Maschinelle Beschlagwortung in der DNB

Der nächste Schritt: Englischsprachige Netzpublikationen

Elisabeth Mödden

Inhaltsverzeichnis

Maschinelle Beschlagwortung

- maschinelle Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND)
- GND-Vokabular mit Qualitätslevel 1 und Teilbestand s
- Technologie:
  Averbis Extraction Platform
  Averbis Terminologie Platform
  der Firma Averbis GmbH
  Freiburg i.Br.

Tp – Person (individualisiert) / 369.015 Datensätze

Ts - Sachbegriff / 184.165 Datensätze

Ts1e – Hinweissatz / 4715 Datensätze

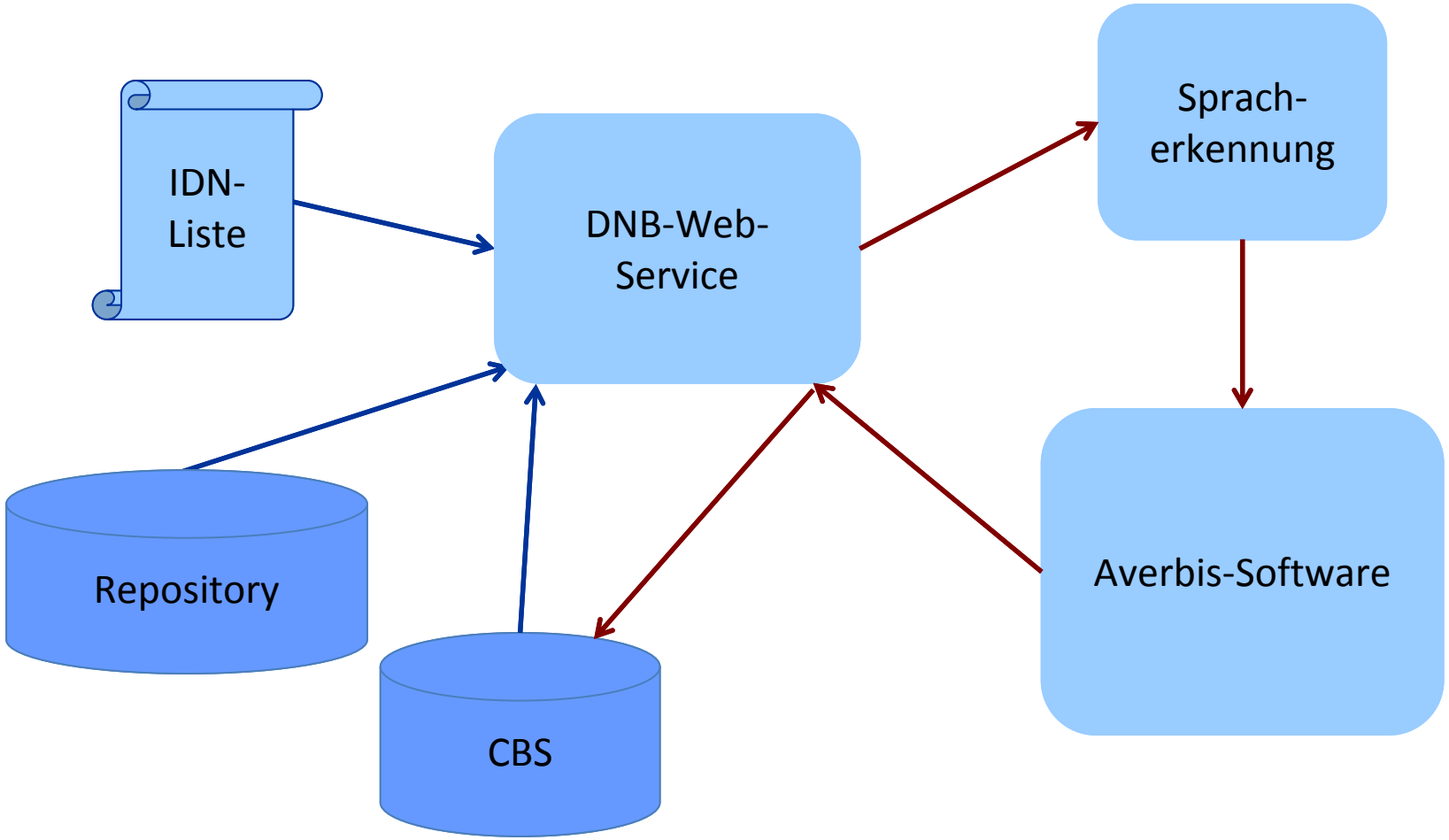
Tg - Geografikum / 205.244 Datensätze

Tb1 – Körperschaften / 142.593 Datensätze

Tf1 - Kongresse / 11.831 Datensätze

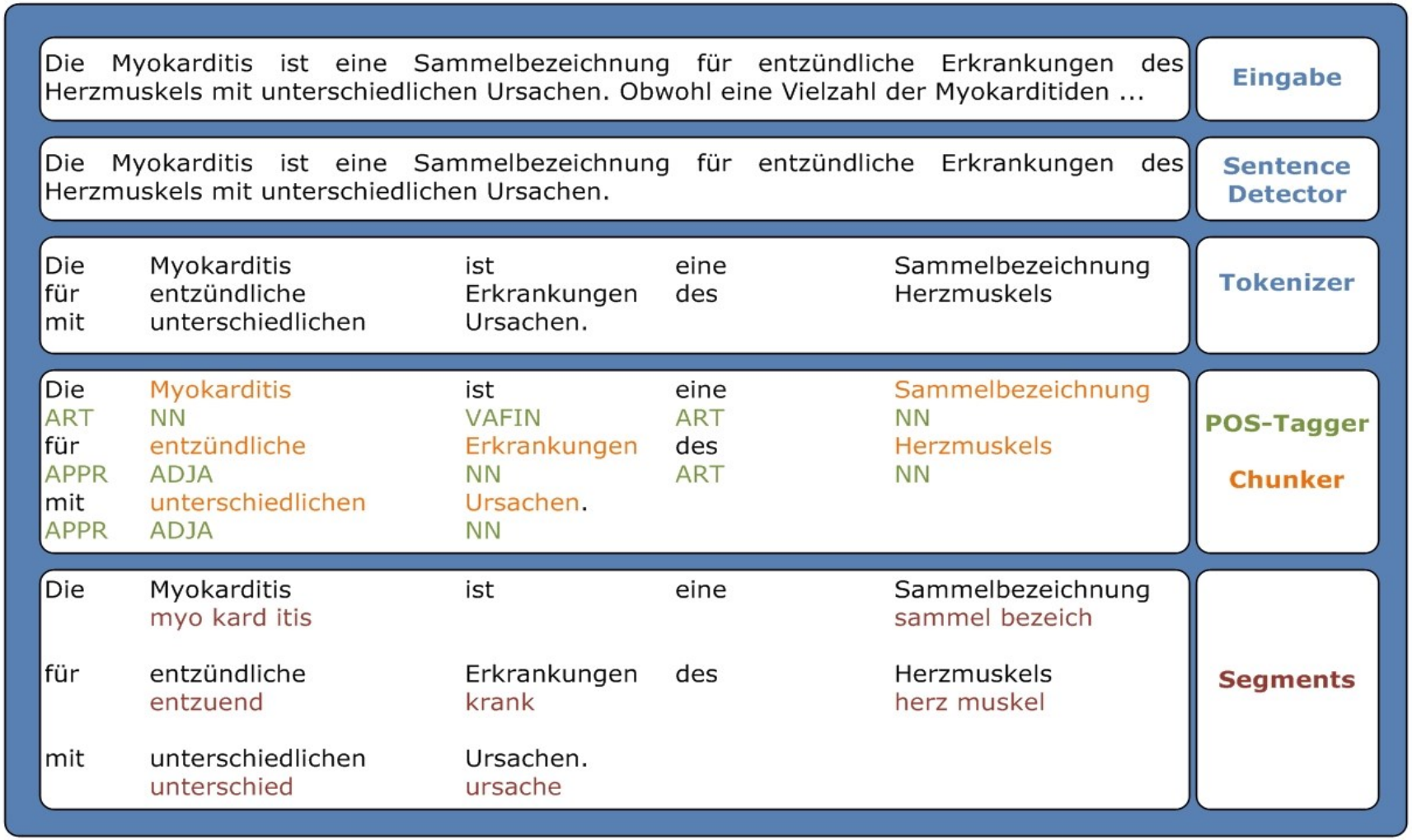
Tu1 – Werke / 87.831 Datensätze

Prozess
