Sandro Uhlmann

Automatische Beschlagwortung mit dem Vokabular der Schlagwortnormdatei (SWD) und der Personennamendatei (PND)

Erfahrungen aus dem DNB-Projekt PETRUS

Automatische Beschlagwortung mit dem Vokabular der SWD und PND

- Projekt PETRUS
- Automatische Beschlagwortung:

  Projektziele und bisheriger Projektverlauf
- Software: *Averbis Extraction Platform*
- Indexierungsqualität und Evaluierung
- Test und Ergebnisse
- Fazit und Ausblick

Das Projekt PETRUS

„Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek“ (2009 – 2011) - Auswahl, Erprobung und Einführung softwaregestützter Verfahren für die formale und inhaltliche Erschließung von Netzpublikationen (NPs)

- Automatischer Abgleich aller beim NP-Import mitgelieferten Verfassernamen mit der Personennamendatei (PND)

- Automatische Übernahme von Erschließungsdaten und Normdatenverknüpfungen aus parallelen Ausgaben

- Automatische Einordnung der NPs in die Systematik der DDC-Sachgruppen

- Automatische Beschlagwortung der NPs mit dem kontrollierten Vokabular von SWD und PND

Automatische Beschlagwortung: Projektziele und Projektverlauf

DEUTSCHE NATIONAL BIBLIOTHEK

- automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der SWD und der PND sowie die Generierung freier Schlagwörter für deutsch- und englischsprachige NPs
- 2010 Durchführung von zeitlich befristeten Funktionstests mit zwei ausgewählten Softwaresystemen, die zeigen sollten, a) wie gut die automatisch vergebenen Schlagwörter das Thema einer Publikation tatsächlich abbilden und ob sie sinnvolle Sucheinstiege liefern sowie b) welcher technisch/methodische Ansatz für die Daten der DNB am besten geeignet ist
- im Herbst 2010 fanden die Erkenntnisse der Testphase Eingang in ein erneutes Beschaffungsverfahren infolge dessen das Softwaresystem der Firma Averbis den Zuschlag erhielt

Software: *Averbis Extraction Platform*

Averbis GmbH, Freiburg i.Br. (www.averbis.de):

– Entwicklung & Vertrieb linguistisch-semantischer Suchtechnologien und Software zur automatischen Analyse von Texten

Automatische Beschlagwortung mit dem Averbis Concept Mapper:

– ein konfigurierbarer, lexikonbasierter Annotator, der auf unterschiedlichen Ebenen einer linguistischen Vorverarbeitung aufsetzt, insbesondere morphologische Analyse (Wortebene) und Syntaxanalyse (Satzebene), kombiniert mit Methoden des maschinellen Lernens

