Projektskizze Semesterprojekt, CAS BGD, FS 2020

Titel: Volltextsuchmaschine für digitalisierte Bernensia

1. Umfeld / Ausgangslage

DigiBern ist eine Plattform für das kulturelle Erbe des Kantons Bern („Bernensia"), die von der Universitätsbibliothek Bern betrieben wird. Auf der Basis der Web Content Management Systems (WCMS) Drupal werden digitalisierte Dokumente und spezifischere Online-Angebote in einem Portal zusammengefasst und mittels grober Kategorien in Form eines Webverzeichnisses erschlossen (Epochen, Regionen und Orte, Personen, Themen).

2. Problemstellung

Insbesondere für die digitalisierten Dokumente existiert keine übergreifende Volltextsuche, da diese Dokumente zum einen auf den Plattformen e-rara (Beispiele) und e-periodica (Beispiel) liegen, zum anderen als PDF-Dateien einfach in das WCMS eingehängt sind. Für die Dokumente auf e-rara sind Volltexte zum Teil als TXT-Dateien verfügbar, zum Teil nur als PDF. Die Dokumente auf e-periodica sind (auf Artikel-Ebene) als PDF-Dateien verfügbar. Beide Plattformen liefern zusätzlich die bibliografischen Metadaten per OAI-PMH-Schnittstelle in XML-basierten Standards aus.

3. Lösungsansatz

Zielsetzung ist der Proof-of-Concept eines Such-Interfaces für die DigiBern-Plattform, das eine Volltext-Suche über die verteilt vorhandenen Digitalisate ermöglicht. Hierzu müssen Metadaten und Volltexte der Plattformen geharvestet werden und in einen gemeinsamen Such-Server geladen werden. ElasticSearch bzw. der ELK Stack bietet sich hier als Lösung an. Per Logstash können verschiedenartige Daten transformatiert und in ElasticSearch geladen werden, ElasticSearch selbst verfügt über einen JSON-basierten Document Store, der sich für die Verwaltung von (nicht übermässig grossen) Textdaten eignet. Mit ElasticSearch können automatisiert (da „schemaless") und manuell per Mapping die für eine performante Suche erforderlichen invertierten Indizes erstellt werden. ElastisSearch kann für die Suche über eine API angesprochen werden und die Suchergebnisse damit einfach in bestehende Webseiten eingebunden werden.

Der ELK-Stack ist über verschiedene Cloud-Computing-Anbieter als Software-as-a-Service (SaaS) verfügbar, etwa bei AWS oder Elastic Cloud. In der Projektarbeit soll auf einer solchen Basis eine Volltextsuchmaschine über die verteilten Bernensia-Digitalisate als Machbarkeitsstudie implementiert werden, d.h. bis zur Ansprechbarkeit der Suchmaschine per API.

4. Personen

Studierender: Woitas, Kathi, Tel. + 41 78 899 4301, kathi.woitas@students.bfh.ch

Ansprechpartner/ Betreuer in der Firma: Prudlo, Marion, Tel. Tel. +41 31 684 95 94, marion.prudlo@unibe.ch