Beschlagwortung deutschsprachiger HSS



Averbis-Software

Linguistik

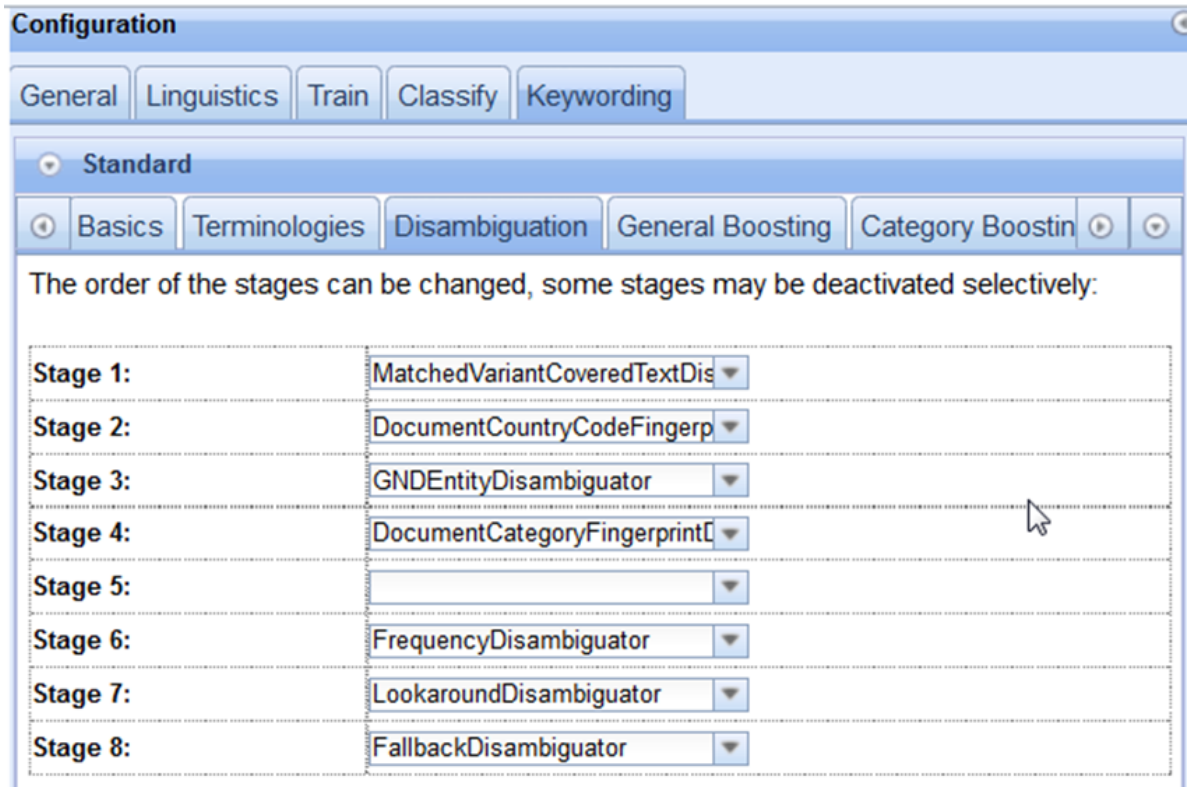

Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen. Obwohl eine Vielzahl der Myokarditiden ...
Eingabe
Die Myokarditis ist eine Sammelbezeichnung für entzündliche Erkrankungen des Herzmuskels mit unterschiedlichen Ursachen.
Sentence Detector
Die Myokarditis ist eine Sammelbezeichnung
für entzündliche Erkrankungen des Herzmuskels
mit unterschiedlichen Ursachen.
Tokenizer
Die Myokarditis ist eine Sammelbezeichnung
ART NN VAFIN ART NN
für entzündliche Erkrankungen des Herzmuskels
APPR ADJA NN ART NN
mit unterschiedlichen Ursachen.
APPR ADJA NN
POS-Tagger
Chunker
Die Myokarditis ist eine Sammelbezeichnung
myo kard itis sammel bezeich
für entzündliche Erkrankungen des Herzmuskels
entzuend krank herz muskel