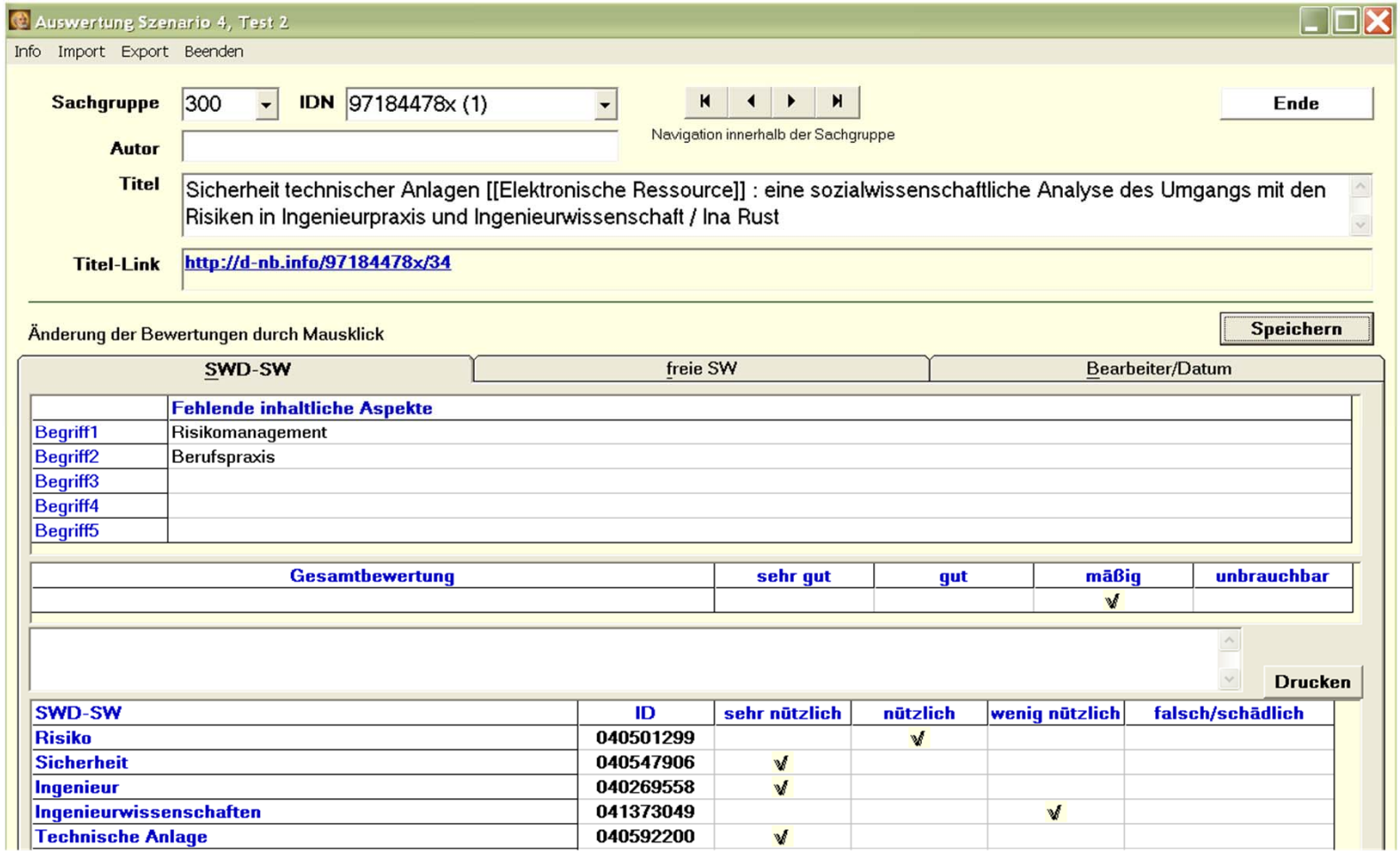
– eine flexible Lexikonstruktur ermöglicht die Aufnahme von Synonymen und verschiedenen Attributen

