Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

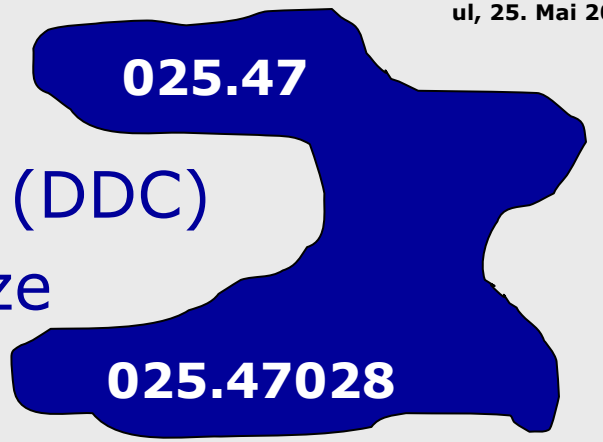
Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen

Ulrike Reiner
Verbundzentrale des Gemeinsamen Bibliotheksverbundes (VZG)

VZG

Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen - Inhalt des Vortrages -

Fundy-Nationalpark
ul, 25. Mai 2008



- **Inhalte erschließen –**
  - Dewey Dezimalklassifikation (DDC)
  - Bibliografische Titeldatensätze

  mit neuen Tools
  - OCLC **classify** (an experimental classification web service)
  - VZG Colibri/DDC ***vc_dcl*** (*vzg colibri_ddc classifier)*

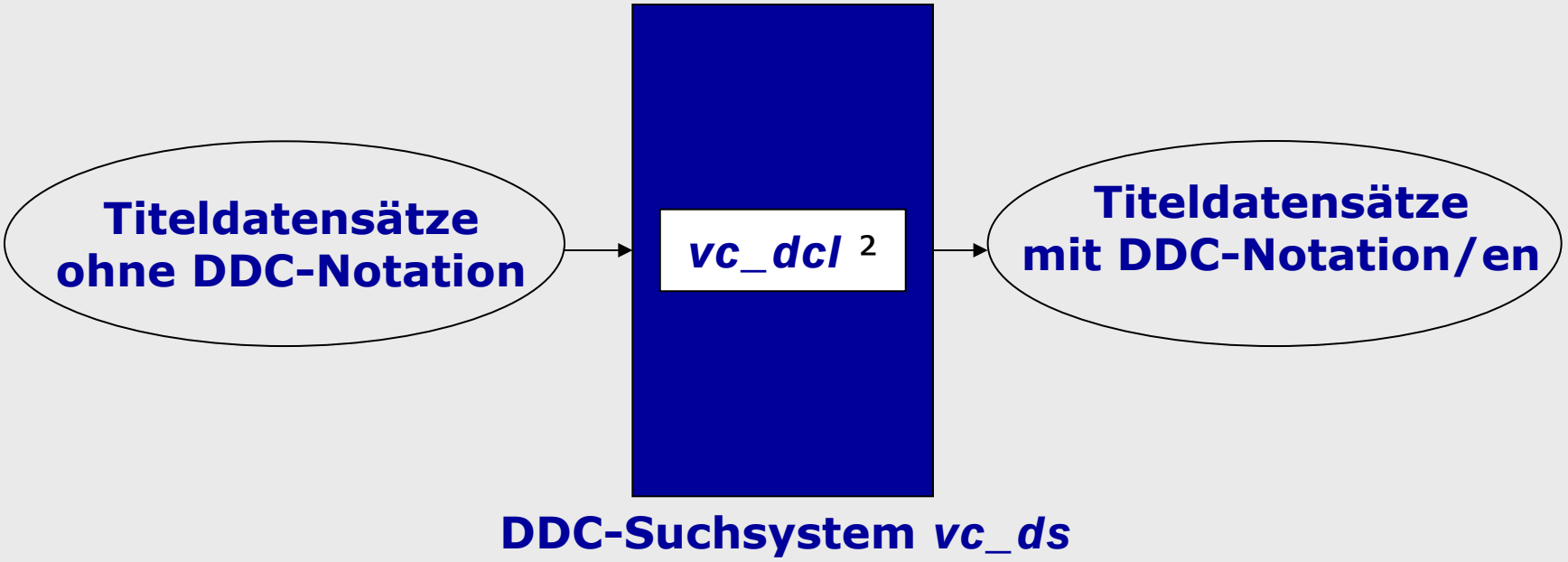
- **VZG-Colibri/DDC-Wettbewerb**

- **Klassifizierungskomponente *vc_dcl*: Modell, Tests, Bewertung, Ergebnisse & Perspektiven**

Colibri/DDC - Forschungsfrage *Q1*

Fundy-Nationalpark
ul, 25. Mai 2008

Ist es möglich, eine inhaltlich stimmige DDC-Titelklassifikation aller GVK-PLUS[1]-Titeldatensätze automatisch zu erzielen?



[1]GVK-PLUS: Gemeinsamer Verbundkatalog (GVK) und Online Contents (OLC); [2]*vzg colibri_ddc classifier*

Dewey-Dezimalklassifikation (DDC)
DDC-Ausschnitt

Fundy-Nationalpark
ul, 25. Mai 2008

Ebene

1. **0**00(Informatik, ...) **1**00 **2**00 **...** **9**00
2. **00**0 (Informatik, ...) **...** **02**0 (Bib.- u. Informationswiss.en) **...**
3. **006** (Spez. Computerverfahren) **025 ...**
4. **006.3** (Künstliche Intelligenz) **... [025.0]** **...**
5. **006.31** (Maschinelles Lernen) **... 025.04** (... Information-Retrieval-Systeme)

12-stellig

12. **...** **025.302855741** **...**
(Einsatz von Dateiorganisation und Dateizugriffsverfahren)

DDC-Notationen: 26.715 (Haupttafeln); 9.356 (Hilfstafeln); 13.919 (mit Regeln gebildete)

DDC-Klassifzierung: ein Thema (Kleidung)– mehrere Systemstellen !

Fundy-Nationalpark
ul, 25. Mai 2008

„Da die einzelnen Teile der DDC nach Fachgebieten und nicht nach Themen geordnet sind, kann ein Thema mehrere Systemstellen haben. So kann z. B. das Thema »Kleidung« unter verschiedenen Aspekten aus mehreren Fachgebieten gesehen werden. Die psychologische Wirkung von Kleidung gehört zu 155.95, als Teil des Fachgebiets Psychologie; mit Kleidung verbundene Bräuche gehören als Teil des Fachs Ethnologie zu 391 und Kleidung im Sinn der Modeschöpfung gehört als Teil des Fachgebiets Künste zu 746.92“

Hervorhebungen (Unterstreichungen, farbliche Markierungen) durch Autorin

[DDC 22 Dewey-Dezimalklassifikation und Register (begr. von Melvil Dewey; hrsg. von Joan, S. Mitchell unter Mitwirk. von Julianne Beall; Giles Martin; Winton E. Matthews, Jr.; Gregroy R. New; Mitarbeit: Heidrun Alex; Anne Betz; Winfried Gödert; Magda Heiner-Freiling; Melanie Jackenkroll; Marlene Lambert; Tina Mengel; Michael Preuss; Esther Scheven; Lars G. Svensson). Dt. Ausgabe (hrsg. von Der Deutschen Bibliothek). Band 1, K.G. Saur, München, 2005, S. I]

DDC-Klassifzierung: ein Thema (Der Apfel) – welche Systemstelle?