mit unterschiedlichen Ursachen.
unterschied ursache
Segments

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

Ambiguität

Das Disambiguierungsverfahren der Software durchläuft mehrere Stufen:



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

======= next document: 968571956.txt
document text: "Prozeßorientierung im Wissensmanagement unter Verwendung grafischer Modelle Wissen im Fluß

Document fingerprint: {4.3=106, 1=82, 30=64, 9.3c=61, 10.11a=61, 6.5=45, 6.2a=44, 5.5=42, 10.11i=38, 28=32, 9.4a=32, 11.3a=30, 10.11b=30, 5.1a=27, 00=26, 4.4=23, 11.2a=22, 9.2b=21, 31.1a=21, 9.3b=21, 5.3=19, 6.7=18, 2.3=18, 14.2=14, 29=12, 10.2b=12, 11.1a=11, 2.2=11,
5.1b
20=
13.3
7.1a
31.3
10.9
19.1
12.1
12.4
14.4
31.7

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

Wörterbuchpflege

- Modus „default“ - Term und Textstelle werden segmentiert
- Modus „exact“ - Term muss exakt so im Text stehen
- Modus „ignore“ – Term wird „stillgelegt“, kein Matching

➡ Modifikationen möglich für einzelne Synonyme, alle Benennungen, ganze Hierarchiebäume

➡ Erstellung komplexer Filter möglich.

Beispiel

4000 Die @soziale Selektivität des lebenslangen Lernens : diskontinuierliche Erwerbsverläufe als exkludierender Faktor / Annika Schlenker.

4204 Berlin, Humboldt Universität zu Berlin, Diss., 2014

5050 370;330**$E**i
5050 370**$E**m**$K**0,996
5052 **$f**330**$F**0,828**$g**570**$G**0,800

5540 [GND]!041343735!Lebenslanges Lernen **$K**0,344
5540 [GND]!040059642!Berufslaufbahn **$K**0,020
5540 [GND]!041176227!Weiterbildung **$K**0,018
5540 [GND]!041342275!Berufsgruppe **$K**0,007

