HOCHSCHULE DER MEDIEN

Altbestandserschließung

Automatische Übernahme von RVK und SWD über Verbundgrenzen hinweg

Prof. Magnus Pfeffer
pfeffer@hdm-stuttgart.de

Überblick

HOCHSCHULE DER MEDIEN

- Ausgangslage
- Projekt
- Ergebnisse
- Bewertung
- Ausblick

Ausgangslage

Anlass

- Retroklassifikation Freihandbestand UB Mannheim
  - Seit 2001
  - 5 große Bibliotheksbereiche statt 11 kleine Bereichsbibliotheken
  - RVK als einheitliche Klassifikation
  - Wunsch nach mehr Fremddaten
    - 2004: Weniger als 50% der Titel mit RVK

Vorarbeiten

HOCHSCHULE DER MEDIEN

- Automatische Vergabe von RVK-Notationen
  - Projekt seit 2004
  - Ansatz: Vergabe aufgrund Ähnlichkeit zu bereits klassifizierten Titeln
  - Ergebnisse
    - Verfahren funktioniert grundsätzlich
    - Hoher Rechenaufwand
    - Liefert meist mehrere Klassen pro Titel
      - Darunter fast immer eine gute Klassifikation
      - Filtern oder Reduzierung der Klassen trifft auch die „guten“ Ergebnisse

→ Anteil „Rauschen“ zu groß für direkte Nutzung

Kenngrößen (Stand 2010)

HOCHSCHULE DER MEDIEN

- SWB
  - 12.777.191 Monografien
  - 3.979.796 (31,1%) mit SWD-Schlagwörtern
  - 3.235.958 (25,3%) mit RVK-Notationen
- HeBIS
  - 8.844.188 Monografien
  - 2.237.659 (25,3%) mit SWD-Schlagwörtern
  - 1.933.081 (21,8%) mit RVK-Notationen

Mehr Statistik

HOCHSCHULE DER MEDIEN

- Verteilung der Titel auf Jahre (SWB)

Jahr	Anzahl	SWD	RVK
sonstige	95.740	9.699	9.746
1000-1599	105.473	1.338	767
1600-1699	194.825	8.078	2.044
1700-1799	367.529	21.406	11.532
1800-1899	890.558	58.683	84.977
1900-1949	1.490.137	152.883	248.658
1950-1979	2.954.648	638.932	802.363
1980-1999	4.304.732	1.846.295	1.354.512
2000-	2.373.515	1.242.461	721.358

Mehr Statistik

HOCHSCHULE DER MEDIEN

- Verteilung der Titel auf Jahre (Hebis)

Jahr	Anzahl	SWD	RVK
sonstige	205.651	35.133	5.670
1000-1599	31.454	534	54
1600-1699	110.596	1.886	332
1700-1799	248.218	4.331	4.800
1800-1899	340.859	26.128	21.605
1900-1949	648.814	54.883	35.351
1950-1979	1.688.942	105.062	306.824
1980-1999	3.260.544	912.868	1.031.428
2000-	2.294.910	1.096.588	526.097

Projekt

Aktuelles Projekt

HOCHSCHULE DER MEDIEN

- Grundidee: Übernahme von SWD und RVK aus
  - Vorauflagen
  - Parallelausgaben
  - Übersetzungen

- Vorhanden aus Vorprojekt
  - Datenaufbereitung
  - Programme
  - Generierte Indexe

→ „fast gleiche“ Titel suchen

Daten

HOCHSCHULE DER MEDIEN

- Ausgangsdaten
  - Verbunddatenbank Südwestverbund und Hebis
  - MAB2 Format
  - ca. 14 GB

- Aufbereitung
  - Datenreduktion auf relevante Felder
  - Expansion von Titelinformation
    - Information von Gesamtaufnahme in Stücktitel
  - ca. 4,2 GB

Umsetzung

HOCHSCHULE DER MEDIEN

- Vergleich auf Basis von
  - Einheitssachtitel
    - Feld 304_
  - Titel und Untertitel
    - Felder 331_, 335_
  - Autoren und Urheber
    - Felder 100_, 104a, 108a, 200_, 204a, 208a
  - beteiligte Personen und Körperschaften
    - Felder 100b, 104b, 108b, 200b, 204b, 208b
- Ansatz:
  Identischer (Einheitsach-)Titel plus **eine** Übereinstimmung bei Person/Körperschaft = Match

Umsetzung

HOCHSCHULE DER MEDIEN

- Algorithmus
  - Berechne für alle Titel
    - Wenn Feld 304_ vorhanden
      - Suche Titel mit identischem Feld 304_
      - Vergleiche Autoren, Urheber und beteiligte
        - MATCH, wenn **eine** Übereinstimmung vorhanden
    - Sonst (nur Feld 331_ und 335_ vorhanden)
      - Suche Titel mit identischen Feldern 331_ und 335_
      - Vergleiche Autoren, Urheber und beteiligte
        - MATCH, wenn **eine** Übereinstimmung vorhanden
- Technische Umsetzung
  - Perl unter Linux
  - Indexstrukturen im Hauptspeicher (>4GB)

