HOCHSCHULE DER MEDIEN

Abgleich von Titeldaten zur Übernahme von Sacherschließungsinformationen über Verbundgrenzen

Prof. Magnus Pfeffer
Hochschule der Medien, Stuttgart
pfeffer@hdm-stuttgart.de

Überblick

- Ausgangslage
- Ansatz
- Erste Projektphase
- Zweite Projektphase
- 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

Inkonsistente Erschließung

HOCHSCHULE DER MEDIEN

- Aus einem deutschen Verbundkatalog
  - Herzfeld, Hans: Der erste Weltkrieg
    - 18 Titelsätze
    - davon 11 mit RSWK, 8 mit RVK
  - Friedell, Egon: Kulturgeschichte der Neuzeit
    - 31 Titelsätze
    - davon 21 mit RSWK, 17 mit RVK
  - Tanenbaum, Andrew S.: Computer Networks
    - 44 Titelsätze
    - davon 19 Deutsch, 15 Englisch, 1 Chinesisch
    - davon 38 mit RSWK, 31 mit RVK

Inkonsistente Erschließung

- Tanenbaum, Andrew S.: Computer Networks
  - RVK Notationen
    - ST 200: 31 Titel
      - Informatik-Monografien-Vernetzung, verteilte Systeme-Allgemeines, Netzmanagement
    - ST 205: 3 Titel
      - Informatik-Monografien-Vernetzung, verteilte Systeme-Internet allgemein
    - QH 500: 2 Titel
      - Wirtschaftswissenschaften-Mathematik. Statistik. Ökonometrie. Unternehmensforschung-Wirtschaftsinformatik. Datenverarbeitung
    - MS 7965: 1 Titel
      - Soziologie-Spezielle Soziologien-Soziologie der Massenkommunikation und öffentlichen Meinung, Mediensoziologie-Internet, neue Medien

Ansatz

Grundidee

- Übernahme von RSWK und RVK aus
  - Vor- und Folgeauflagen
  - Parallelausgaben
  - Übersetzungen

- Annahmen
  - Titelgleichheit über Auflagen und Ausgaben
  - Mindestens ein Autor/Herausgeber bleibt bei Neuauflage

Grobes Matching

HOCHSCHULE DER MEDIEN

- Ausgangsdaten: MAB2
  - Nur monografische Titel
- 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

Match bei
identischem (Einheitsach-)Titel
UND einer Übereinstimmung
bei Person/Körperschaft

Konsolidierung der Erschließung

HOCHSCHULE DER MEDIEN

- Clustering
  - Basis: Matching-Ergebnisse
  - Ergebnis: Inhaltlich konsistente Cluster
    - „Werksebene“
- Verarbeitung innerhalb der Cluster
  - Sammeln der Erschließungsinformationen
  - Verteilen auf alle Elemente des Clusters

Erste Projektphase

Projektzeitraum: 2010-2011

HOCHSCHULE DER MEDIEN

- Ausgangsdaten: Verbunddatenbanken
  - Katalog des Südwestdeutschen Bibliotheksverbundes (SWB)
    - 12.777.191 Monografien
    - 3.979.796 (31,1%) mit RSWK-Schlagwörtern
    - 3.235.958 (25,3%) mit RVK-Notationen
  - Katalog des Hessischen Bibliotheks- und Informationssystems (HeBIS)
    - 8.844.188 Monografien
    - 2.237.659 (25,3%) mit RSWK-Schlagwörtern
    - 1.933.081 (21,8%) mit RVK-Notationen

Umsetzung

- 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 / Linux
  - Indexstrukturen im Hauptspeicher (>4GB)

Ergebnisse: SWB

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

Ergebnisse: Hebis

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

Bereitstellung der Ergebnisse

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
- Daten an die Verbundzentralen
  - Titel und gefundene SWD-IDs und RVK-Notationen

Evaluation

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

Nutzung

HOCHSCHULE DER MEDIEN

- 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)

Zweite Projektphase

Projektzeitraum: 2012

HOCHSCHULE DER MEDIEN

- Aggregation möglichst vieler Fremddaten
- Daten
  - SWB
    - Katalog des Südwestdeutschen Bibliotheksverbundes
  - Hebis
    - Katalog des Hessischen Bibliotheks- und Informationssystems
  - HBZ
    - Katalog des Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen
  - B3Kat
    - Gemeinsamer Verbundkatalog von Bibliotheksverbund Bayern und dem Kooperativen Bibliotheksverbund Berlin-Brandenburg

Ergebnisse

HOCHSCHULE DER MEDIEN

Katalog	Monografien	Anteil RVK	Anteil RSWK	Zuwachs RVK	Zuwachs RSWK
SWB	13.330.743	4.217.226	4.083.113	581.780	957.275
Hebis	8.844.188	1.933.081	2.237.659	1.097.992	1.308.581
HBZ	13.271.840	1.018.298	3.322.100	2.272.558	1.080.162
B3Kat	22.685.738	5.750.295	6.055.164	1.097.992	1.308.581

Ausblick

Mehr...

HOCHSCHULE DER MEDIEN

- Mehr Titeldaten
  - Gemeinsamer Bibliotheksverbund (GBV)
  - Deutsche Nationalbibliothek (DNB)
  - Schweizer und Österreichische Katalogdaten
  - Open Data aus anderen (Verbund-)Katalogen
- Mehr Klassifikationssysteme
  - LCC
  - LCSH
  - DDC / UDC

… und nachhaltiger

- Probleme
  - Eigenentwicklung ist weder wartbar noch portabel
  - Datenmengen wachsen rapide
    - >100 Mio. Titeldatensätze als Open Data verfügbar
  - Vielzahl von Statistiken / Clusteringmethoden für unterschiedlichste Anwendungen

→ Aufbau einer offenen Infrastruktur für die Analyse von Metadaten

Culturegraph.org

HOCHSCHULE DER MEDIEN

- Initiative von DNB und HBZ
  - Ziel: Zusammenführen von bibliografischen Informationen, die als Linked Open Data zur Verfügung stehen
- Open Source Infrastruktur
  - Parametrisierbare Metadatenverarbeitung
  - Erweiterbar (Java)
  - Skalierbar (Hadoop)
  - http://sourceforge.net/projects/culturegraph/

Weitere Anwendungen

HOCHSCHULE DER MEDIEN

- Konkordanzen zwischen Erschließungssystemen
- Analyse der Nutzung von Erschließungssystemen
  - Statistiken
  - Doppelstellen / Unscharfe Klassentrennung
- Verknüpfungen und Anreicherungen
  - Explizite Beziehungen zwischen Titeln
  - Nicht-bibliografische Ontologien
    - z.B. Ortsnamen

→ Ihre Ideen sind gefragt!

Danke für Ihre Aufmerksamkeit!

Folien online unter
http://www.slideshare.net/MagnusPfeffer/