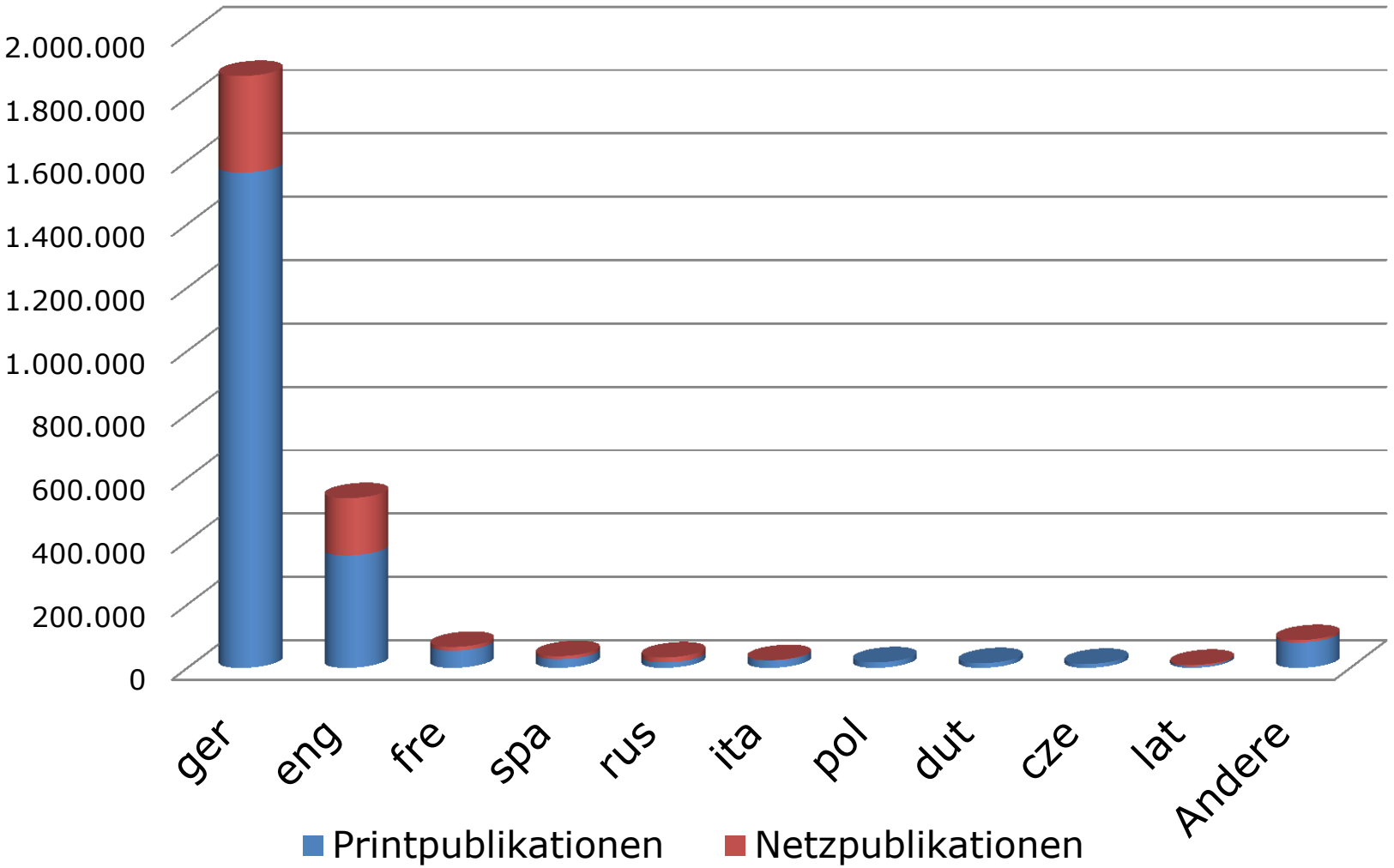
4000 Die @soziale Selektivität des lebenslangen Lernens : diskontinuierliche Erwerbsverläufe als exkludierender Faktor / Annika Schlenker.
4204 Berlin, Humboldt Universität zu Berlin, Diss., 2014

5050 370;330**$E**i
5050 370**$E**m**$K**0,996**$D**2015-03-25
5052 **$f**330**$F**0,828**$g**570**$G**0,800**$D**2015-03-25

5100 !041343735!Lebenslanges Lernen
5101 !040059642!Berufslaufbahn
5102 !04300539X!Ausgrenzung
5103 !040694674!Bildungssystem
5104 !042122856!Beschäftigungssystem

5540 [GND]!041343735!Lebenslanges Lernen **$K**0,344
5540 [GND]!040059642!Berufslaufbahn **$K**0,020
5540 [GND]!041176227!Weiterbildung **$K**0,018
5540 [GND]!041342275!Berufsgruppe **$K**0,007