Ergebnisse

Ergebnisse: SWD

HOCHSCHULE DER MEDIEN

- 5.809.349 Titel mit mindestens einem Match
  - Davon
    - 3.269.340 ohne SWD
    - 3.627.017 ohne RVK
  - Anreicherung durch Übernahme möglich bei
    - 636.462 mit SWD
    - 959.419 mit RVK

Ergebnisse: Hebis

HOCHSCHULE DER MEDIEN

- 4.535.618 Titel mit mindestens einem Match
  - Davon
    - 3.068.968 ohne SWD
    - 3.071.022 ohne RVK
  - Anreicherung durch Übernahme möglich bei
    - 1.179.133 mit SWD
    - 992.046 mit RVK

Verteilung der neuen Daten

HOCHSCHULE DER MEDIEN

- Exemplarisch am SWB

Jahr	Anzahl Titel	SWD	RVK
sonstige	205.651	4.027	6.464
1000-1599	31.454	5.162	6.094
1600-1699	110.596	8.253	3.984
1700-1799	248.218	17.020	15.612
1800-1899	340.859	36.135	51.303
1900-1949	648.814	71.309	96.607
1950-1979	1.688.942	161.587	221.072
1980-1999	3.260.544	197.147	328.531
2000-	2.294.910	135.822	229.752

Bereitstellung

HOCHSCHULE DER MEDIEN

- Daten zum Download
  - Textformat, bz2-Archiv
  - Titel-ID und gefundene Matches
- Linked Open Data
  - RDF-Tripel der Form ID-equalsForClassification-ID
  - http://data.bib.uni-mannheim.de
    - Mehr dazu morgen früh :-)
- Daten an die Verbundzentralen
  - Titel und gefundene SWD-IDs und RVK-Notationen

Bewertung

Prüfung

HOCHSCHULE DER MEDIEN

- Online im Linked-Data Web
  - Verbünde erlaubten Titeldarstellung
  - Matches untereinander verlinkt
  - Wer: Externe Interessierte
- Testdatenbanken der Verbünde
  - Einspielung der gelieferten Daten in Auszügen
  - Stichproben und Recherchen möglich
  - Wer: Sacherschließer und interessierte Verbundnutzer

→ Hohe Qualität der Ergebnisse bestätigt

Mehr Quellen – mehr Daten?

HOCHSCHULE DER MEDIEN

- Beispiel Schlagwörter im SWB
  - 451.677 angereicherte Titel bei Daten nur aus SWB
  - 636.462 bei SWB plus Hebis
- Beispiel RVK UB Mannheim
  - Bibliotheksbereich A5, Sozialwissenschaften
  - 63.300 Titel zu bearbeiten
  - 44.991 Titel mit RVK-Notationen im SWB
  - 45.610 Titel mit Übernahme aus SWB und Hebis
  - 48.454 Titel mit Übernahme aus SWB, Hebis, BVB
    - (Nur experimentell; Suchen der Titel über den BVB-Verbundkatalog)

Mehr Quellen – mehr Daten?

HOCHSCHULE DER MEDIEN

- Beispiel Schlagwörter im SWB
  - 451.677 angereicherte Titel bei Daten nur aus SWB
  - 636.462 bei SWB plus Hebis
- Beispiel RVK UB Mannheim
  - Bibliotheksbereich A5, Sozialwissenschaften
  - 63.300 Titel zu bearbeiten
  - 44.991 Titel mit RVK-Notationen im SWB
  - 45.610 Titel mit Übernahme aus SWB und Hebis
  - 48.454 Titel mit Übernahme aus SWB, Hebis, BVB
    - (Nur experimentell; Suchen der Titel über den BVB-Verbundkatalog)

Ausblick

Realisierung

HOCHSCHULE DER MEDIEN

- Hebis
  - Daten im Testsystem geprüft
  - Einspielung ins Produktivsystem geplant/erfolgt
- SWB
  - Daten im Testsystem geprüft
  - Einspielung ins Produktivsystem läuft aktuell

Weitere Arbeiten

HOCHSCHULE DER MEDIEN

- Verbesserungen Algorithmus
  - Übersetzungen erkennen
  - Personen in unterschiedlicher Ansetzung erkennen
  - Transitive Hülle bilden
    - Wenn A = B und B = C, dann auch A = C
    - Wichtig, wenn z.B. nur A erschlossen
- Verbesserungen Datenmodell
  - Art der Verknüpfung explizit dokumentieren
    - Vorauflage
    - Parallelausgabe
    - Übersetzung
    - Verlagswechsel

Mehr Daten

HOCHSCHULE DER MEDIEN

- Weitere Verbünde
  - Nur Verbundabzug und Erlaubnis zur Weitergabe der Sacherschließungselemente erforderlich
    - Ideal: Open Data
- Erweiterung auf andere Erschließungsarten
  - Dewey Decimal Classification
  - LoC Classification
  - LoC Subject Headings
    - British National Bibliography ist als Open Data verfügbar
    - Erste Abgleichversuche gestartet

Fragen und Diskussion