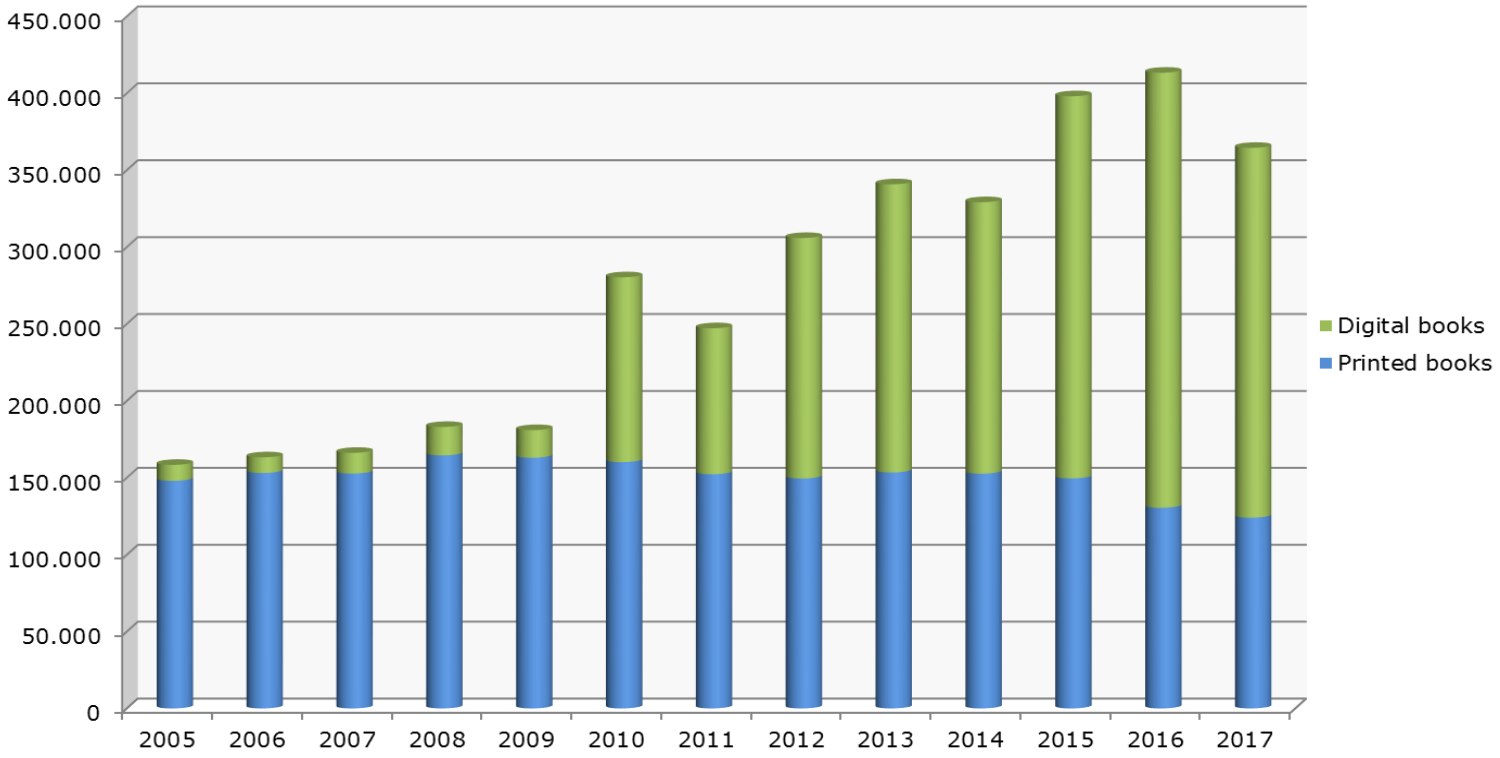
Bücher, Hochschulschriften, Hörbücher, Zeitschriften, Zeitungen, Karten, Musikalien, Tonträger …