Fundy-Nationalpark
ul, 25. Mai 2008

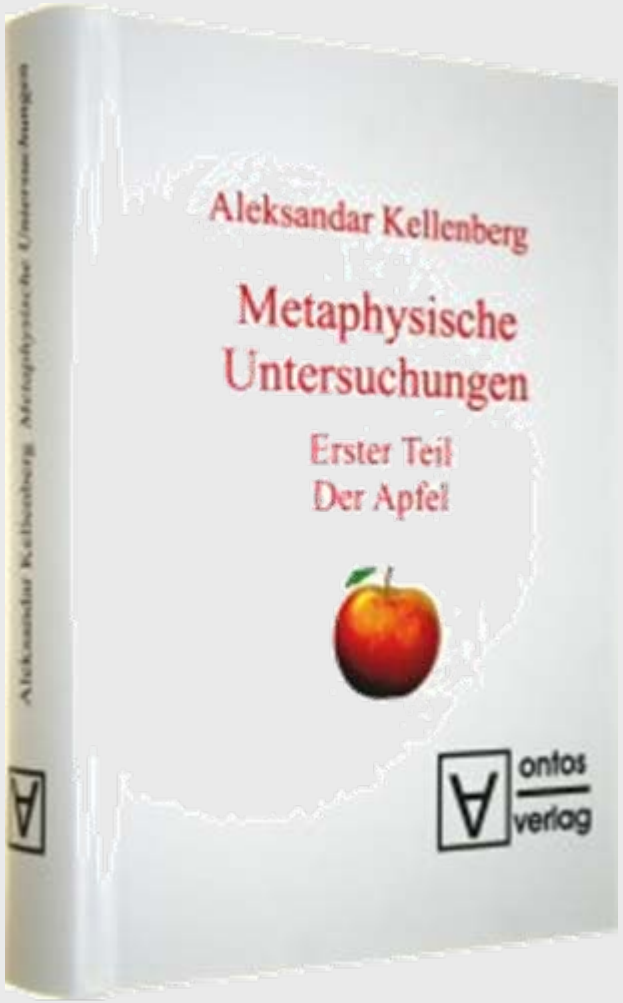
DNB-Titeldatensatz DNB984784829 im MAB2-Format

```
001 984784829
002a20070628
003 20071011111524
004 20071023
010 985958774
025a984784829
026 DNB984784829
030 g|1iaz|z|||||
037bger
050 a|b||||||||||
070 1245
070aDNB
070b1250
089 Teil 1.
090 11
331 Der Apfel
425 2007
425a2007
433 471 S.
540aISBN 978-3-938793-62-6 Pp. : EUR 98.00
540aISBN 3-938793-62-7 Pp. : EUR 98.00
544aFÎ2007 A 65008
553a9783938793626
568 07,N30,0128
574 07,A45,0115
655e□qtext/html□uhttp://deposit.d-nb.de/cgi-bin/dokserv?id=2979287&prov=M&
dok_var=1&dok_ext=htm□3Inhaltstext□A2
700 |100ÎDNB
705a□a110□c110□eDDC22ger
```

Intellektuelle DDC-Klassifizierung

Der Apfel: 110 (Metaphysik)

Fundy-Nationalpark
ul, 25. Mai 2008



[http://cover.deutschesfachbuch.de/books/3938793627/bx.jpg]

OCLC Classify

(an experimental classification web service)

Der Apfel {372.133}

Fundy-Nationalpark
ul, 25. Mai 2008

Classify - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://deweybrowser.oclc.org/classify2/Classify?swid=198826610

Jockweg, Bernd der apfel

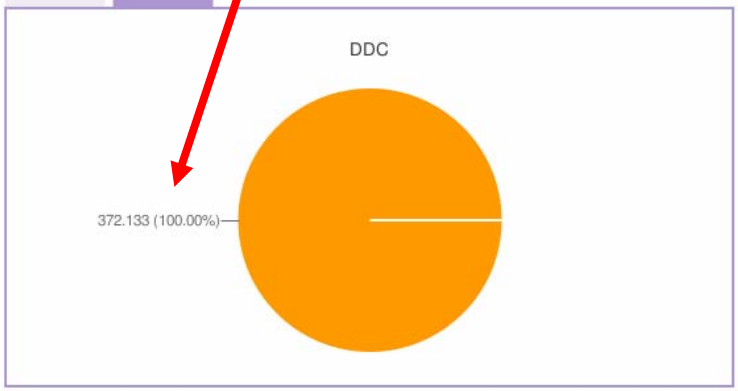
Title: Der Apfel
Author: Jockweg, Bernd
Format: Books & Audio Books Editions: 1 Total Holdings: 1

Jockweg, Bernd: Der Apfel
DDC Klasse: 372.133 (100%)
(Unterrichtsmaterialien--Primärbildung)

Classification Summary

DDC:	Class Number	Holdings
Most Frequent	372.133	1
Most Recent	372.133	1

All DDC



Edition Details

Editions: 1 to 1 of 1

[http://www.curriculum-online.de/itemsimages/9783867230179.jpg]

Title / Author | Lang. | Holdings | Tag | Class # | Library | Src.

Suchen: zelle Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung Das Seitenende wurde erreicht, Suche vom Seitenanfang fortgesetzt

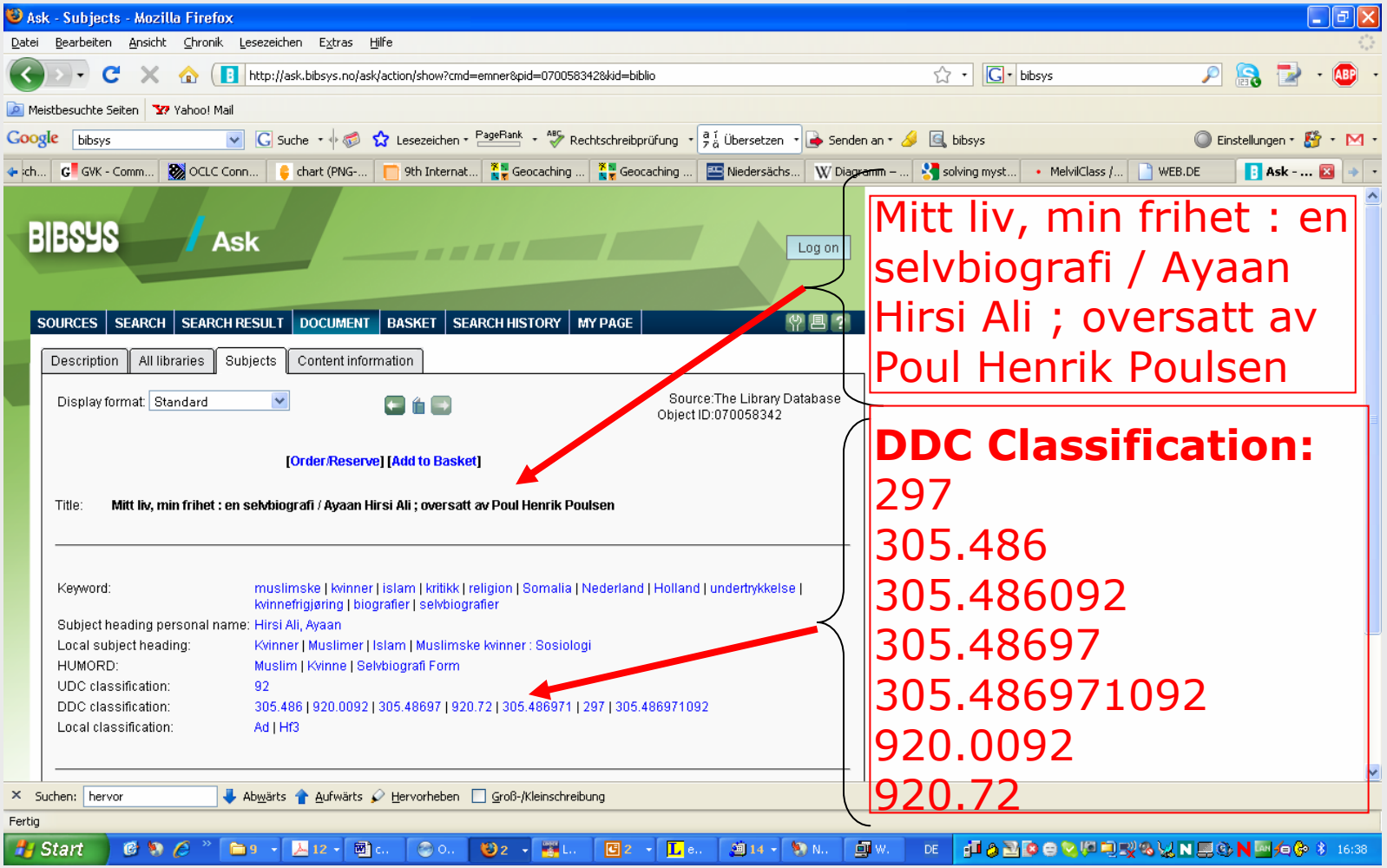
Fertig

Intellektuelle DDC-Klassifizierung

Mitt liv, min frihet : {297,...,920.72}

Fundy-Nationalpark
ul, 25. Mai 2008

Beispiel aus: [Ingebjørg Rype; Magdalena Svanberg: Dewey in Scandinavia: exploring new translation models of Dewey]. Vortrag auf 3. EDUG-Symposium „Dewey goes Europe - On the Use and Development of the Dewey Decimal Classification (DDC) in European Libraries", Vienna 28 April, 2009.
[http://www.onb.ac.at/events/files/rype.ppt], p. 8



Mitt liv, min frihet : en selvbiografi / Ayaan Hirsi Ali ; oversatt av Poul Henrik Poulsen

DDC Classification:
297
305.486
305.486092
305.48697
305.486971092
920.0092
920.72

OCLC Classify
Mein Leben, meine Freiheit : {324.2092}

Fundy-Nationalpark
ul, 25. Mai 2008

Classify - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://deweybrowser.oclc.org/classify2/Classify?swid=180710878

oclc classify

Meistbesuchte Seiten Yahoo! Mail

Google oclc classify Suche Lesezeichen PageRank Rechtschreibprüfung Übersetzen Senden an oclc classify Einstellungen

OCLC Conn... Classify 9th Internat... Geocaching ... Geocaching ... Niedersächs... Diagramm – ... solving myst... MelvilClass /... WEB.DE Ask - Subjects Classify