Sprachverteilung
Publikationen mit Sprachencode



Zwei Lösungsansätze

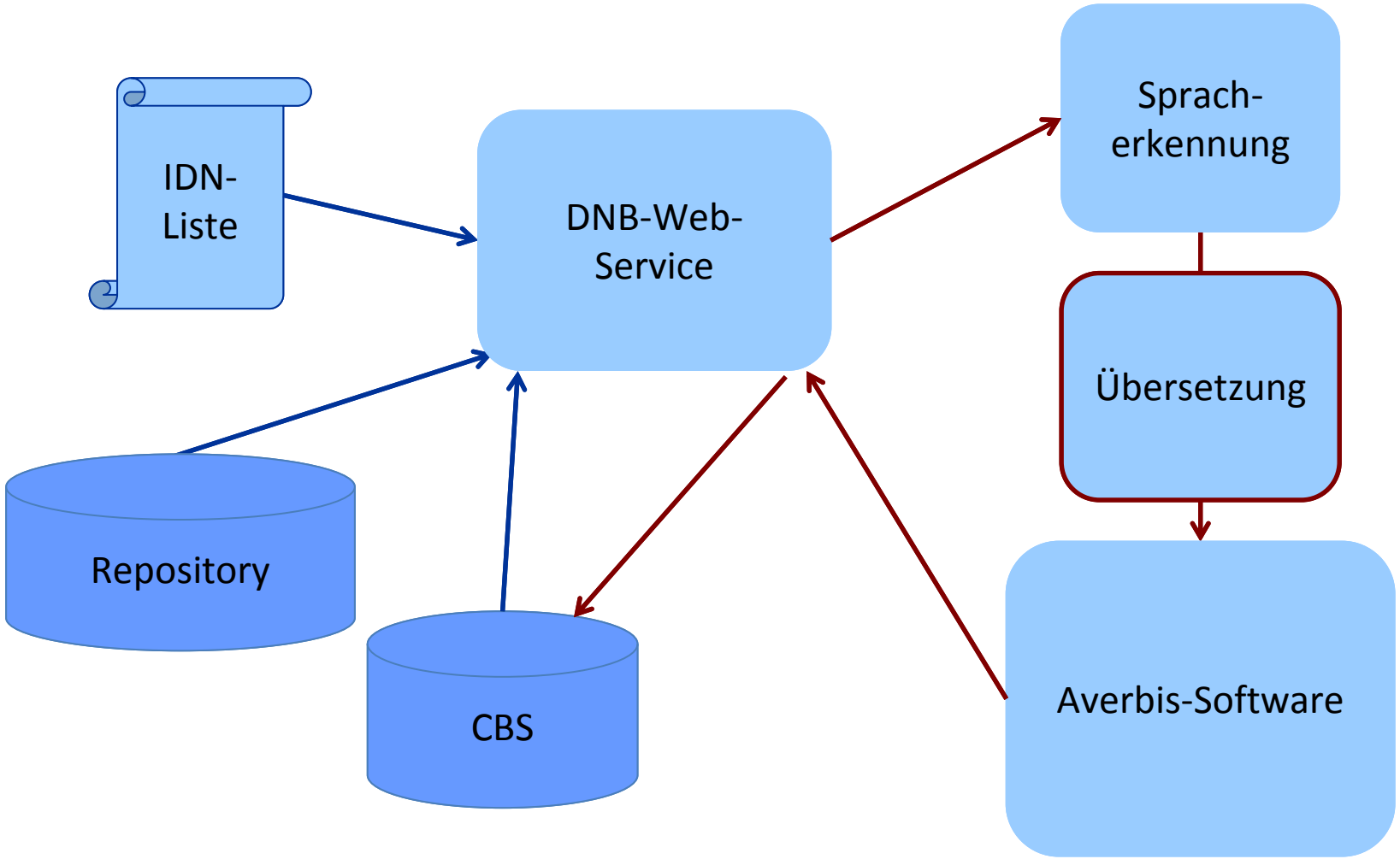
Lösungsvariante Einsatz eines Übersetzungstools:

Die englischsprachige Netzpublikation wird in die deutsche Sprache übersetzt und dann mit dem Vokabular der GND maschinell beschlagwortet.