Unkörperliche Medienwerke

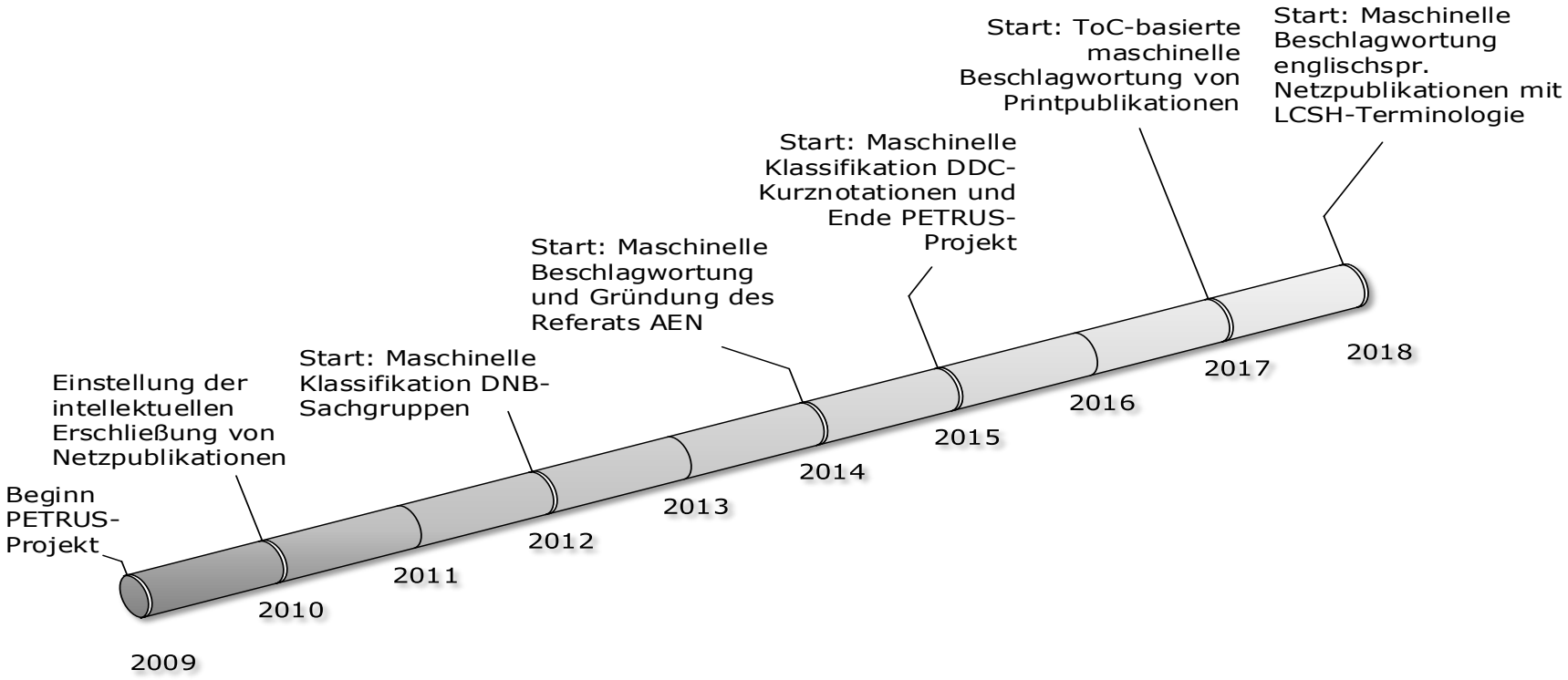
= Darstellungen in öffentlichen Netzen (Netzpublikationen)

E-Books, E-Paper,
E-Journals,
Websites
…

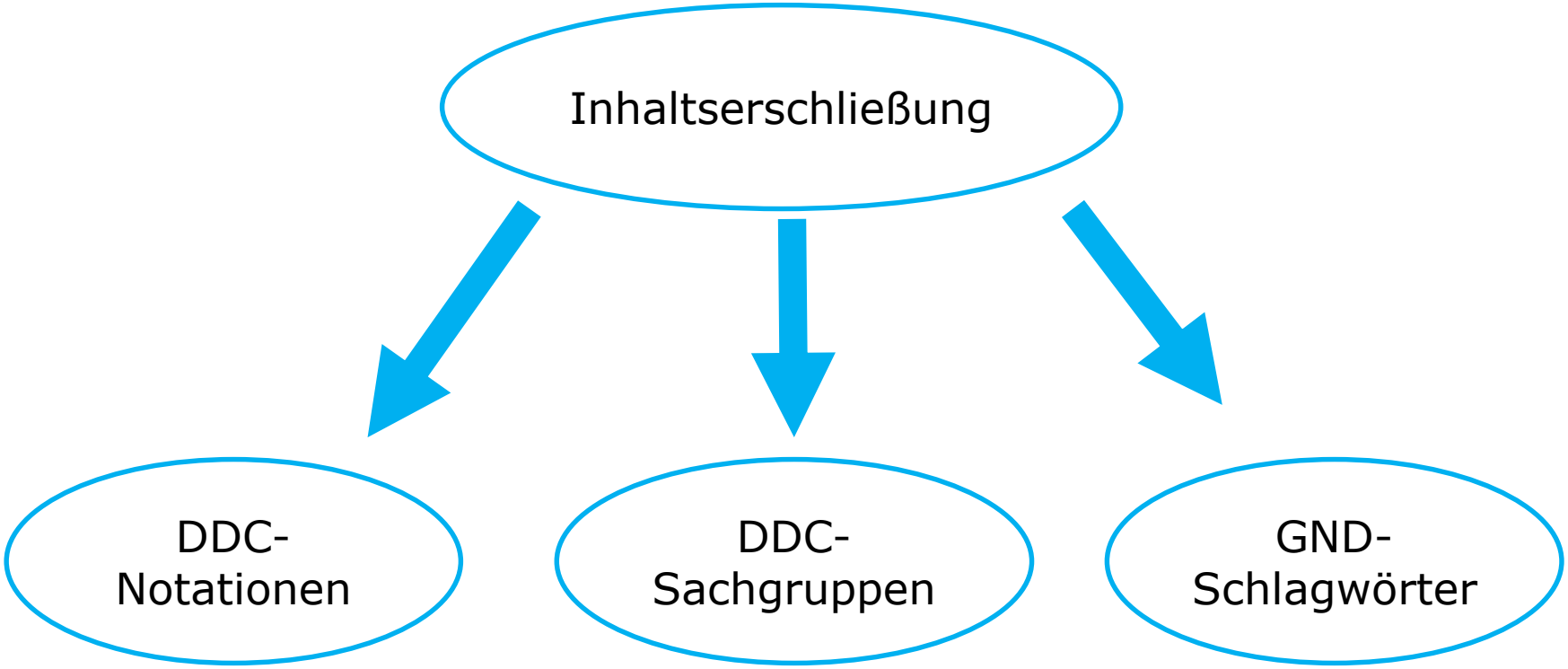
Zugang monographischer Print- und Netzpublikationen 2005–2017