Summary

Title: Mein Leben, meine Freiheit Auszüge aus der Autobiographie ; gekürzte Lesung
Author: Hirsi Ali, Ayaan
Format: Computer Files **Editions:** 1 **Total Holdings:** 1

Hirsi Ali, Ayaan: **Mein Leben, meine Freiheit** (Holdings: 1)
DDC Klasse: 324.2092 (100%)
(Politiker--Biografien)

Classification Summary

DDC:	Class Number	Holdings
Most Frequent	324.2092	1
Most Recent	324.2092	1

All DDC

DDC

324.2092 (100.00%)

Edition Details

Editions: 1 to 1 of 1

Suchen: hervor Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung

Fertig

Start DE 17:24

VZG Projekt Colibri/DDC

VZG

OCLC Classify: The caged virgin - an emancipation proclamation for women and Islam {297.082,...,922.97}

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Classify - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://deweybrowser.oclc.org/classify2/Classify?swid=058843388

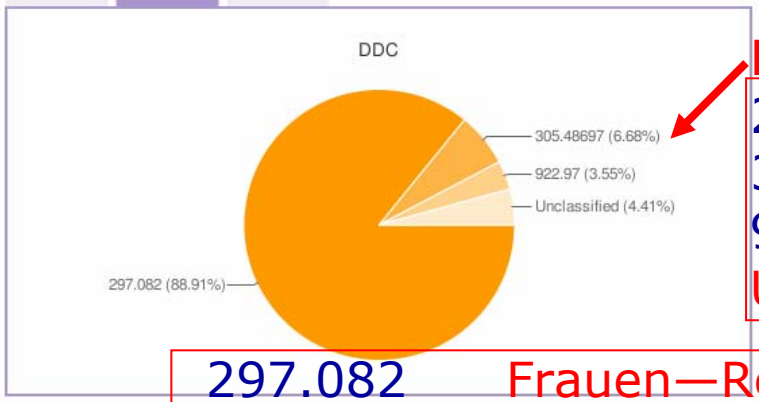
Title: The caged virgin : an emancipation proclamation for women and Islam /
Author: Hirsi Ali, Ayaan,
Format: Books & Audio Books **Editions:** 16 **Total Holdings:** 1407

Hirsi Ali, Ayaan: The caged virgin – an emancipation proclamation for women and Islam (Holdings: 1407)

Classification Summary

DDC:	Class Number	Holdings
Most Frequent	297.082	1251
Most Recent	305.48697	94
Latest Edition: 22	297.082	1251
Latest Edition: 22	305.48697	93
Latest Edition: 22	922.97	50
LCC:	**Class Number**	**Holdings**
Most Frequent	BP173.4	1216
Most Recent	BP173.4	1216

All DDC LCC



DDC Klassen:

297.082	(88.91%)
305.48697	(6.68%)
922.97	(3.55%)
Unclassified	(4.41%)

297.082	Frauen—Religion—Islam
305.48697	Musliminnen--soziale Gruppe, . . .
922.97	(Adherents of Islam)

Edition Details

Editions: 1 to 16 of 16

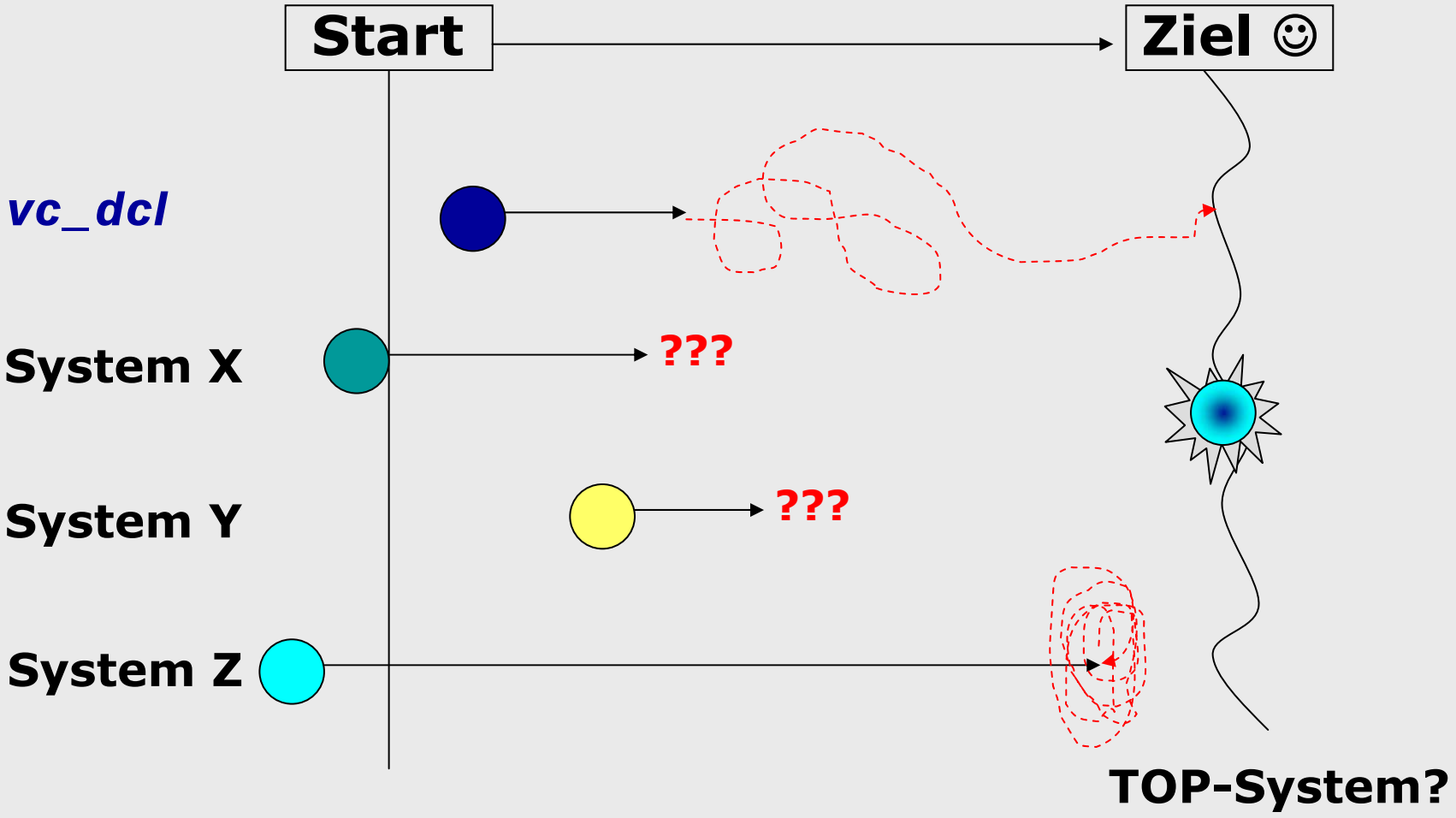
Title / Author | Lang. | Holdings | Tag | Class # | Library | Src.

Suchen: zelle Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung Das Seitenende wurde erreicht, Suche vom Seitenanfang fortgesetzt

Fertig

Initiative Colibri/DDC-Wettbewerb (Juni 2009)
Ziel: bester automatischer DDC-Klassifizierer für bibliografische Titeldatensätze gesucht

Fundy-Nationalpark
ul, 25. Mai 2008



Initiative Colibri/DDC-Wettbewerb (Juni 2009)
Ziel: bester automatischer DDC-Klassifizierer für bibliografische Titeldatensätze gesucht

Fundy-Nationalpark
ul, 25. Mai 2008

Systemtest[1]

- **Modell des Systems oder detaillierte Beschreibung des Systems und seiner Komponenten**
- **Zu testende Hypothesen**
- **Bewertungskriterien und Maße, die diese Kriterien widerspiegeln**
- **Methoden, Daten zu ermitteln und zu bewerten**

[1] **[Salton 1983] Gerard Salton; Michael J. McGill, : Introduction to Modern Information Retrieval. McGraw-Hill, New York u.a. , 1983. S. 158**

Automatische DDC-Klassifizierung (1)
Colibri/DDC-Modell (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Deskriptor (*descr*): Pica+/MAB2-Kategorie, deren Werte zur inhaltlichen Charakterisierung beitragen

Pica+ : {..., 021A, ..., 044K, ...}; MAB2: {..., 310, ..., 410, ...}

Deskriptorwert (*descr_val*): Wert eines Deskriptors

{Apfel, Apfelbeere, Aronia}

DDC-Klasse: Menge von Deskriptorwerten

634:= {..., <021A>-aronia, ...}

Titeldatensatz: Menge von Deskriptorwerten

DNB991499077 := {..., <331>-aronia, <902s>-aronia, ...}

DDC-Datenbasis *vc_DB*: Menge von DDC-Klassen, repräsentiert durch DDC-Notationen (dnos)

{000, 006.31, 025.302855741, 634 , ..., 999.23}

Automatische DDC-Klassifizierung (2)
Colibri/DDC-Modell (2)


Fundy-Nationalpark
ul, 25. Mai 2008

IR[1]-Komponente von *vc_dcl*

Vektorprodukt[2] als Ähnlichkeitsmaß s: $s_{uc} = \sum_{i=1}^{l} u_i c_i$

Bestimmung der Ähnlichkeit zwischen den binären Vektoren *u*[3] und *c*[4] mit *s*: similarity (Ähnlichkeit); i: i-te Gewicht (1: Deskriptorwert vorhanden; 0: Deskriptorwert nicht vorhanden); l : Anzahl der Deskriptorwerte von *u*.