Qualität und Evaluierung

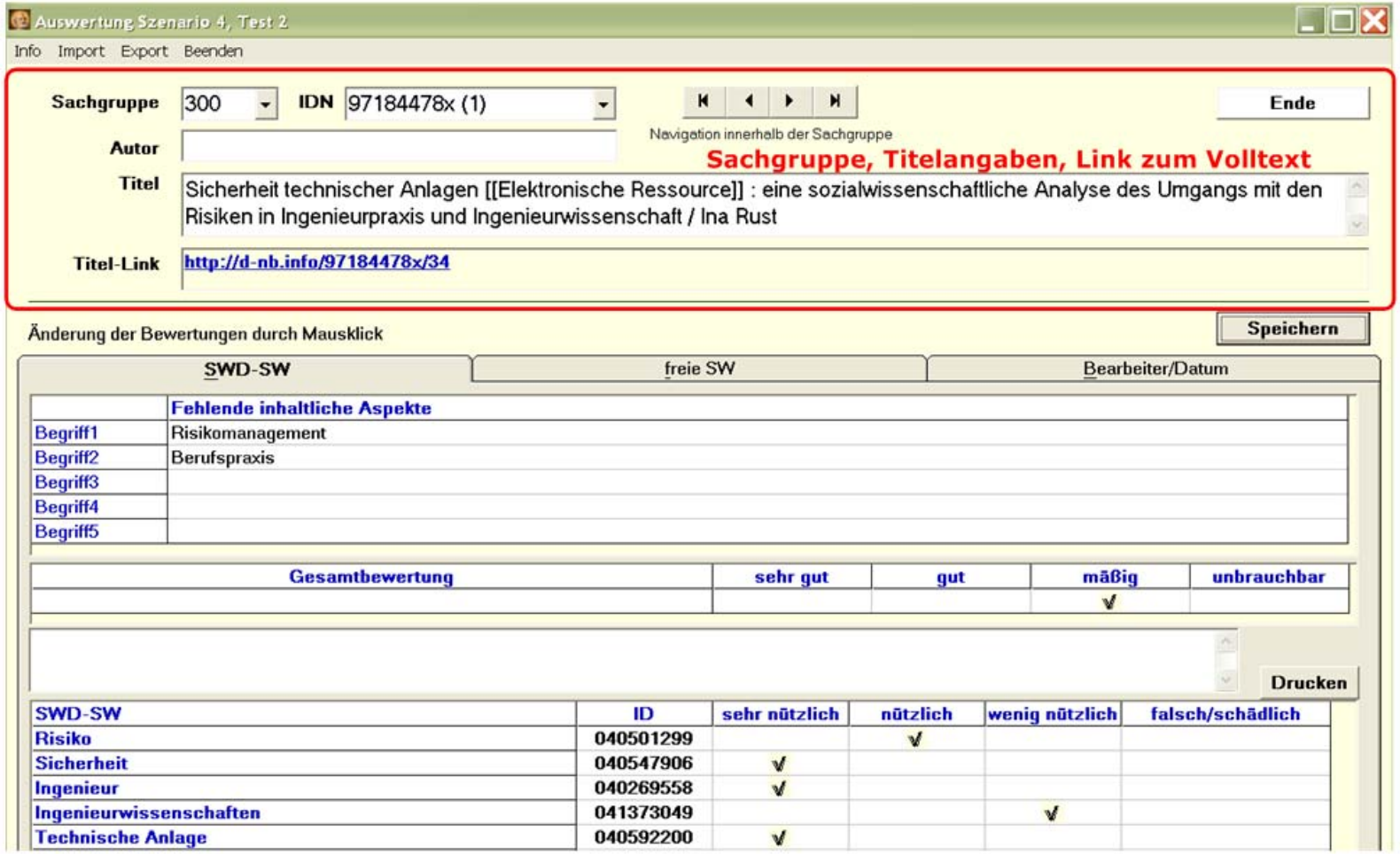
DEUTSCHE NATIONAL BIBLIOTHEK

- intellektuelle Beurteilung der inhaltlichen Übereinstimmung zwischen den automatisch vergebenen Schlagwörtern und dem Thema des Dokuments anhand von Stichproben
- in einer Datenbank bekommen die FachreferentInnen der Abteilung Inhaltserschließung den Autor, Titel, einen Link zum Volltext sowie eine Liste der Schlagwörter vorgelegt
- es wird jedem Schlagwort auf einer 4-Punkte-Skala ein Wert in den Kategorien *sehr nützlich, nützlich, wenig nützlich, falsch/schädlich* zugewiesen
- statistische Auswertung der intellektuellen Beurteilungen: Berechnung von Nützlichkeit (Precision) und Vollständigkeit (Recall) sowie der Average generalized Precision (F-Measure)