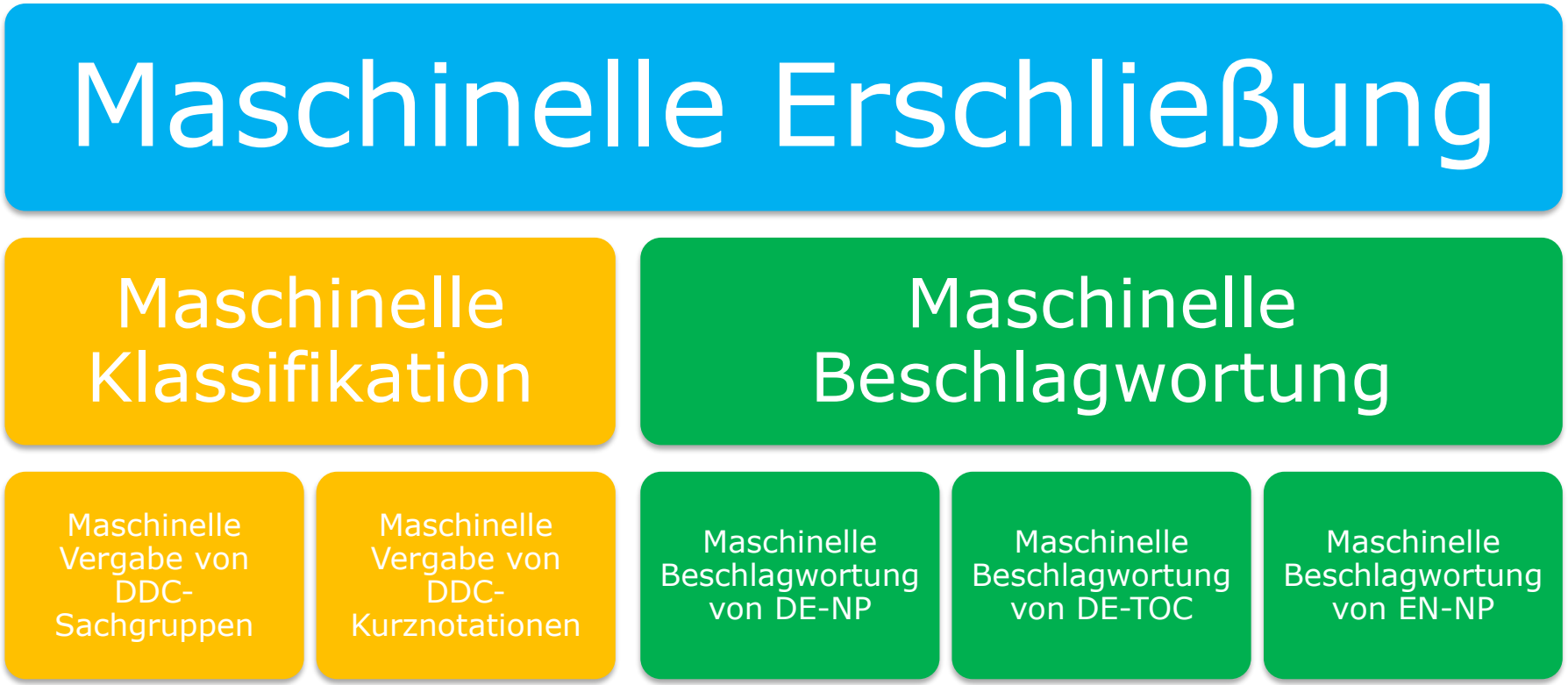
Maschinelle Erschließung in der Entwicklung