DDC-Notationskandidat(en) für einen Titeldatensatz

DDC-Notationskandidat: DDC-Klasse mit größtem Ähnlichkeitswert zwischen *u* und *c*: $s_{uc\,\max}$

Menge von DDC-Notationskandidaten (*dno_cand_set*):
DDC-Klassen mit gleichen Ähnlichkeitswerten

[1] **IR: Information Retrieval;** [2] **[Salton 1968] Gerard Salton: Automatic Information Organization and Retrieval. McGraw-Hill, New York, 1968, p. 237;**

[3] ***u*: unclassified (Deskriptorwerte eines nicht klassifizierten Titeldatensatzes);**

[4] ***c*: classified (Deskriptorwerte einer DDC-Klasse)**

Automatische DDC-Klassifizierung (3)
Colibri/DDC-Modell (3)


Fundy-Nationalpark
ul, 25. Mai 2008

KI[1]-Komponente von *vc_dcl* (1)
Heuristische Funktion *cutoff_val*[2]

Annahme: Deskriptorwerte, die in zu vielen DDC-Klassen auftreten, sind (mit bestimmter Ausnahme) für die automatische DDC-Klassifizierung ungeeignet.

cutoff_val
Obergrenze für Berücksichtigung von Häufigkeitswerten von Deskriptorwerten

cutoff_val_dyn
Wert wird zur Laufzeit durch heuristische Regeln dynamisch bestimmt

cutoff_val_stat
statischer (= konstanter) Wert für Testzwecke

in_descr_val_lim = 6
Anzahl der zu berücksichtigenden Deskriptorwerte (Anfangswert)

[1] **KI: Künstliche Intelligenz;** [2] **[Reiner 2009] , S. 12ff**

Automatische DDC-Klassifizierung (4)
Colibri/DDC-Modell (4)


Fundy-Nationalpark
ul, 25. Mai 2008

KI-Komponente von *vc_dcl* (2)

Heuristische Regeln, z. B.[1]

H2. Berücksichtigung spezifischer Begriffe

Wenn Differenz zwischen zwei Häufigkeitswerten größer als „200“
dann *cutoff_val_dyn* := kleinerer Wert der beiden Häufigkeitswerte

H3. Berücksichtigung auch allgemeiner Begriffe

Wenn Summe der 1- bis 3-stelligen Häufigkeiten kleiner als Anzahl der größer als 3-stelligen Häufigkeiten
dann *cutoff_val_dyn* := größter Häufigkeitswert (allgemeine Begriffe überwiegen im Titeldatensatz).

[1] **[Reiner 2009] , S. 12**

Automatische DDC-Klassifizierung (5)
VZG Colibri/DDC-Suchsystem *vc_ds*


Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Anfragen

DNB991499077 (Aronia, Folie 56) ohne automatisch ermittelte DDC-Notation

615.32373

{600,610,615, 615.3,615.32, 615.323, 615.32373,583, 583.7,583.73}

615.32373

vc_dcl
vzg colibri_ddc classifier

vc_day
vzg colibri_ddc number analyzer

vc_dsy
vzg colibri_ddc number synthesizer

vc_dqa
vzg colibri_ddc question answerer

Antworten

DNB991499077 mit DDC-Notationskandidat 615.321 (Folie 30)

{600,610,615, 615.3,615.32, 615.323, 615.32373,583, 583.7,583.73}

615.32373

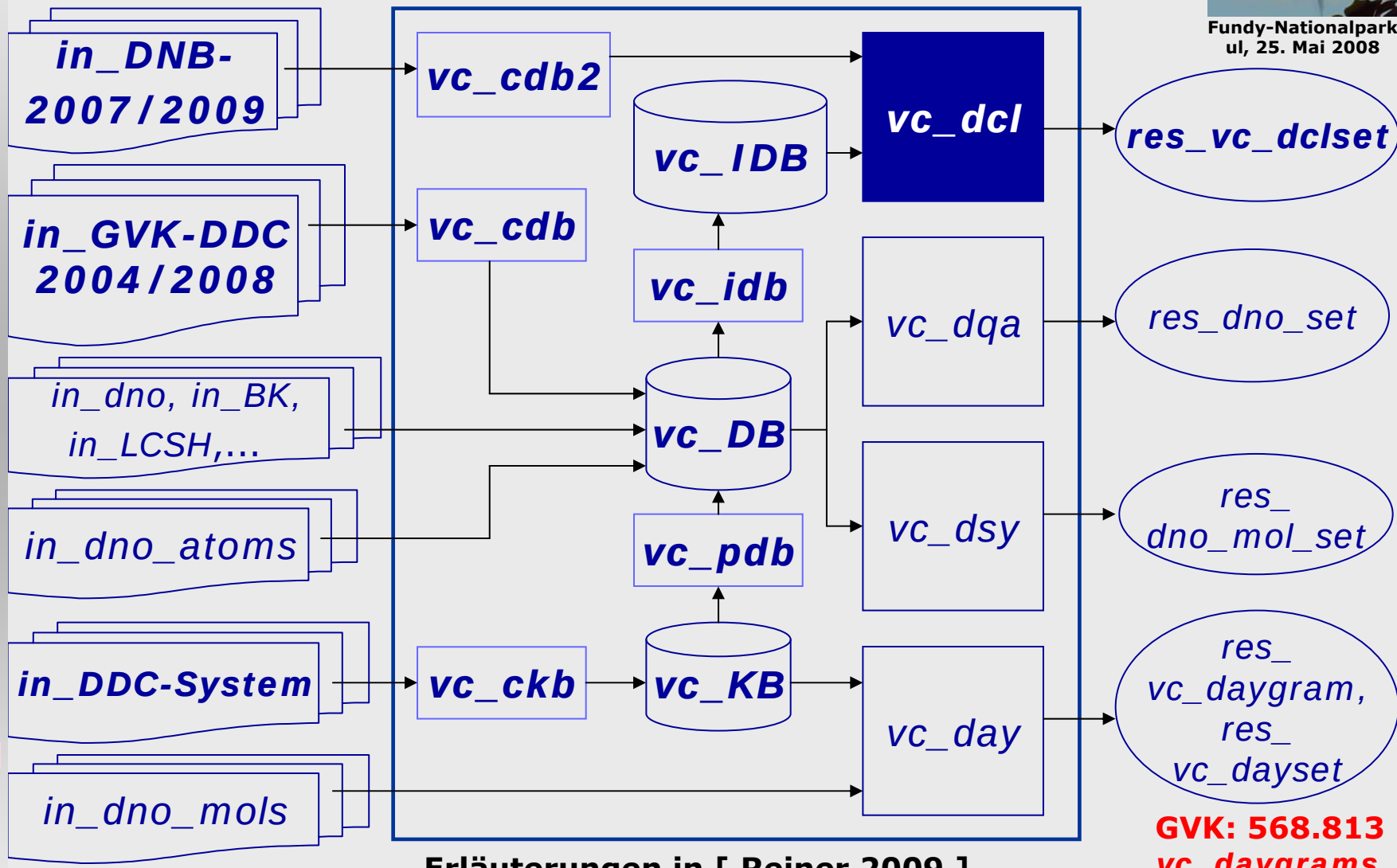
LCC: RM300-666 (Drugs and their actions)

VZG

Automatische DDC-Klassifizierung (6)

Softwaresystem-Architektur DDC-Suchsystem *vc_ds*

Fundy-Nationalpark
ul, 25. Mai 2008



Erläuterungen in [Reiner 2009]

GVK: 568.813
vc_daygrams
(24.3.09)

VZG Projekt Colibri/DDC

VZG

Automatische DDC-Klassifizierung (7)
Standard-Testbestände: Information Retrieval

Fundy-Nationalpark
ul, 25. Mai 2008

- **Cranfield** (1950)[2]
  1398 Kurzfassungen (Aerodynamik-Zeitschriftenartikel), 225 Anfragen, Relevanzurteile

- **TREC** (NIST, 1992)[2]
  "Ad Hoc track" für TREC1 – TREC8 (1992-1999), 6 CD's: 1.89 Mio. Dokumente, 450 Anfragen ("topics"), Relevanz-urteile, "TREC 6-8": 528.000 Artikel, 150 Anfragen

- **GOV2** (2004)[1]
  27 Mio. WWW-Seiten, 15 KB durchschnittliche Dokumentengröße

- **Cross Language Evaluation Forum (CLEF)** (2000)[3]
  Europäische Sprachen, sprachübergreifendes Information Retrieval

- **REUTERS** (1996-2004)[2]
  Reuters-21578: 21.578 Artikel von Nachrichtenagenturen
  RCV1 (Reuters Corpus Volume, 1GB): 806.791 Dokumente (z. B. aus Politik, Wirtschaft, Sport, Wissenschaft)

- **20 NEWSGROUPS**[2]
  1000 Artikel von 20 Usenet-Newsgroups

[1] [Voorhees/Harman 2005], S. 21-52; [3] [CLEF]; [2] [Manning/Raghavan/Schütze 2008], S. 153. Online: http://nlp.stanford.edu/IR-book/pdf/08eval.pdf;

Automatische DDC-Klassifizierung (8)

Colibri/DDC-Systemtest: Testbestände

Fundy-Nationalpark
ul, 25. Mai 2008

Basis für die automatische Klassifizierung

DDC-Datenbasis[1]

vc_DB-2004*, *vc_DB-2008

DDC-Wissensbasis[1]

vc_KB-2004

DDC-Testbestände (Testdokumente)

in_DNB-2007*, *in_DNB-2009

Andere Kollektionen als DDC-Testbestände?