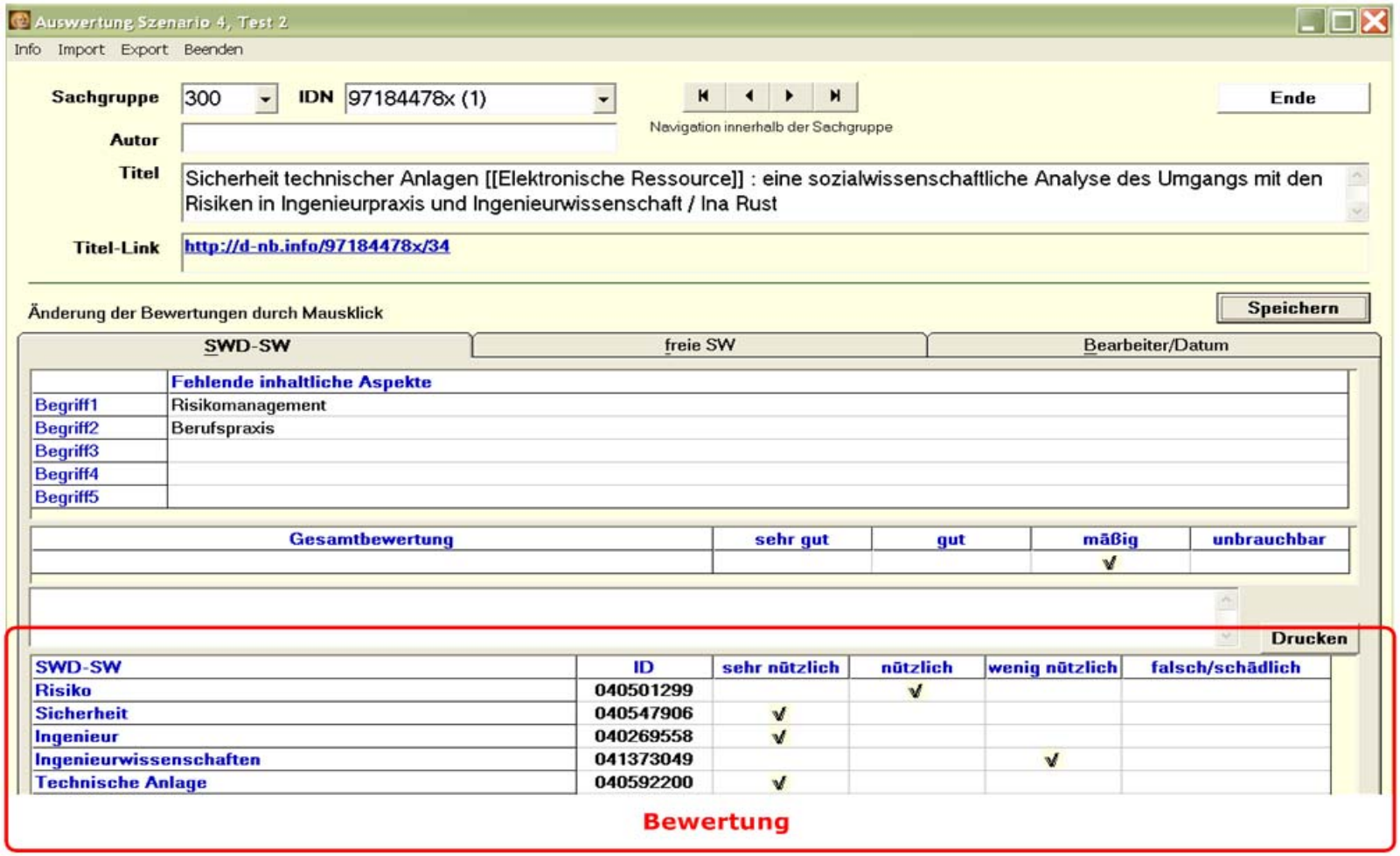
Qualität und Evaluierung



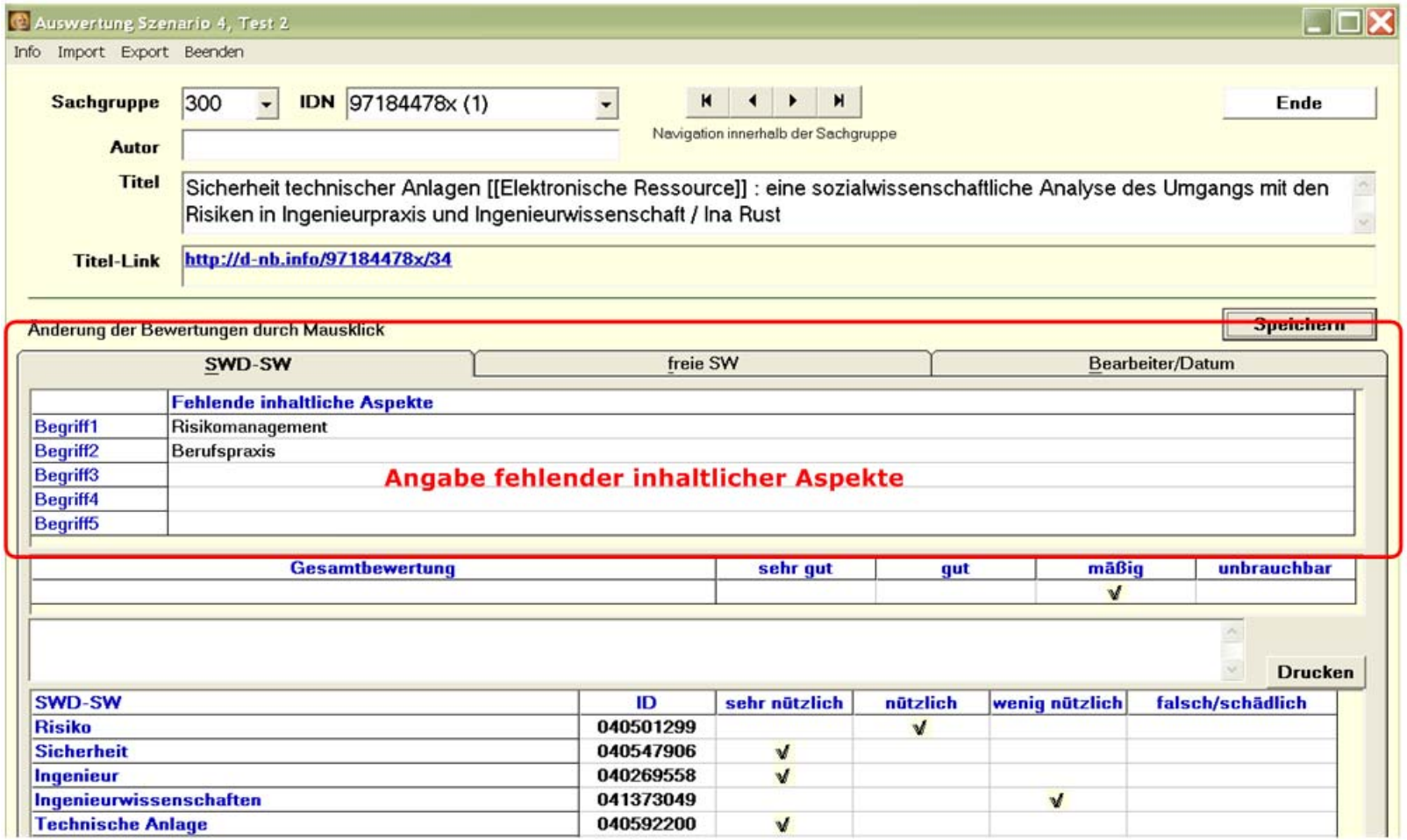
Qualität und Evaluierung



Qualität und Evaluierung



Qualität und Evaluierung



Test und Ergebnisse: Release 2

Testobjekte: deutschsprachige elektronische Volltexte (v.a. Hochschulschriften)

Testbedingungen: im Averbis-Dictionary enthalten sind 170.000 Sachschlagwörter und 153.000 Geografika/Ethnografika der SWD sowie 311.000 individualisierte Personennamen der PND ; 10 Sachgruppen ; Ausgabe von 10 Schlagwörtern pro Dokument

Sachgruppen	Anzahl beschlagworteter Objekte	Anzahl Stichproben
004 Informatik	273	25
100 Philosophie	55	25
320 Politik	55	25
330 Wirtschaft	284	25
340 Recht	221	25
370 Erziehung, Schul- und Bildungswesen	213	25
510 Mathematik	177	25
610 Medizin	238	25
830 Deutsche Literatur	58	25
900 Geschichte	40	25
Gesamt: 10	*1614*	*250*

Test und Ergebnisse: Release 2

Beispiel 1 http://d-nb.info/969596510

Titel: Theorie und Praxis der Nutzung militärischer Macht im Vergleich der Systeme USA und Sowjetunion [Elektronische Ressource] : eine Anwendung der Methodologie von Clausewitz auf den Einsatz militärischer Macht in der Epoche des Ost-West-Konflikts 1945 - 1991 / Frank Kostelnik

Sachgruppe(n): 320 Politik

SW	IDN	KW	Bewertung
|s|Militärische Macht	041699505	1.0	sehr nützlich
|s|Militärischer Einsatz	041145992	0.381	nützlich
|s|Ost-West-Konflikt	040757706	0.373	sehr nützlich
|g|Sowjetunion	040775488	0.252	sehr nützlich
|s|Macht	040368246	0.194	nützlich
|s|Klausenit	958190623	0.173	falsch
|g|USA	040787044	0.168	sehr nützlich
|s|Witz	040666808	0.143	falsch
|s|Systemvergleich	040588130	0.028	sehr nützlich
|s|Internationales politisches System	041254880	0.022	nützlich