Lösungsvariante Nutzung von Crosskonkordanzen:

Die englischsprachige Netzpublikation wird mit dem Vokabular der LCSH maschinell beschlagwortet und über Crosskonkordanzen, z.B. MACS-Verlinkungen, DBpedia und geeignete Thesauri, zum Vokabular der GND geführt.

Einsatz eines Übersetzungstools



Titel: Bank lending and monetary policy transmission in Germany

5100 !954405110!Europäische Zentralbank
5101 !040199029!Geldpolitik
5110 !040118827!Deutschland

5111 !040199029!Geldpolitik
5112 !041321596!Kreditgewährung
5113 !042885337!Vektor-autoregressives Modell
5114 !042330483!Fehlerkorrekturmodell

Translator 1:
|s|Kredit; 0.37311; |s|Geldpolitik; 0.35706; |s|Bank; 0.14161; |s|Kointegration; 0.10451; |s|Geld; 0.09899; |s|Zinsfuß; 0.09674; |s|Geldmarkt; 0.08666; |s|Bankkredit; 0.07489; |s|Fehlerkorrekturmodell; 0.05687; |s|Markt; 0.04832

Translator 2:
|s|Geldpolitik; 0.51309; |s|Kredit; 0.44286; |s|Bank; 0.34711; |s|Mindestreserve; 0.23942; |s|Zinsfuß; 0.22356; |s|Geldmarkt; 0.22247; |s|Darlehen; 0.21957; |s|Notenbank; 0.14618; |s|Tagesgeld; 0.14365; |s|Euro <Währung>; 0.11334

Einsatz eines Übersetzungstools

Offene Fragen:

- Welches Übersetzungstool eignet sich für welche Sprache am besten?
- Qualität bei der Disambiguierung?
- Auswirkungen von Übersetzungsfehlern?

Vorteile:

- Für alle Sprachen mit den entsprechenden Übersetzungstools einsetzbar.
- Technisch einfach im Prozess zu integrieren.