100.000 BASE-Titeldatensätze[2]
426.254 NORBOK-Titeldatensätze[3]

[1] [Reiner 2009] ; analog zu DDC-Daten-/Wissensbasis auch Fallbasis/Trainingsdokumente (Maschinelles Lernen); z. B. [Pfeffer 2008]; [Oberhauser 2004]; [Wille 2006]; [Mehler/Waltinger 2009a]; [Mehler/Waltinger 2009b];

[2] [http://base.ub.uni-bielefeld.de/en/lab_browse_menu.php?menu=5]

[3] [http://nabo.nb.no/trip?_b=baser&navn=norbok&_h=0]

Automatische DDC-Klassifizierung (9)

Colibri/DDC-Systemtest

Fundy-Nationalpark
ul, 25. Mai 2008

Kriterium für Tests / Experimente

Wiederholbarkeit!

Verwendete Hard- und Software (*colibri2.gbv.de*)

HP Proliant DL585 G1, 4xAMD Opteron 275, 2.2 GHz, 16GB Hauptspeicher. SuSE Linux Enterprise 10, gawk-3.1.5.

vc_dcl_srv.awk **(Server): 1.222 Zeilen Programmcode;**
vc_dcl_cli.awk **(Client): 27 Zeilen Programmcode.**

Automatische DDC-Klassifizierung (10)
Eingabedaten


Fundy-Nationalpark
ul, 25. Mai 2008

in_DDC-System

Elektronische Form als XML-Datei **(22. Aufl., in Englisch, Januar 2004)**

in_GVK-DDC-2004 **(Pica+ - Format)**

3,0 Mio. Titeldatensätze

in_GVK-DDC-2008[1] **(Pica+ - Format)**

4,3 Mio. Titeldatensätze[2]

in_DNB-2007 **bzw.** ***in_DNB-2009*** **(MAB2-Format)**

12 DNB-Wochen/Monatslieferungen der Deutschen Nationalbibliografie der Reihen A, B und H mit intellektuell vergebenen DDC-Notationen aus den Jahren 2007 bzw. 2009

- *in_DNB-2007* (25.653 Titeldatensätze, 10,5 Deskriptorwerte im ∅)
- *in_DNB-2009* (30.717 Titeldatensätze, 11,0 Deskriptorwerte im ∅)

[1] **Zum Vergleich: GVK: 28,2 Mio. Titeldatensätze, Nov. 2008;** [2] **LoC: 54,7%; BNB: 23,8%; Quelle nicht rekonstruierbar: 23,5%**

Automatische DDC-Klassifizierung (11)
Datenkonvertierung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Datenkonvertierung

- Eliminierung: irrelevante Deskriptorwerte, Sonderzeichen

- Deskriptorwerte: Transliterierung, Kleinschreibung

Berücksichtigte MAB2-Felder (*vc_cdb2*)

026 (Regionale Identifikationsnummer); 037 (Sprachencode nach ISO 639); 100 (Name der 1. Person in Ansetzungsform); 310 (Hauptsachtitel in Ansetzungsform); 331 (Hauptsachtitel in Vorlageform oder Mischform); 335 (Zusätze zum Hauptsachtitel); 341 (1. Parallelsachtitel in Vorlageform oder Mischform); 370 (Weitere Sachtitel); 410 (Ort(e) des 1. Verlegers, Druckers usw.); 412 (Name des 1. Verlegers, Druckers usw.); 451 (1. Gesamttitel); 540 (Internationale Standardbuchnummer (ISBN)); 542 (Internat. Standardnr. für fortlauf. Sammelwerke); 700 (Systematik der katalogisierenden Institution); 705 (DDC analytisch); 902/12/22 s/g, 907/17/27 s/g (Sach-/geographisch-ethnographische Schlagworte);

Automatische DDC-Klassifizierung (12)
Datenkonvertierung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

Berücksichtigte Pica+ - Kategorien (*vc_cdb*)

001A (Kennung der Ersterfassung); 003@ (Pica production number); 004A (ISBN); 004B (2. und weitere ISBN); 004D (formal falsche ISBN); 005A (ISSN); 006G (DNB-Nummer); 006L (Weitere Verbundidentifikationsnummern); 006Y (Verbundidentifikationsnummer); 007G (Identifikationsnummer der ersterfassenden Institution); 021A (Hauptsachtitel, Verfasser); 022A/01(Einheitssachtitel); 027D (Titel in Bandsätzen); 028A (1. Verfasser); 028B (2. und weitere Verfasser); 028C (Sonstige beteiligte Personen); 028E (Interpreten); 033A (Ort, Verlag); 036C (Gesamtheit und Abteilungen in Vorlageform); 039B (Verknüpfung zur größeren Einheit); 041A (Kettenglied einer RSWK-Kette); 044A (Library of Congress Subject Headings (LCSH)); 044C (Medical Subject Headings (MESH)); 044E (PRECIS); 044F (DNB-Schlagwörter); 044G (British Library Subject Headings (BLSH)); 044K (Einzelschlagwort); 044L (Einzelschlagwort (Projekte)); 045A (Library of Congress Classification (LCC)); 045F (DDC); 045Q (Basisklassifikation); 045U ZDB (Notation bei Zeitschriften); 144Z/244Z (Lokale Schlagwörter); 145Z/245Z (Lokale Notationen);

Automatische DDC-Klassifizierung (13)
Erstellung der DDC-Daten-/Wissensbasis

Fundy-Nationalpark
ul, 25. Mai 2008

vc_ckb **(*vzg colibri_create ddc knowledge base*)**

Erstellung der DDC-Wissensbasis *vc_KB-2004* aus Daten des DDC-Systems (Januar 2004)

vc_cdb **(*vzg colibri_create ddc data base*)**

Erstellung der DDC-Datenbasis *vc_DB* aus GVK-DDC-Titeldatensätzen (Pica+ -> *vc_DB*-Repräsentation)

vc_cdb2 **(*vzg colibri_create ddc data base2*)**

Konvertierung der DNB-Titeldatensätze in *vc_DB*-Repräsentation (MAB2 -> *vc_DB*-Repräsentation)

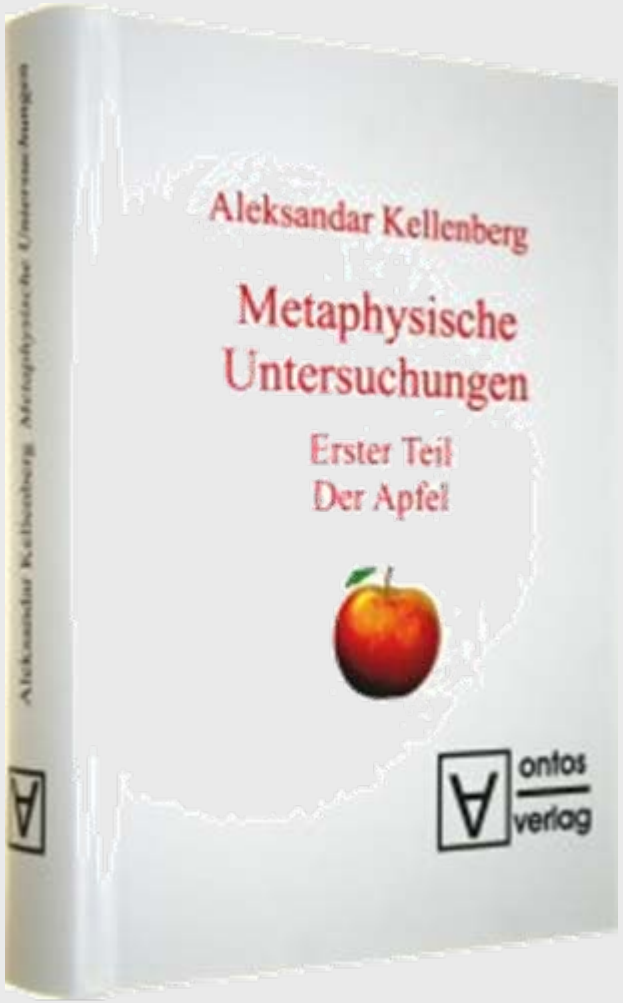
vc_pdb **(*vzg colibri_prepare ddc data base*) und *vc_idb*** **(*vzg colibri_create inverted ddc data base*)**