Gesamtbewertung: *Mäßig*

Fehlende Aspekte: |p|Clausewitz, Carl von, |s|Außenpolitik

Test und Ergebnisse: Release 2

Beispiel 2 http://d-nb.info/96963756X

Titel: Das Strafgerichtswesen im kurpfälzischen Territorialstaat [Elektronische Ressource] : Entwicklungen der Strafgerichtsbarkeit in der Kurpfalz ; dargestellt anhand von ländlichen Rechtsquellen aus vier rechtsrheinischen Zenten / von Melanie Julia Hägermann

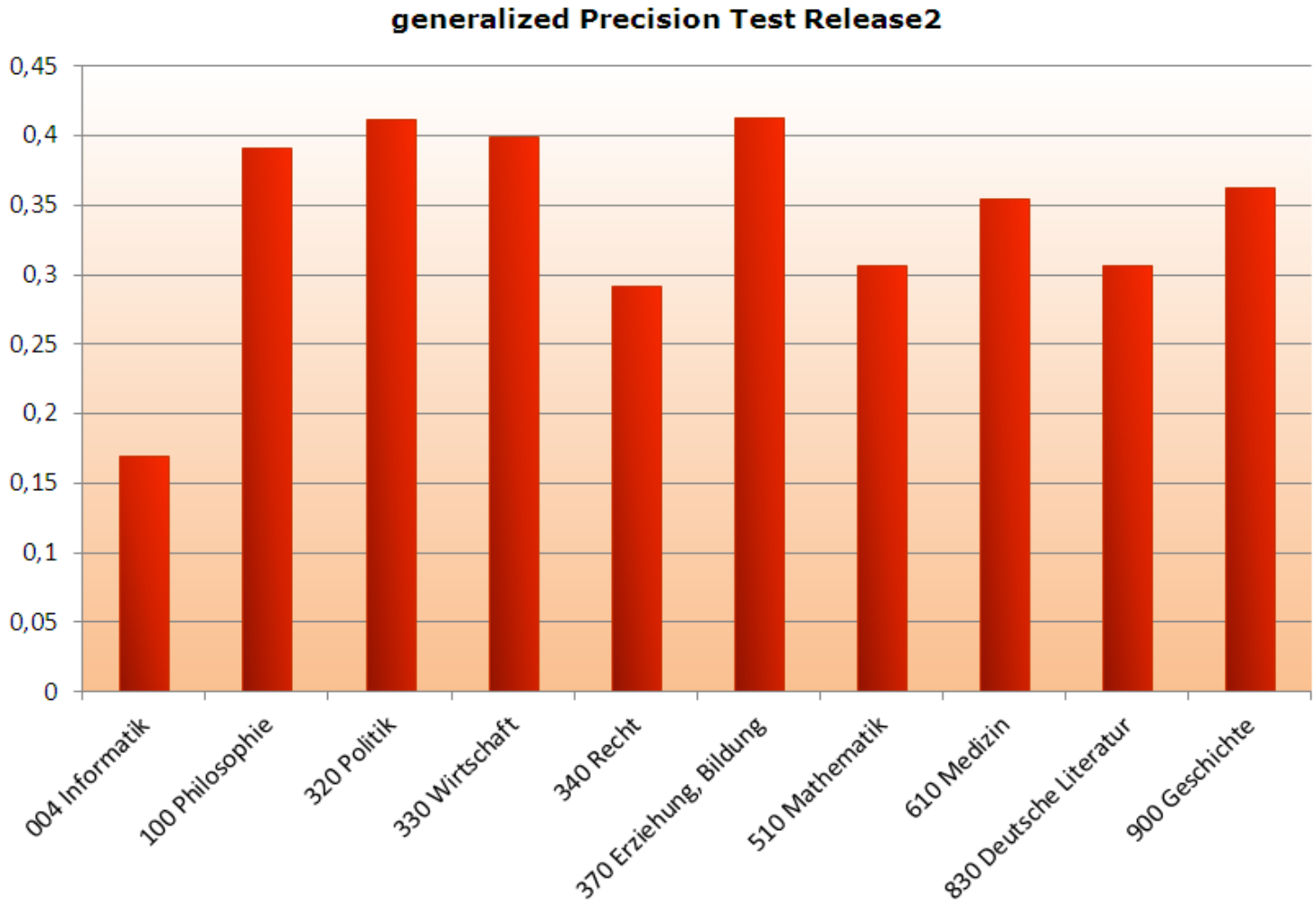
Sachgruppe(n): 340 Recht

SW	IDN	KW	Bewertung
|g|Pfalz	040760316	1.0	sehr nützlich
|s|Zentenar	041907019	0.360	nützlich
|s|Strafgerichtsbarkeit	040577899	0.219	sehr nützlich
|s|Wesen	041897242	0.116	falsch
|s|Zent	042322723	0.106	sehr nützlich
|s|Weistum	04065267X	0.059	wenig nützlich
|s|Entwicklung	041134508	0.057	falsch
|g|Alzey <Oberamt>	043020852	0.044	falsch
|g|Schriesheim	040532968	0.042	wenig nützlich
|s|Gerichtsbarkeit	040203425	0.032	nützlich