Intellektuelle Inhaltserschließung in der DNB



Maschinelle Erschließung: Anwendungsfälle



Maschinelle Klassifikation

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Klassifikation: Anwendungsfälle

Maschinelle Klassifikation

Maschinelle Vergabe von DDC-Sachgruppen

Maschinelle Vergabe von DDC-Kurznotationen

DDC-Sachgruppen

- Seit 2004
- Basiert auf der Dewey-Dezimalklassifikation (DDC)
- Aktuell 102 Klassen
- Sachgruppen Übersicht

DEUTSCHE NATIONAL BIBLIOTHEK

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

Beispiel DDC-Kurznotationen

Thema:
Studie
Übergewicht bei Kindern
Kiel
2000-2009

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

DEUTSCHE NATIONAL BIBLIOTHEK

Maschinelle Klassifikation

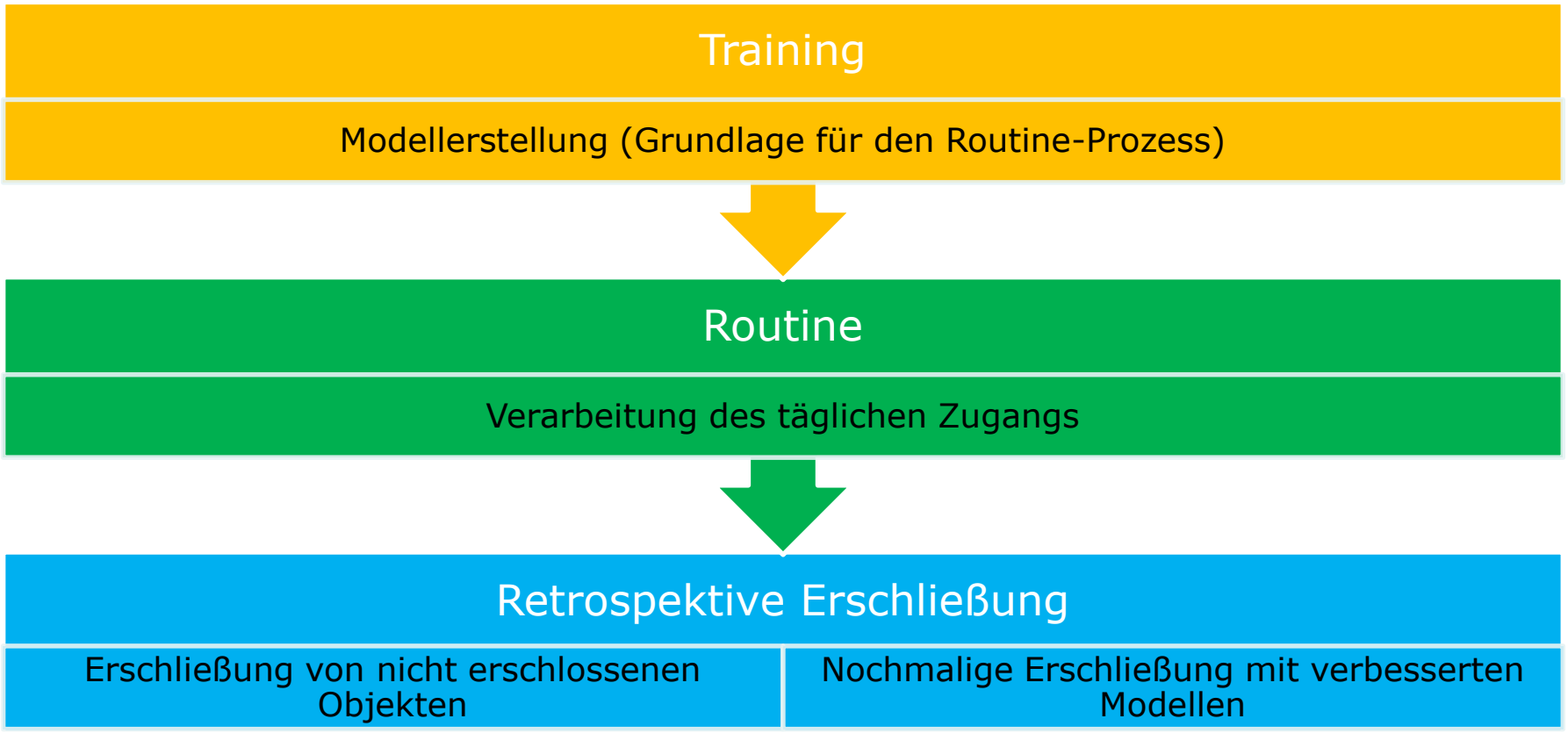
- Start: 2012 Sachgruppe / 2015 Kurznotationen
- Methode: Maschinelles Lernen / SVM
- Averbis Extraction Plattform (AEP)
- Dokumentarten:
  - Alle NPs ohne Belletristik
  - Formate PDF (2012) & Epub (2015)
  - Sprache Ger/Eng
- Umfang: 1.747.620 Publikationen (11/2018)