Erstellung der invertierten DDC-Datenbasis *vc_IDB*

vc_IDB-2004 (510 MB): ca. 3 Min. Einlesezeit in den Hauptspeicher
vc_IDB-2008 (712 MB): ca. 5 Min. Einlesezeit in den Hauptspeicher

Intellektuelle DDC-Klassifizierung
Der Apfel: 110 (Metaphysik)

Fundy-Nationalpark
ul, 25. Mai 2008



[http://cover.deutschesfachbuch.de/books/3938793627/bx.jpg]

Automatische DDC-Klassifizierung (14)

Der Apfel: {334.683411, … ,391.0092}

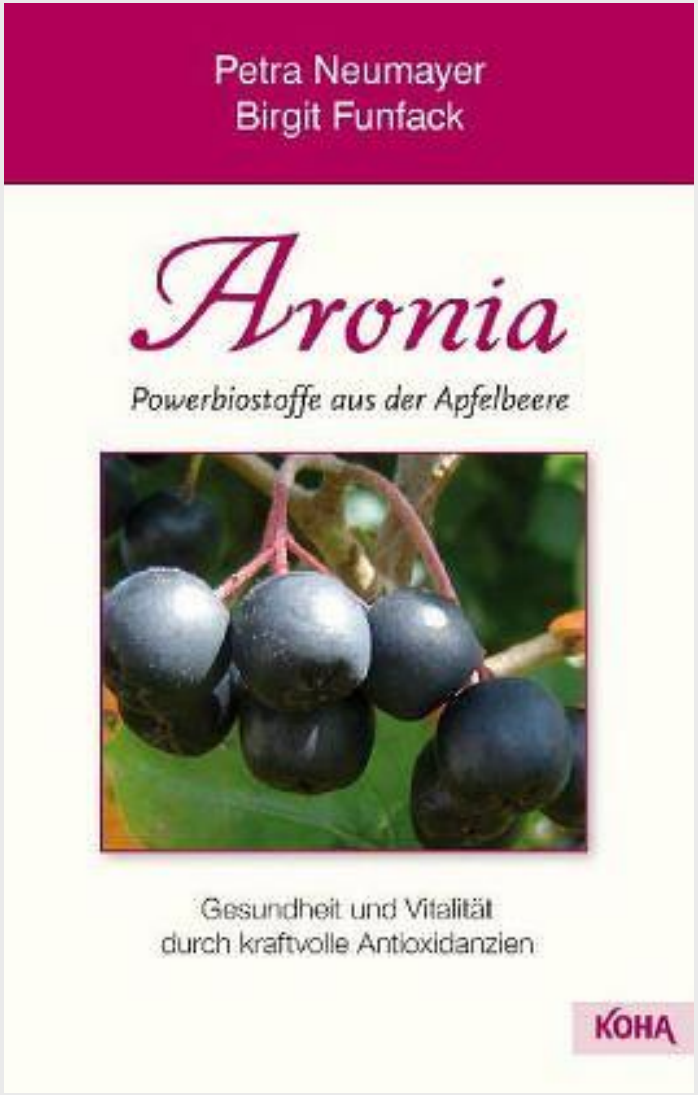
Fundy-Nationalpark
ul, 25. Mai 2008

Ethik

```
number of ddc-classified title:      1197
identifier (dno,schedno):            DNB0984784829 (110,110) DNB DDC
notation (MAB2 field 700):  {100}
DDC notation (MAB2 field 705):       {110}
calculated cutoff value:             31
title:                               Der Apfel
considered descriptor values:        |2| {<331>-apfel[31], <540a>-3-938793-
62-7[0]}
matched descriptor values:           |1| {apfel}
max. match value of matched descriptor values: |1|
calculated1 ddc classes (subdiv):    |31| {070.924, 170, 300, 334.683411,
338.108, 338.10942, 338.13, 338.17411, 338.174110942, 343.73084, 370,
380.1414110943, 391.0092, 581.12, 634.11, 634.116, 634.117, 634.1193,
634.11943, 635.08, 641.341109748, 791.430233092, 813.54, 822.33,
823.0872909, 823.7, 823.914, 823.92, 833.914, 839.3135, 892.493}
calculated1 ddc classes (sections): |7| {300, 334, 338, 343, 370, 380, 391}
calculated1 ddc classes (main):      |1| {300}
calculated2 ddc classes (subdiv):    |7| {334.683411[1], 338.108[1],
338.10942[1], 338.13[1], 338.17411[1], 343.73084[1], 391.0092[1]}
calculated2 ddc classes (sections): {338[5]}
calculated2 ddc classes (divisions):{330[6]}
calculated2 ddc classes (main):      {300[11]}
correlation(dnb_A0745_DNB0984784829#ger#dno_i{110}#dno_a{M300,D330,S338,s33
4.683411,s338.108,s338.10942,s338.13,s338.17411,s343.73084,s391.0092}#consi
: 2#matched: 1,1{apfel}): 00x.xxx xxx xxx xxx (0)
```

Automatische DDC-Klassifizierung (15) ?

Fundy-Nationalpark
ul, 25. Mai 2008



[http://picture.yatego.com/images/428b84fecc19b0.4/pid4748589.jpg]

VZG Projekt Colibri/DDC

VZG

Automatische DDC-Klassifizierung (16)

Powerbiostoffe aus der Apfelbeere : {615.321}

Fundy-Nationalpark
ul, 25. Mai 2008

```
number of ddc-classified title:      996
identifier (dno,schedno):            DNB0991499077 (615.32373,615.32373)
DNB DDC notation (MAB2 field 700):   {610}
DDC notation (MAB2 field 705):       {615.32373}
calculated cutoff value:             79
title:                               Aronia
title (remainder):                   Powerbiostoffe aus der Apfelbeere ;
Gesundheit und Vitalitaet durch kraftvolle Antioxidanzien
considered descriptor values:        |11| {<100>-petra#neumayer[0], <331>-
aronia[1], <335>-apfelbeere[0], <335>-gesundheit[823], <335>-vitalitaet[79],
<335>-kraftvolle[1], <335>-antioxidanzien[0], <335>-powerbiostoffe[0],
<412@410>-<033A>-koha@burgrain[0], <540a>-978-3-86728-084-6[0],
<902s1>-naturheilmittel[10]}
matched descriptor values:           |2| {naturheilmittel, vitalitaet}
max. match value of matched descriptor values: |2|
calculated1 ddc classes (subdiv):    |1| {615.321}
calculated1 ddc classes (sections):  |1| {615}
calculated1 ddc classes (main):      |1| {600}
calculated2 ddc classes (subdiv):    |1| {615.321[1]}
calculated2 ddc classes (sections):  {615[1]}
calculated2 ddc classes (divisions):{610[1]}
calculated2 ddc classes (main):      {600[1]}
correlation(dnb_A0912_DNB0991499077#ger#dno_i{615.32373}#dno_a{M600,D610,S61,
s615.321}#consi: 11#matched: 2,2{naturheilmittel, vitalitaet}):
111.110 00x xxx xxx (0.625)
```

Automatische DDC-Klassifizierung (17)
Bewertung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Projekt Colibri/DDC[1]: **Korrelationsmaße *C*, *CP*[2] und *CN*[2]**

Stellenweiser Ziffernvergleich von links nach rechts zwischen intellektuell vergebener (*dno_i*) und automatisch ermittelter DDC-Notation (*dno_a*). Annahme: *dno_i* ist optimal. *L_i*: Länge von *dno_i*.

***C* (Correlation)**: Anzahl der übereinstimmenden Ziffern in *dno_i* und *dno_a*.

***CP* (Correlation Pattern)**: 16-stelliges Muster mit „." (Dewey Punkt) an Stelle 4; „1", wenn *dno_i* und *dno_a* an Stelle s übereinstimmen; „0", wenn sie nicht übereinstimmen; „x" an Stellen größer *L_i* .