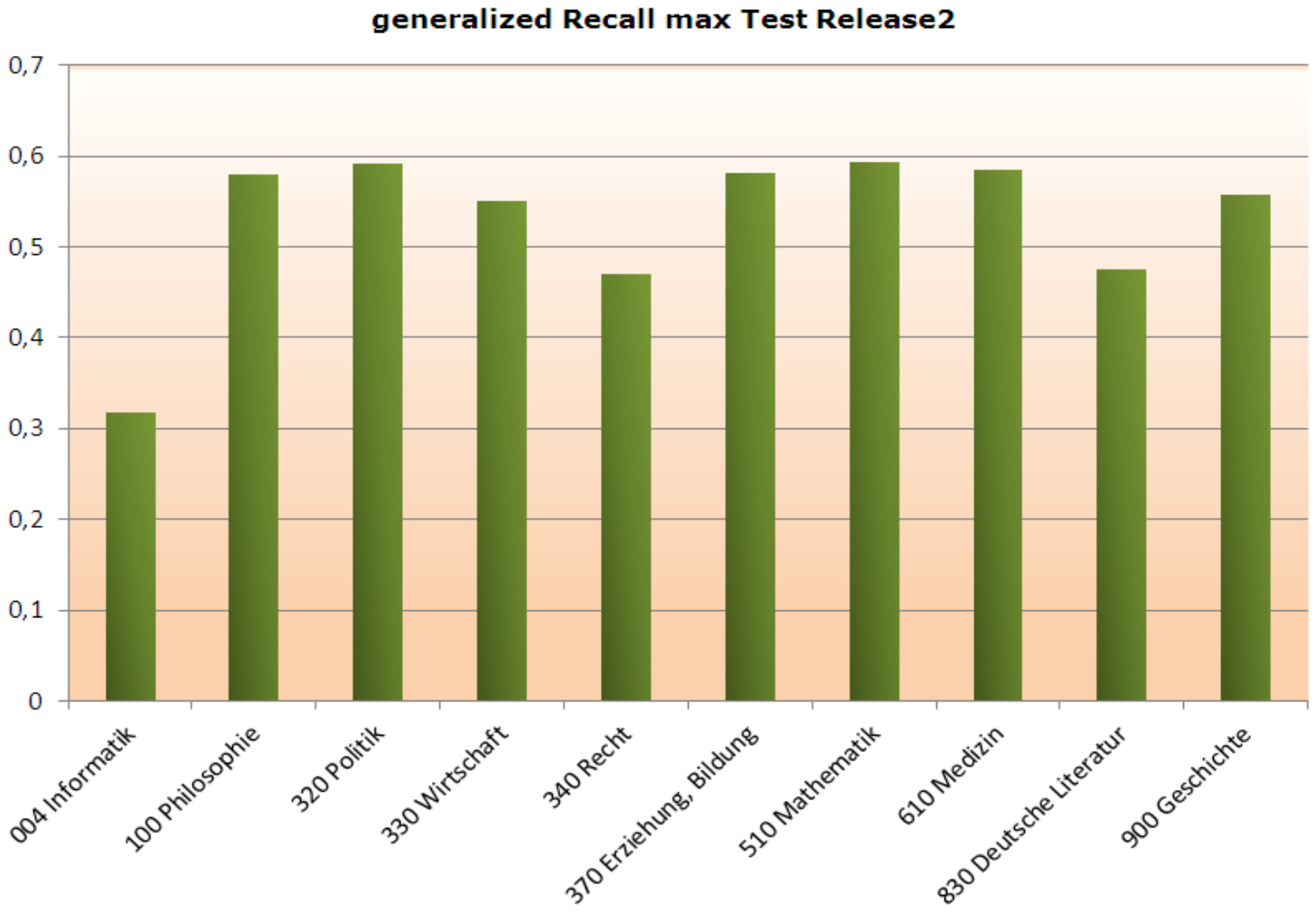
Gesamtbewertung: *Mäßig*

Fehlende Aspekte: Geschichte

Test und Ergebnisse: Release 2



Test und Ergebnisse: Release 2



Mehrdeutigkeit im Vokabular

800 **|s|Bank <Möbel>** 808 |a|M 1. 810 13.6 816 645.4#2# 816 684.1#2# 816 749.3#2# 830 |s|Sitzbank 850 |s|Sitzmöbel	800 **|s|Möbel** 808 |a|M, Au=DB 810 13.6;31.13 816 645.4#3# 816 684.1#3# 816 749#4# 830 |s|Mobiliar 830 |s|Möbelkunst 830 |s|Wohnmöbel 830 |s|Holzmöbel *Quasisynonym 850 |s|Wohnungseinrichtung	
800 **|s|Bank** 808 |a|M unter Banken 810 10.9b 816 332.1#3# 816 339.53#2# 830 |s|Bankbetrieb 830 |s|Banken 860 |s|Geschäftsbank 860 |s|Kreditinstitut 860 |s|Kreditwesen	800 |s|Turnbank 808 |a|Sport-B. 810 34.3 815 saz 816 372.860284#2# 816 796.440284#2# 818 796.40284#2# 818 796.44#1# 830 **|s|Bank <Geräteturnen>** 830 |s|Langbank 850 |s|Turngerät	800 **|s|Geräteturnen** 808 |a|B 1986, Du. 810 34.3 815 saz 816 796.442#4# 818 796.44#3# 830 |s|Boden- und Geräteturnen 830 |s|Gerätturnen *Sport-B 830 |s|Kunstturnen 850 |s|Turnen 860 |s|Turngerät 860 |s|Geräteturner
800 **|g|Bank <Herzogenrath>** 808 |a|Orts-Mü. 29 811 XA-DE-NW 815 gik 830 |g|Herzogenrath-Bank	800 **|g|Herzogenrath** 808 |a|Orts-Mü. 27, GKD 811 XA-DE-NW 815 gik	

Fazit und Ausblick

DEUTSCHE NATIONAL BIBLIOTHEK