Maschinelles Lernen

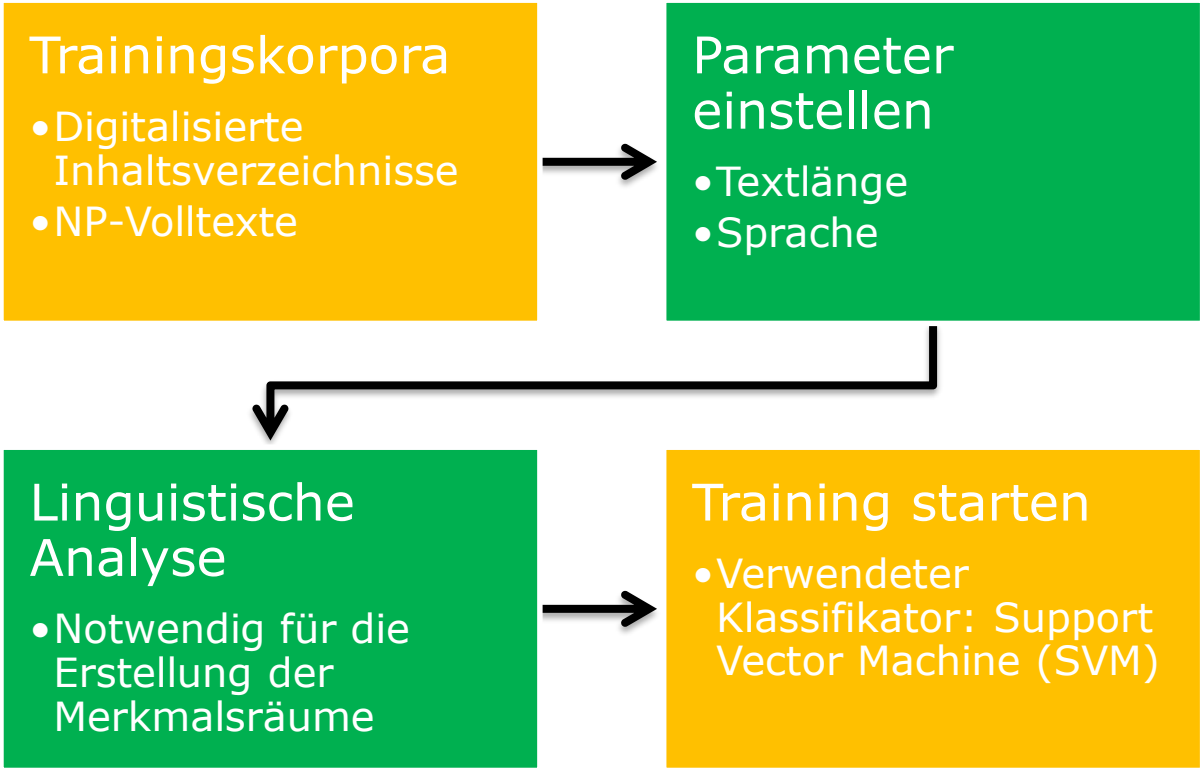
- Lernen aus Beispielen
- Erkennen von Mustern
- Verallgemeinerung der Muster
- Unbekannte Objekte können klassifiziert werden

