Panel 3a

Figure 1: AVefi_Logo

- Anforderungen
- Komponenten
- Indizierung
- Match und Merge

Anforderungen

- Bestände mit PIDs versehen.
- Exemplare, Manifestationen und Werke als eigenständige Einheiten erfassen.
- Werke sollen Datengeberübergreifend einheitliche PIDs erhalten.
- Exemplare sollen über Filterkriterien auf allen drei Ebenen recherchierbar sein.
- Datenzugriff per Webinterface und API.

Systemkomponenten

- PID-System + Typeregistry zur Datenvalidierung.
- Suchindex (Elasticsearch) als Plattform für die Recherche und den Abgleich von Werksdatensätzen.
- Spiegelungsdienst zwischen PID-System und Suchindex.
- AVefi Service Ebene („unser Backend“) als one-stop-shop für unser Frontend und bald auch API-gestützte Zugriffe von außen.
- Single-Sign-On-Authentifizierung über Academiccloud.

Mehrere Suchindizes für verschiedene Anforderungen

- Primärindex erhält Echtzeitupdates vom PID-System:
  - Zu jedem PID gibt es genau ein indiziertes Dokument.
  - Eignet sich besonders für Werksabgleich, da neu registrierte PIDs sofort für Abfragen bereitstehen.

Denormalisierter Index für facettierte Suche

- Ein Werk wird mit den zugehörigen Manifestationen und Exemplaren als ein einziges, geschachteltes Dokument indiziert.
- Bestimmte Informationen werden in besser indizierbarer Form dupliziert.
  - Beispiel: `has_date=1950~/1952` wird ergänzt durch `production_in_year>=1949<=1952`.
- Änderungen an den PIDs erreichen diesen Index nur zeitverzögert.
- Ohne die Datendenormalisierung wäre die facettierte Suche wegen vieler Hintergrundabfragen weniger performant.

Beispielabfrage

- Wie viele Fassungen hat welcher Datengeber zum Suchbegriff „Sommer“, gefiltert nach Format 35mm und Schwarzweiß?
- Die Filterung erfolgt für:
  - Suchbegriff auf Werksebene,
  - Format und Farbe auf Exemplarebene,
  - Datengeber auf Manifestationsebene.

Werksabgleich (Matching)

- Der Abgleich erfolgt anhand der vier Kriterien: Titel, Land, Jahr, Regie.
- „Deutschland“, „Bundesrepublik Deutschland“ und „BRD“ werden aktuell nicht als Match erkannt.
- Identifier aus Normdateien helfen bei Land und Regie, weitere Hilfsmittel denkbar.
- Titel scheint besonders herausfordernd zu sein.

Titel die Erfassung

- Elasticsearch bietet erprobte und konfigurierbare Bordmittel zur Textanalyse.
- Leichte Abweichungen wie Groß-/Kleinschreibung und eine konfigurierbare „Unschärfe“ werden toleriert.
- Beispiel: “TUNGUSKA - DIE KISTEN SIND DA” passt zu “Tunguska - die Kisten sind da”.
- Es kann aber sein, dass die vier Match-Kriterien nicht ausreichen.

Beispiele aktueller False Positives

- “Das alte Lied”:
  - Trailer ist nach den Kriterien nicht vom Hauptwerk unterscheidbar.
- “Die Wespe”:
  - Serienwerk matched gegen seine Episoden.

Bekannte False Negatives

- “Solo Sunny”
- “Menschen am Sonntag – Das Dokument der Gegenwart”:
  - Abgleich funktioniert nur in eine Richtung (kurzer Titel gegen langen Titel).