Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Aus der Programmierungsküche eines bibliothekarischen Nichtprogrammierers

Mechanismen zur Gewinnung, Filterung und Veredelung neuer Fachmetadaten durch Web- und Schnittstellenharvesting, maschinelle Sprachbestimmung und Übersetzungsverfahren sowie automatische Volltextindizierung

45. Internationale Arbeits- und Fortbildungstagung der der ABDOS e.V.

Herne, den 3. Mai 2016

Überblick

- Worum geht es wirklich?
  - → Nutzer und ihr Suchverhalten
- Fachorientierte Metadaten
  - → Gewinnung und Filterung
- Metadatenveredelung und -anreicherung
  - → Automatische Sprachbestimmung
  - → Maschinelle Übersetzung
  - → OCR-Mechanismen und Volltextindizierung
- Fazit

Worum geht es hier wirklich?

- Parallele Suche des Slavistik-Portals (als fertiges Suchsystem)
- Benutzer und ihr Suche-Verhalten
  - → Benutzer als hohes Gut

- **Fazit**: Das Suchsystem muss besser werden
  - → Problem: IT-Spezialisten haben keine Zeit oder sie haben keine speziellen Fachkenntnisse
- Antwort: „Do it yourself“
  - → Werkzeuge: weit verbreitete Skriptsprache, leichtfüßige Datenbank, Windows-Rechner
  - → Dadurch: größere Flexibilität bei der (Um)Strukturierung der Daten (Beispiel: skriptsprachenbasierter XML-Parser vs. Harvester-Software) und Effizienz (viele Daten auf einmal)

Beispiele für Fachmetadaten I

- Sammlung elektronischer Bücher vom „Інститут історії України Національної академії наук України“, ca. 3800 BE, (JSON)
- BRDRL = Bibliografija rabot po drevnerusskoj literature, Bde 1-7, 1917-2002, 24.925 BE, bereitgestellt durch Puškinskij Dom (HTML) ≥
- Руниверс-ToCs (XHML): Чтения в ИОИДР, 1846-1908, ca. 3550 BE; Архив Юго-Западной России, 34 Bde (1500 BE) u.a.
- Starieknigi.info, 15.200 BE (HTML) ≥
- Archive.org, 14.700 BE (SOLR, Indiziert mit Fuzzy-Faktor)
- Federacja Bibliotek Cyfrowych (OAI-MHP): Sammlungen aus JBC, eBUW, MBC, WBC, Polona (BN), SBC u.a. werden fachlich und sprachlich gefiltert. (Auch spezielle Sammlungen wie die Zss. Pamietnik Literacki, Jezyk Polski sind möglich)

Fachorientierte Metadaten

- Warum „Fach"-Daten? >
- Strukturierte vs. unstrukturierte (Meta)Daten
  - → Daten im freien Zugang
  - → Lokale Speicherung -> Durchlaufen von Daten -> Mustererkennung -> (DB-Zwischenlagerung) -> XML-Erstellung
  - → Strukturierung anhand von Mustern, nicht von konkretem Material, da die Quelle sich jederzeit verändern kann
- Gewinnung
  - → Schnittstellen (Z39.50, SRU, SOLR, OpenSearch)
  - → Harvesting (OAI, UnAPI, Z39.88-2004, XML)
  - → Crawling (XHTML, HTML)
- Filterung
  - → Anhand von Sprach- und Erschließungsdaten
  - → XML-Filterung via Streaming-Parser

Metadatenveredelung und -anreicherung

- Automatische Sprachbestimmung
  - → Sprachbestimmung anhand von Trigramm-Algorithmus >
  - → Retransliterierung auf der Grundlage der Sprachbestimmung
- Maschinelle Übersetzung
  - → Einsatz von „Яндекс.XML“ (80.000 Anfragen pro Tag) >
  - → Caching und Wiederverwendung von Schnittstellendaten
- OCR-Mechanismen und Volltextindizierung
  - → Tesseract-OCR (für slawische Sprachen, inkl. Frakturschrift)
  - → Tika-Textextraktion (z.B. Kievskaja starina, 3262 BE, bereitgestellt von Universität Kiev)
- Anreicherung der ISSNs-Angaben bei Zeitschriften

Beispiele für Fachmetadaten

- ToCs: Izvestija Otdelenija russkogo jazyka = IORJaS, über 100 Bde, 1852-1998, ca. 5000 BE, bereitgestellt von Feb-web.ru (HTML)
- Bazhum - czasopisma humanistyczne i społeczne w internecie, 193.311 DS, (XML, UnAPI-JSON)
- HathiTrust, Metadaten der Volltexte mit Einschränkung auf den US-amerikanischen Raum, 40.000 BE (OAI, hier auch Retransliterierung) >
- ToCs + Volltext: Kievskaja starina, 1882-1907, 291 Hefte, 3.262 DS (HTML, Tika-Textextract der ersten 3 Seiten)
- Cyberleninka, ca. 50.000 BE, Fachausschnitt: Sprache, Literatur, Kultur, Geschichte, Volkskunde (OAI, durch Base-search.net nicht vollständig indiziert)

Fazit

- Manuell gesteuertes Harversting und Crawling bringen Struktur in die Daten hinein, die sonst für die Benutzer*innen überwiegend unübersichtlich und nur über „Google“ zugänglich wären.
- Durch Verbesserung der Qualität der Metadaten wird unsere Suche auf die Bedürfnisse der Nutzer noch mehr abgestimmt.
- Das lokale Vorhalten der qualitativ hochwertigen Daten reduziert die Umweltbelastung durch Server, Rechner und Traffic-Wärme, macht die Suche effizienter und das Suche-Erlebnis positiv erlebbar.
- >

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Vielen Dank für Ihre Aufmerksamkeit!

Dr. Vladimir Neumann

Osteuropa-Abteilung, SBB

Vladimir.Neumann@sbb.spk-berlin.de

Tel. 030 – 266 435640