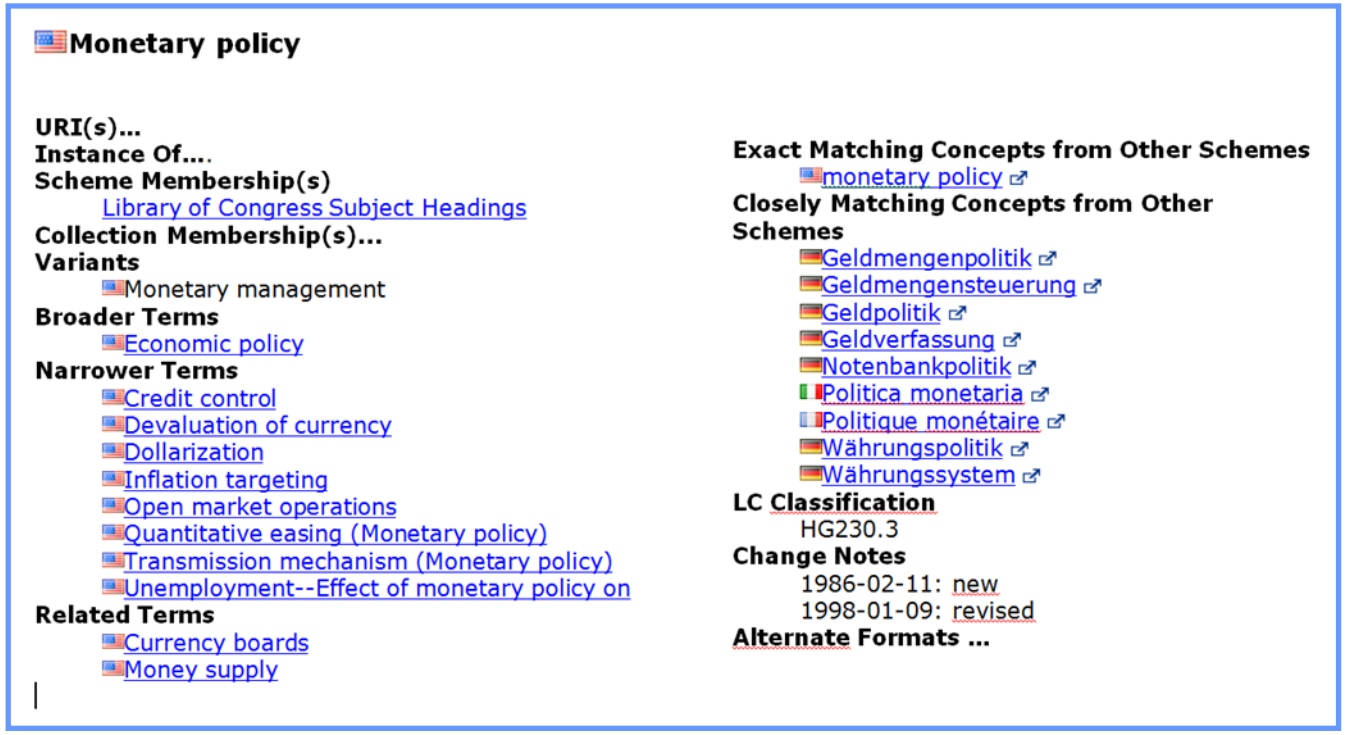
Nutzung von Crosskonkordanzen

Nutzung von Crosskonkordanzen

Vorgehen:

- Erweiterung der Terminologiebasis mit der LCSH
- Anpassungen Linguistik
- Anpassungen Disambiguierung

Library of Congress Subject Headings (LCSH)



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 ...

Nutzung von Crosskonkordanzen

DEUTSCHE NATIONAL BIBLIOTHEK

- Nutzung der MACS-Mapping-Links (aus dem Projekt: Multi Lingual Access to Subjects)
- Nutzung des Mappings zwischen GND und Standard Thesaurus Wirtschaft (STW)
- Nutzung weiterer Mappings
- Erstellen von Mappings mit DBpedia Englisch und weiteren geeigneten Thesauri

Nutzung von Crosskonkordanzen

Offene Fragen:

- LCSH / GND basieren auf unterschiedlichen Regelwerken – welche Folgen ergeben sich daraus?
- Technische Umsetzung des Mappings?
- Qualität der Mappings? Fehlerquote?
- Mapping nur für einen GND-Teilbestand möglich. Welche Lösung gibt es für den „Rest“?
- Der Einsatz von Crosskonkordanzen wird zu besseren Ergebnissen führen als der Einsatz eines Übersetzungstools. Führt aber eine Kombination beider Lösungsvarianten zu noch besseren Ergebnissen?

Ausblick

Realisierung in zwei Projektphasen:

Phase 1:
Anforderungserhebung für die Anpassung der Software

- Tests, Evaluierung und Vorbereitung möglicher Erschließungsszenarien
- Anforderungsbeschreibung
- Entwicklungsauftrag

Ausblick

Phase 2:

Entwicklung, Konsolidierung & Vorbereitung der Produktivnahme

- Entwicklungsphase
- Abnahme und Konsolidierung
- Vorbereitung Produktivnahme

Ziel:

Start der maschinellen Beschlagwortung englischsprachiger Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND) 2018

Vielen Dank

für Ihre Aufmerksamkeit

Elisabeth Moedden

Automatische Erschließungsverfahren, Netzpublikationen

Deutsche Nationalbibliothek

Telefon: +49-69-1525-1533

E-Mail: e.moedden@dnb.de