Workflow

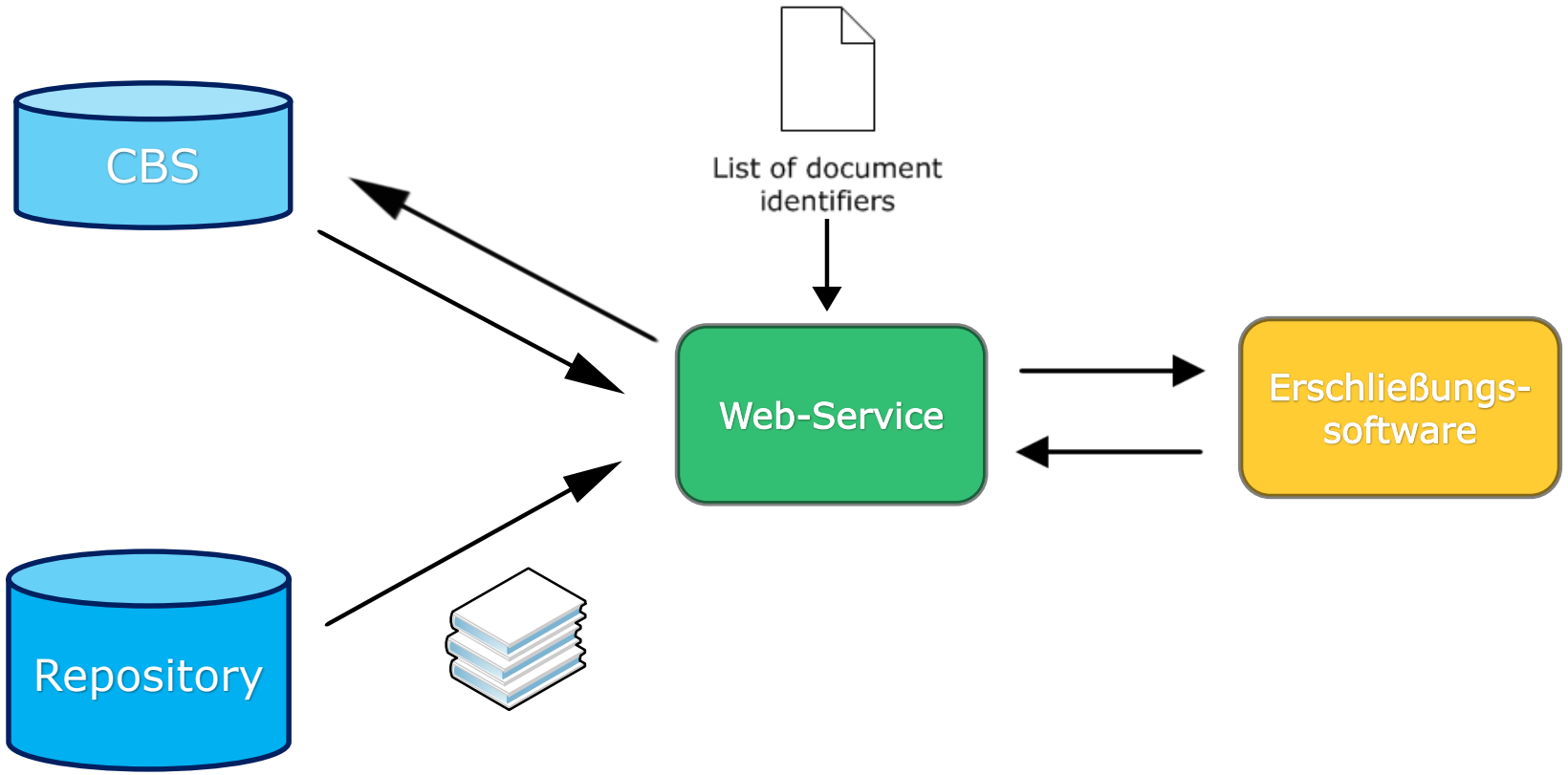
Workflow Übersicht



Training


Trainingskorpora
•Digitalisierte Inhaltsverzeichnisse
•NP-Volltexte
Parameter einstellen
•Textlänge
•Sprache
Linguistische Analyse
•Notwendig für die Erstellung der Merkmalsräume
Training starten
•Verwendeter Klassifikator: Support Vector Machine (SVM)