***CN* (Correlation Number)**: auf *L_i* normiertes Korrelationsmaß *CN* = *C* / *L_i* .

[1] **Mathematische Definitionen in [Reiner 2009] , S. 13ff;** [2] **eingeführt in [Reiner 2008], S. 127**

Automatische DDC-Klassifizierung (18)
Bewertung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

Projekt Colibri/DDC[1]

dno_i **= 150** **(Psychologie)**
dno_a **= 158.1** **(Persönliche Weiterentwicklung und Analyse)**

CP **= 110.xxx xxx xxx xxx;** ***CN*** **= (1+1+0)/3 = 0.66666**

dno_i **= 158.1** **(Persönliche Weiterentwicklung und Analyse)**
dno_a **= 158** **(Angewandte Psychologie)**

CP **= 111.0xx xxx xxx xxx;** ***CN*** **= (1+1+1+0)/4 = 0.75**

dno_i **= 591.513** **(Intelligenz) [Oberklasse: 590 (Tiere)]**
dno_a **= 156.39** **(Intelligenz bei Tieren--vergleichende Psychologie, ...)**

CP **= 000.000 xxx xxx xxx;** ***CN*** **= (0+0+0+0+0+0)/6 = 0**

[1] **[Reiner 2009] , S. 15**

Automatische DDC-Klassifizierung (20)
Bewertung (4)

Fundy-Nationalpark
ul, 25. Mai 2008

Projekt Pfeffer/RVK[1]

„Bewertung

– Vergleich der automatischen und manuellen Klassifikation

– Suche des nächsten gemeinsamen Vaterknoten im RVK-Baum

- **Perfekt: Übereinstimmung**
- **Gut: Abstand 1-3**
- **Mäßig: Abstand >3, aber noch gleiches Fach**
- **Schlecht: anderes Fach“**

[1] **[Pfeffer 2008], S. 10**

Automatische DDC-Klassifizierung (21)
Bewertung (5)


Fundy-Nationalpark
ul, 25. Mai 2008

Colibri/DDC					Scorpion/DDC							Pfeffer/RVK
Bsp.dno_i	dno_a	CP	C	CN	S1	S2	S3	S4	S8	S9	S10	P
8. 529.326	529.326	111.111	6	1	x	x	x		x	x	x	P1
9. 529	529.3	111	3	1	x	x	x	G		x	x	P2
10. 529.3	529	111.0	3	0.75	x	x	x	Sp		x	x	P2
11. 111	115	110	2	0.66	x	x					x	P2
12. 520	529	110	2	0.66	x	x					x	P2
13. 571.68	571.58	111.00	3	0.60	x	x	x				x	P2
14. 111.8	110	110.0	2	0.50	x	x					x	P2
15. 571.5929	571	111.0000	3	0.43	x	x	x	Sp		x	x	P3
16. 111.85	110	110.00	2	0.40	x	x					x	P2
17. 111.850952	111	111.000000	3	0.33	x	x	x	Sp		x	x	P4
18. 572.6	500	100.0	1	0.25	x						x	P2
19. 111.85	100	100.00	1	0.20	x						x	P3
20. 529.326	500	100.000	1	0.16	x						x	P3
21. 571.5929	500	100.0000	1	0.14	x						x	P3
22. 100	500	000	0	0								P4
23. 170	570	000	0	0								P4

Abb. 8: Vergleichende Betrachtung mit unterschiedlichen Bewertungsmaßen; Auszug aus [Reiner 2009], S. 18

Automatische DDC-Klassifizierung (23)
Bewertung (7)


Fundy-Nationalpark
ul, 25. Mai 2008

Bewertungsmasse[1] (2)

- **Precision** **P** = a / (a+b)
- **Recall** **R** = a / (a+c)
- **Fallout** **F** = b / (b+d)
- **F-Measure** = 2*P*R / (P+R)

[1] **[Salton 1968]; [Sasaki 2007] Yutaka Sasaki: The truth of the F-measure. School of Computer Science, University of Manchester MIB, 131 Princess Street, Manchester, M1 7DN, October 26, 2007. Online: http://personalpages.manchester.ac.uk/staff/yutaka.sasaki/F-measure-YS-26Oct07.pdf**

Automatische DDC-Klassifizierung (24)
Bewertung (8)


Fundy-Nationalpark
ul, 25. Mai 2008

Weitere Bewertungsmasse[1] (3)

- Accuracy = a+d / (a+b+c+d)
- Error = 1 – Accuracy

- Percent too specific
- Percent too general
- Average overlap
- Accuracy at level

} [2]

- Eleven-point average precision
- Precision-recall breakeven point

[1] [Oberhauser 2004], S. 21 ff. ; [2] [Frank/Paynter]: Predicting Library of Congress Classifications From Library of Congress Subject Headings. Journal of the American Society for Information Science and Technology, Vol. 55, No. 3; p. 222

Automatische DDC-Klassifizierung (26)

Klassifizierungsergebnisse mit *vc_dcl*

Fundy-Nationalpark
ul, 25. Mai 2008

Aus Gründen der Bewertung wird eine automatische Klassifizierung nur durchgeführt, wenn der Titeldatensatz

- **eine korrekte DDC-Notation enthält,**
- **noch nicht klassifiziert wurde (Prüfung: MAB2-Feld 026)**
- **nicht in der DDC-/Wissensbasis enthalten ist (Prüfung: MAB2-Felder 540a, 540b, 004A mit Pica+ - Kategorien 004A, 004B,004D, 005A)**

Name der Ergebnisdatei *res...*	res (Anz.)[1]	tit (Anz.)[2]	t[3]
res_vc_IDB-2004_in_DNB-2007	16.694	25.653	133
res_vc_IDB-2008_in_DNB-2007	15.365	25.653	136
res_vc_IDB-2004_in_DNB-2009	21.591	30.717	120
res_vc_IDB-2008_in_DNB-2009	21.422	30.717	140

[1]Anzahl der Klassifizierungsergebnisse; [2]Anzahl der Titeldatensätze;
[3]Laufzeit der automatischen Klassifizierung in Minuten

Automatische DDC-Klassifizierung (27)
Automatisch bewertete Ergebnisse (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Hypothesen

a) **Unterschied bei unterschiedlichen Daten-/Wissens-basen ist signifikant**

b) **Unterschied bei verschiedenen Testbeständen ist nicht signifikant**

Test-bestand	Daten-/Wissensbasis vc_DB-2004	vc_DB-2008	Differenz
in_DNB-2007	57.33%	62.84%	+5.51%
in_DNB-2009	57.26%	63.85%	+6.59%
Differenz	-0.07%	+1.01%	

***CN*-Werte > 0**

Übereinstimmung mindestens in der DDC-Hauptklasse

Automatische DDC-Klassifizierung (28)

Automatisch bewertete Ergebnisse (2)

res_vc_IDB-2008_in_DNB-2009

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

---------- CN for all dnos ----------

CN=0:	7743;	36,15%	
0<CN<1:	10954;	51,13%	63,85%
CN=1:	2725;	12,72%	

Übereinstimmung mindestens in der DDC-Hauptklasse

----------C ----------

C=0	C=1	C=2	C=3	C=4	C=5	C=6	C=7	C=8	C=9
36,15%	13,71%	26,29%	10,27%	6,07%	3,29%	2,81%	0,98%	0,29%	0,13%

Verteilung der Übereinstimmungen

VZG

Automatische DDC-Klassifizierung (29)

Automatisch bewertete Ergebnisse (3)

res_vc_IDB-2008_in_DNB-2009

Fundy-Nationalpark
ul, 25. Mai 2008

Hypothese: Es gibt signifikante Unterschiede zwischen den DDC-Klassen

CN (Anzahl pro DDC-Klasse)

	dno0	dno1	dno2	dno3	dno4	dno5	dno6	dno7	dno8	dno9
CN=0	249	311	252	849	178	1426	3188	639	141	510
0<CN<1	230	174	347	3015	142	1326	4471	528	207	514
CN=1	77	60	73	421	55	290	1136	204	77	332
CN>0	307	234	420	3436	197	1616	5607	732	284	846

CN (Prozentwerte pro DDC-Klasse)

	dno0	dno1	dno2	dno3	dno4	dno5	dno6	dno7	dno8	dno9
CN=0	44,78%	57,06%	37,50%	19,81%	47,47%	46,88%	36,25%	46,61%	33,18%	37,61%
0<CN<1	41,37%	31,93%	51,64%	70,36%	37,87%	43,59%	50,84%	38,51%	48,71%	37,91%
CN=1	13,85%	11,01%	10,86%	9,82%	14,67%	9,53%	12,92%	14,88%	18,12%	24,48%
CN>0	55,22%	42,94%	62,50%	80,18%	52,54%	53,12%	63,76%	53,39%	66,83%	62,39%

Automatische DDC-Klassifizierung (30)
Automatisch bewertete Ergebnisse (4)
res_vc_IDB-2008_in_DNB-2009