- Fehler bei der Schlagwortvergabe liegen insbesondere in der unzureichenden Auflösung von Mehrdeutigkeiten ; Lösungsansatz: Entwicklung eines mehrstufigen Disambiguierungsverfahrens, u.a. mittels Nutzung der Homonymenzusätze, Oberbegriffe, verwandte Begriffe, Ländercodes und SWD-Sachgruppen aus dem SWD/PND-Vokabular und der Anwendung von Methoden der maschinellen Textanalyse
- Universalität des SWD-Wortschatzes: problematisch sind bspw. die vielen vorhandenen aber inhaltlich oft wenig relevanten Allgemeinbegriffe ; Lösungsansatz: eine thematische Vorauswahl von Sachschlagwörtern, die in Korrelation zur (automatisch) vergebenen Sachgruppe stehen, mittels Nutzung der am Schlagwortdatensatz vorhandenen Notationen der SWD-Sachgruppen und der Dewey-Dezimalklassifikation

Fazit und Ausblick

DEUTSCHE NATIONAL BIBLIOTHEK

- Probleme bereitet auch die generelle Unterscheidung zwischen Personen, Geografika und Sachbegriffen im Text ; Lösungsansatz: mittels Eigennamenerkennung (Named Entity Recognition) sollen Geografika und Personennamen als solche erkannt werden und die entsprechenden geografischen Schlagwörter bzw. Personenschlagwörter nur hier zur Anwendung kommen
- stärkere Berücksichtigung der formalen Struktur einzelner Objektgruppen und Anpassung der sprachlichen Verarbeitung
- aktuell: Averbis Software Release 3
- bis Jahresende: Weiterentwicklung und Anpassung der Software für die Übernahme in den DNB-Betrieb

DEUTSCHE NATIONAL BIBLIOTHEK

Vielen Dank für Ihre Aufmerksamkeit.

Fragen, Anregungen, Kritik an
Sandro Uhlmann
Deutsche Nationalbibliothek
Abt. Inhaltserschließung
Tel.: +49-341-2271-385
Email: s.uhlmann@dnb.de