Routine


CBS
List of document identifiers
Web-Service
Erschließungs-software
Repository

Kennzeichnung

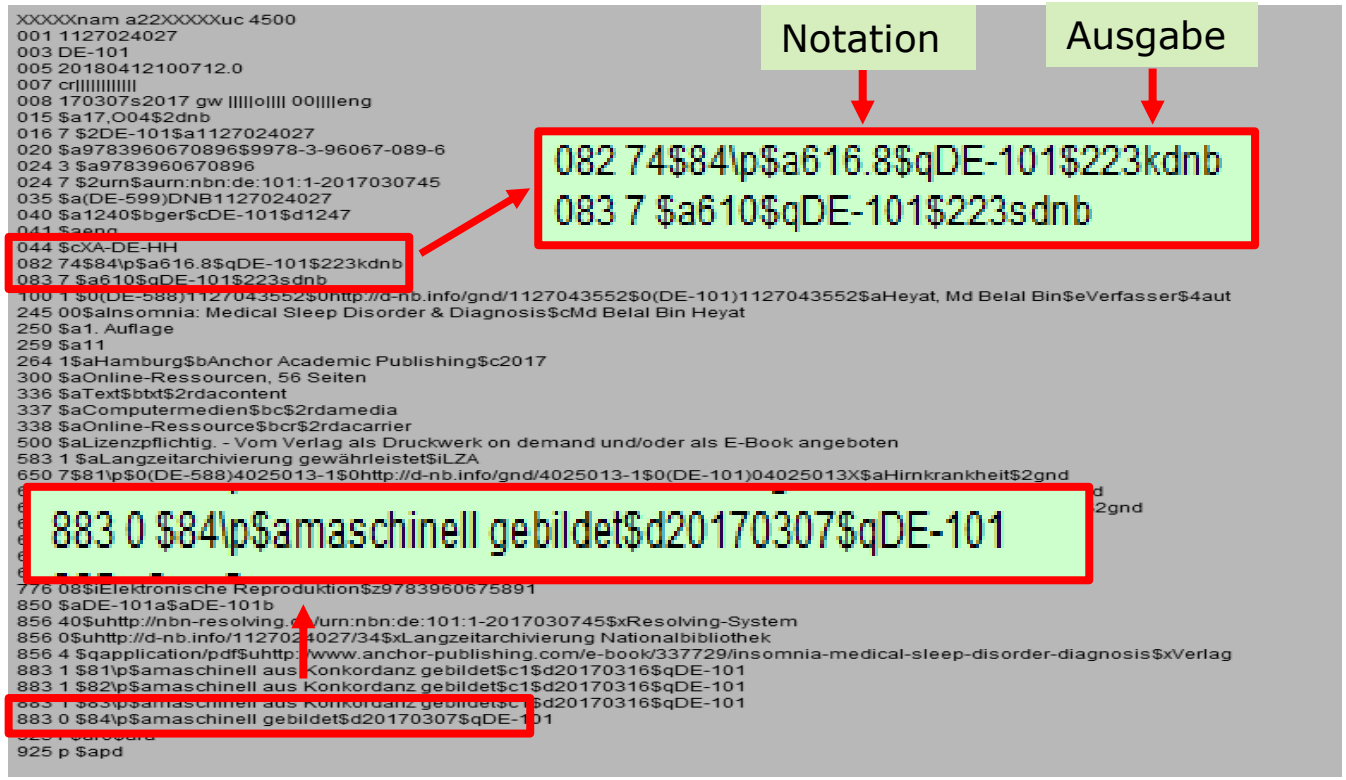
Kennzeichnung

- Kennzeichnung der maschinell vergebenen Notationen
- Datenauslieferung in MARC 21
- Kennzeichnung und Anzeige im DNB Portal

Kennzeichnung Marc 21

```
XXXXXnam a22XXXXXuc 4500
001 1127024027
003 DE-101
005 20180412100712.0
007 cr||||||||||||
008 170307s2017 gw |||||o|||| 00||||eng
015 $a17,O04$2dnb
016 7 $2DE-101$a1127024027
020 $a9783960670896$9978-3-96067-089-6
024 3 $a9783960670896
024 7 $2urn$aurn:nbn:de:101:1-2017030745
035 $a(DE-599)DNB1127024027
040 $a1240$bger$cDE-101$d1247
041 $aeng
044 $cXA-DE-HH
082 74$84\p$a616.8$qDE-101$223kdnb
083 7 $a610$qDE-101$223sdnb
100 1 $0(DE-588)1127043552$0http://d-nb.info/gnd/1127043552$0(DE-101)1127043552$aHeyat, Md Belal Bin$eVerfasser$4aut
245 00$aInsomnia: Medical Sleep Disorder & Diagnosis$cMd Belal Bin Heyat
250 $a1. Auflage
259 $a11
264 1$aHamburg$bAnchor Academic Publishing$c2017
300 $aOnline-Ressourcen, 56 Seiten
336 $aText$btxt$2rdacontent
337 $aComputermedien$bc$2rdamedia
338 $aOnline-Ressource$bcr$2rdacarrier
500 $aLizenzpflichtig. - Vom Verlag als Druckwerk on demand und/oder als E-Book angeboten
583 1 $aLangzeitarchivierung gewährleistet$iLZA
650 7$81\p$0(DE-588)4025013-1$0http://d-nb.info/gnd/4025013-1$0(DE-101)04025013X$aHirnkrankheit$2gnd
650 7$82\p$0(DE-588)4171595-0$0http://d-nb.info/gnd/4171595-0$0(DE-101)041715950$aNeuropsychiatrie$2gnd
650 7$83\p$0(DE-588)1068493003$0http://d-nb.info/gnd/1068493003$0(DE-101)1068493003$aNervenkrankheit$2gnd
653 $a(Produktform)Electronic book text
653 $a(BISAC Subject Heading)TEC007000
653 $aInsomnia;Power Spectral Density;Diagnosis;Sleep Disorder;Short Time Frequency;EEG Signal
653 $a(VLB-WN)1684
776 08$iElektronische Reproduktion$z9783960675891
850 $aDE-101a$aDE-101b
856 40$uhttp://nbn-resolving.de/urn:nbn:de:101:1-2017030745$xResolving-System
856 0$uhttp://d-nb.info/1127024027/34$xLangzeitarchivierung Nationalbibliothek
856 4 $qapplication/pdf$uhttp://www.anchor-publishing.com/e-book/337729/insomnia-medical-sleep-disorder-diagnosis$xVerlag
883 1 $81\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 1 $82\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 1 $83\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 0 $84\p$amaschinell gebildet$d20170307$qDE-101
925 r $aro$ara
925 p $apd
```

