HOCHSCHULE DER MEDIEN

Die Anreicherung der B3Kat-Echtdatenbank mit RVK-Notationen

Prof. Magnus Pfeffer
pfeffer@hdm-stuttgart.de

Übersicht

- Grundidee
- Experimentelle Umsetzung und Projekt B3Kat
- Aktueller Stand
- Ausblick

Grundidee

Beobachtung

HOCHSCHULE DER MEDIEN

- Zahlreiche Werke tauchen mehrfach im Katalog auf
  - Unterschiedliche Ausgaben
    - Format und Bindung
    - Verlag
    - Übersetzungen
  - Mehrere Auflagen

- Durch die verteilte Arbeit in den Verbünden werden die Mehrfachaufnahmen nicht immer einheitlich erschlossen

Beispiele

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

Beispiele

- Tanenbaum, Andrew S.: Computer Networks
  - 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

Fragestellung

HOCHSCHULE DER MEDIEN

- Idee: Schließen der „Lücken“ in der Erschließung durch Übernahme aus
  - Vorauflagen
  - Parallelausgaben
  - Übersetzungen

- Aber:
  - Lohnt sich das überhaupt?
  - Sind es nicht nur Einzelfälle?

Experimentelle Umsetzung

Experiment

- Vergleich von Monografien
  - 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

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

Experiment: Algorithmus

HOCHSCHULE DER MEDIEN

- Berechne für alle monografischen 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

Experiment: Umsetzung

HOCHSCHULE DER MEDIEN

- Programme
  - Einfache Perlskripte unter Linux
  - Selbst entwickelte Indexstrukturen
  - Berechnung auf einem Einzelrechner
- Daten
  - MAB2-Vollabzug Südwestverbund
  - MAB2-Vollabzug Hebis

Ausgangsdaten

HOCHSCHULE DER MEDIEN

- 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

Verarbeitung

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

Ergebnis: SWB

- 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

Ergebnis: Hebis

- 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

Datenbereitstellung

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

Datenprüfung

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

Stand Anfang 2011

- SWB und Hebis
  - Projekt abgeschlossen
  - Ergebnisse wurden in Produktivdatenbank eingespielt

Zweite Runde

HOCHSCHULE DER MEDIEN

- Verwendung von Abzügen von vier Verbünden
- 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

- Anreicherung
  - RVK neu: 2.969.381 Einträge
  - RSWK neu: 2.765.967 Einträge
- Übermittlung Ende 2012
- Einspielung 7.2.2014

Aktueller Stand

Benannte Probleme

HOCHSCHULE DER MEDIEN

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

→ Bedarf einer offenen Infrastruktur für die Analyse und Verarbeitung großer Mengen 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 „Metafacture“
  - Parametrisierbare Metadatenverarbeitung
  - Erweiterbar (Java)
  - Skalierbar (Hadoop)

Neue Probleme

HOCHSCHULE DER MEDIEN

- Verteiltes Rechnen benötigt eine neue Herangehensweise
  - Datenspeicherung und -addressierung
  - Lösen von Datenabhängigkeiten
  - Aufteilung der Berechnungsschritte

- Komplexität
  - Hadoop ist unglaublich mächtig, aber sperrig
  - Metafacture ist einfach zugänglich, aber limitiert

Zwischenschritt

HOCHSCHULE DER MEDIEN

- Nutzung einer dokument-orientierten Datenbank
  - Ansatz: Schlüsselbasierter Zugriff auf semistrukturierte Daten (Dokumente)
  - Speicherung der Daten als Attribut-Wert-Paare
    - Werte können auch Listen oder ganze Dokumente sein
  - Passt gut auf bibliothekarische Daten
    - Schlüssel: Identnummern
    - Attribut: Feldnummer / Feldbezeichnung
    - Wert: Feldinhalt, bei Wiederholung als Liste
  - Neue Felder können ohne Schemaänderung zugefügt werden

Zwischenschritt

HOCHSCHULE DER MEDIEN

- Vorteile
  - Mehrere Open-Source Lösungen vorhanden
  - Einfache Nachnutzung und Erweiterung
    - Programme übersichtlicher und verständlicher
      - Nur noch Datenverarbeitung, keine Speicherung
    - Schnittstelle zur Datenbank gut dokumentiert

    → Änderungen schnell umsetzbar
  - Höhere Geschwindigkeit
  - Datenhaltung sehr ähnlich zu Hadoop

    → Portierung nach Metafacture einfacher möglich

Stand Herbst 2014

HOCHSCHULE DER MEDIEN

- Neuimplementierung mit Software „MongoDB“
- Nutzung der Amazon Web Services
  - Virtuelle Server mit großem Hauptspeicher
  - Keine Kosten dank Förderprogramm für Wissenschaftler
- Nutzung von Open Data
  - SWB Marc21
  - B3Kat Marc21
  - HBZ Mab2XML

Aktuelle Projekte

HOCHSCHULE DER MEDIEN

- Anreicherung der Daten von IDS/Basel mit
  - RVK
  - RSWK
  - MeSH
- Besonderheiten
  - Nutzung der neuen Infrastruktur
  - Marc21-Anreicherungs-Komponente
  - Komplexe Verarbeitung der RSWK-Schlagwortfolgen
    - Uneinheitliche Verarbeitung der Altdaten in Marc21

Aktuelle Projekte

HOCHSCHULE DER MEDIEN

- Automatisch generierte Konkordanzen zwischen inhaltlichen Erschließungssystemen
- Ansatz
  - Clustering der Daten
  - Konsolidierung der Erschließung innerhalb der Cluster
  - Auszählen des gemeinsamen Auftretens von Erschließung
  - Filtern und Gewichten dieser Zahlen
  - Vergleich mit vorhandenen (Teil-)Konkordanzen und Bewertung
- Aktuell laufende Umsetzung
  - RVK → Basisklassifikation
  - Partner: ÖNB

Aktuelle Projekte

- Homogene Daten innerhalb der Cluster
- Ansatz
  - Erweiterung des Ansatzes auf Formalerschließung
  - Hier: Verknüpfung mit individualisierten GND-Einträgen
- Laufende Umsetzung
  - Daten grundsätzlich vorhanden
  - Auswertung durch studentische Abschlussarbeit
    - Bislang keine/n Freiwillige/n gefunden :)

Fragen