Fundy-Nationalpark
ul, 25. Mai 2008

Hypothese: Es gibt signifikante Unterschiede zwischen den DDC-Klassen

	a	b	c	d	a+b	a+c	Precision	Recall	Fallout	F-Measure
dno0	307	453	241	20421	760	548	0,404	0,560	0,022	0,469
dno1	234	751	311	20126	985	545	0,238	0,429	0,036	0,306
dno2	420	334	251	20417	754	671	0,557	0,626	0,016	0,589
dno3	3436	3613	842	13531	7049	4278	0,487	0,803	0,211	0,607
dno4	197	174	178	20873	371	375	0,531	0,525	0,008	0,528
dno5	1616	1123	1388	17295	2739	3004	0,590	0,538	0,061	0,563
dno6	5607	1801	3130	10884	7408	8737	0,757	0,642	0,142	0,695
dno7	732	601	630	19459	1333	1362	0,549	0,537	0,030	0,543
dno8	284	1428	141	19569	1712	425	0,166	0,668	0,068	0,266
dno9	846	939	506	19131	1785	1352	0,474	0,626	0,047	0,539

Automatische DDC-Klassifizierung (31)

Automatisch bewertete Ergebnisse (5)

res_vc_IDB-2008_in_DNB-2009

Fundy-Nationalpark
ul, 25. Mai 2008

<u>Hypothese:</u> Es gibt keinen signifikanten Unterschied zwischen deutschen und englischen Titeldatensätzen

```
----------------- CN for ger -----------------------------------
ger:  CN=0:  6338; 36,55%
ger:0<CN<1:  8923; 51,46%
ger:  CN=1:  2079; 11,99%  } 63,45%
----------------------------------------------------------------

----------------- CN for eng -----------------------------------
eng:  CN=0:  1400; 33,14%
eng:0<CN<1:  2188; 51,79%
eng:  CN=1:   637; 15,08%  } 66,86%
----------------------------------------------------------------
```

Automatische DDC-Klassifizierung (32)

Automatisch bewertete Ergebnisse (6)

Fundy-Nationalpark
ul, 25. Mai 2008

<u>Hypothese</u>: Es gibt signifikante Unterschiede zwischen den Reihen A, B und H

Name der Ergebnisdatei *res…*	A	B	H
res_vc_IDB-2004_in_DNB-2007	62.32%	50.37%	55.24%
res_vc_IDB-2008_in_DNB-2007	67.72%	58.69%	60.92%
res_vc_IDB-2004_in_DNB-2009	59.56%	49.35%	57.10%
res_vc_IDB-2008_in_DNB-2009	67.42%	56.33%	62.96%

***CN*-Werte > 0**

Übereinstimmung mindestens in der DDC-Hauptklasse

Automatische DDC-Klassifizierung (33)

Automatisch bewertete Ergebnisse (7)

res_vc_IDB-2008_in_DNB-2009

Fundy-Nationalpark
ul, 25. Mai 2008

<u>Hypothese:</u> Es gibt signifikante Unterschiede hinsichtlich der Stelligkeit der DDC-Notationen

	1-3-digit	4-digit	5-digit	6-digit	7-digit	8-digit	9-digit
SUM:	2922	5319	5009	4924	1985	764	317
CN=0:	28,27%	37,75%	38,07%	36,94%	37,38%	35,21%	33,12%
0<CN<1:	47,23%	47,77%	52,67%	52,36%	54,06%	57,98%	58,68%
CN=1:	24,50%	14,48%	9,26%	10,70%	8,56%	6,81%	8,20%
CN>0:	71,73%	62,25%	61,93%	63,06%	62,62%	64,79%	66,88%

VZG Projekt Colibri/DDC

VZG

Stand: Automatisches Klassifizierungsverfahren mit der Klassifizierungskomponente *vc_dcl*

Fundy-Nationalpark
ul, 25. Mai 2008

- Ermittlung der DDC-Notationskandidaten: Algorithmus verwendet IR[1]- und KI[2]-Verfahren
- IR: einfachstes Ähnlichkeitsmaß (binäre Vektoren, Vektorprodukt); KI: heuristische Regeln
- 2 Klassenaggregationen für Ergebnisausgabe
- keine Volltexte, sondern einzelne - ggf. mehrere zusammenhängende - Wörter
- keine linguistischen Verfahren, kein Lexikon

[1] **IR: Information Retrieval;** [2] **KI: Künstliche Intelligenz**

Perspektiven zur automatischen DDC-Klassifizierung (1)

Fundy-Nationalpark
ul, 25. Mai 2008

- Vergrößerung der DDC-Daten-/Wissensbasis
- Verbesserung der Sacherschliessung bei unzureichend erschlossenen Titeldatensätzen
- Erweiterung der heuristischen Funktion, Verwendung weiterer (KI/IR)-Algorithmen, Lexikonerstellung
- Eliminierung weiterer irrelevanter Deskriptorwerte
- Andere Methode der (Klassenaggregation zur) Ergebnisausgabe
- Anreiz durch **Colibri/DDC-Wettbewerb** ☺

Literatur:
Information Retrieval & Bewertung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

[SALTON 1971] **The SMART Retrieval System – Experiments in Document Processing** (ed. Gerard Salton). Prentice-Hall, Englewood Cliffs, NJ, 1971.

[Jones 1981] Karen Spärck Jones: **Information Retrieval Experiment.** Butterworths, London, 1981.

[Jones 1996] Karen Spärck Jones; Julia R. Galliers: **Evaluating Natural Language Processing Systems. An Analysis and Review.** Lecture Notes in Artificial Intelligence 1083. Springer, Berlin, 1996.

[Voorhees/Harman 2005] **TREC: Experiment and Evaluation in Information Retrieval** (ed. by Ellen M. Voorhees; Donna K. Harman). MIT Press, Cambridge Massachusetts, 2005.

Literatur:
Information Retrieval & Bewertung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

[Moens 2000] Marie-Francine Moens: **Automatic Indexing and Abstracting of Document Texts.** Kluwer Academic Publishers, London, 2000.

[Manning/Raghavan/Schütze 2008] Christopher D. Manning; Prabhakar Raghavan; Hinrich Schütze: **Introduction to Information Retrieval.** Cambridge University Press, Juli 2008. Online: http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html.

[CLEF] **Cross-Language Evaluation Forum (CLEF)** . Online: http://www.clef-campaign.org/.

Literatur:
Automatische Klassifizierung & Bewertung (1)

Fundy-Nationalpark
ul, 25. Mai 2008

[Reiner 2008] Ulrike Reiner: **DDC-based Search in the Data of the German National Bibliography.** In: New Perspectives on Subject Indexing and Classification. Essays in Honour of Magda Heiner-Freiling. Deutsche Nationalbibliothek. Leipzig, Frankfurt am Main, Berlin, 2008, pp. 121-129.

[Reiner 2009] Ulrike Reiner: **Bewertung von automatisch DDC-klassifizierten Titeldatensätzen der Deutschen Nationalbibliothek (DNB).** VZG-Colibri-Bericht 1/2008. Online: http://taipan.dyndns.org/~ul/colibri05.pdf.

[Oberhauser 2004] Otto Oberhauser: **Automatisches Klassifizieren. Verfahren zur Erschliessung elektronischer Dokumente.** Master's Thesis. Zusatzstudiengang Bibliotheks- und Informationswissenschaft. Fakultät für Informations- und Kommunikationswissenschaften, Fachhochschule Köln, 2004.

Literatur:
Automatische Klassifizierung & Bewertung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

[Wille 2006] Jens Wille: **Automatisches Klassifizieren bibliographischer Beschreibungsdaten - Vorgehensweise und Ergebnisse**. Diplomarbeit. Studiengang Bibliothekswesen Fakultät für Informations- und Kommunikationswissenschaften, Fachhochschule Köln, 2006.

[Pfeffer 2008] Magnus Pfeffer: **Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen.** Vortrag: 97. Deutscher Bibliothekartag. 5. Juni 2008, Mannheim.

[Mehler/Waltinger 2009a] Alexander Mehler; Ulli Waltinger: **Automatic Enrichment of Metadata**. Vortrag: „9th International Bielefeld Conference". 4. Februar 2009, Bielefeld.

[Mehler/Waltinger 2009b] Alexander Mehler; Ulli Waltinger: **Enhancing Document Modeling by Means of Open Topic Models: Crossing the Frontier of Classification Schemes in Digital Libraries by Example of the DDC.** Wird publiziert in: Library Hi Tech, 2009.

DNB-Titeldatensatz zu Aronia

Fundy-Nationalpark
ul, 25. Mai 2008

001 991499077
002a20081128
003 20090303090118
004 20090310
025a991499077
026 DNB991499077
...
037bger
...
100 Neumayer, Petra
102a120295911
104aFunfack, Birgit
106a137378009
331 Aronia
335 Powerbiostoffe aus der Apfelbeere ; Gesundheit und Vitalität durch kraftvolle Antioxidanzien
359 Petra Neumayer ; Birgit Funfack
...
540aISBN 978-3-86728-084-6 kart. : EUR 7.95 (DE), EUR 8.20 (AT)
...
700 |610ÎDNB
705a□a615.32373□c615.32□d583.73□eDDC22ger
902s 7636533-5 Aronia
902s1 4288415-9 Naturheilmittel
902f11|Ratgeber
903 213

intellektuell vergebene DDC-Notation

Fundy-Nationalpark
ul, 25. Mai 2008

E N D E