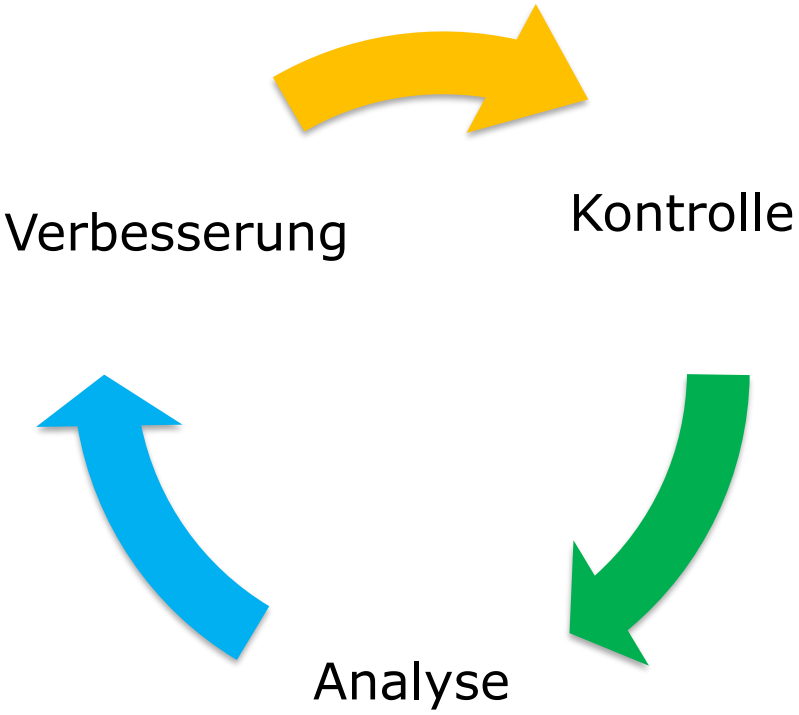
Kennzeichnung Marc 21



XXXXXnam a22XXXXXuc 4500
001 1127024027
003 DE-101
005 20180412100712.0
007 cr|||||||||||
008 170307s2017 gw |||||o|||| 00||||eng
015 $a17,O04$2dnb
016 7 $2DE-101$a1127024027
020 $a9783960670896$9978-3-96067-089-6
024 3 $a9783960670896
024 7 $2urn$aurn:nbn:de:101:1-2017030745
035 $a(DE-599)DNB1127024027
040 $a1240$bger$cDE-101$d1247
041 $aeng
044 $cXA-DE-HH
082 74$84\p$a616.8$qDE-101$223kdnb
083 7 $a610$qDE-101$223sdnb
100 1 $0(DE-588)1127043552$0http://d-nb.info/gnd/1127043552$0(DE-101)1127043552$aHeyat, Md Belal Bin$eVerfasser$4aut
245 00$aInsomnia: Medical Sleep Disorder & Diagnosis$cMd Belal Bin Heyat
250 $a1. Auflage
259 $a11
264 1$aHamburg$bAnchor Academic Publishing$c2017
300 $aOnline-Ressourcen, 56 Seiten
336 $aText$btxt$2rdacontent
337 $aComputermedien$bc$2rdamedia
338 $aOnline-Ressource$bcr$2rdacarrier
500 $aLizenzpflichtig. - Vom Verlag als Druckwerk on demand und/oder als E-Book angeboten
583 1 $aLangzeitarchivierung gewährleistet$iLZA
650 7$81\p$0(DE-588)4025013-1$0http://d-nb.info/gnd/4025013-1$0(DE-101)04025013X$aHirnkrankheit$2gnd
776 08$iElektronische Reproduktion$z9783960675891
850 $aDE-101a$aDE-101b
856 40$uhttp://nbn-resolving.de/urn:nbn:de:101:1-2017030745$xResolving-System
856 0$uhttp://d-nb.info/1127024027/34$xLangzeitarchivierung Nationalbibliothek
856 4 $qapplication/pdf$uhttp://www.anchor-publishing.com/e-book/337729/insomnia-medical-sleep-disorder-diagnosis$xVerlag
883 1 $81\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 1 $82\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 1 $83\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 0 $84\p$amaschinell gebildet$d20170307$qDE-101
925 p $apd

Qualitätsmanagement

Qualitätsmanagement



Ergebnisse Klassifikation

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

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. Informationswiss.	58% Übereinstimmung

Ausblick

Ausblick

- Verbesserung der Ergebnisse
- Anpassung und Neuorganisation der Geschäftsprozesse
- DCC-Kurznotationen für alle DDC-Sachgruppen

Vielen Dank für Ihre Aufmerksamkeit!

Fragen?

Frank Busse
Deutsche Nationalbibliothek
Automatische Erschließungsverfahren, Netzpublikationen
Telefon: 069-1525-1550
mailto:f.busse@dnb.de