

Abschlussbericht: Effizientes Retrieval auf Web of Science-Daten mit Elasticsearch

Zeljko Carevic¹ and Philipp Mayr¹

¹*GESIS — Leibniz-Institut für Sozialwissenschaften*

February 22, 2019

1 Einleitung

Das Projekt "Effizientes Retrieval auf Web of Science-Daten mit Elasticsearch" hat das Ziel die umfangreichen XML-Daten des Web of Science (WoS), die in einer SQL-Datenbank vorliegen, in einen Elasticsearch Index zu überführen. Auf diese Weise sollen diese Daten effizienter recherchierbar und leichter zugänglich gemacht werden. Im Rahmen dieses Projekts wurden insgesamt 12.319.703 Millionen Einträge aus dem WoS in Elasticsearch indexiert. Das Projekt teilt sich in folgende Phasen auf:

- Installation und Konfiguration des Suchservers bei FIZ Karlsruhe: 1,25 PM.
- Transformation der WoS-Daten: 0,25 PM.
- Installation und Konfiguration des Webfrontends: 1,5 PM.

Der ursprünglich definierte Projektplan sah eine Installation und Konfiguration des Suchservers bei GESIS vor. In der letzten Phase sollte dann auf den Server des FIZ Karlsruhe migriert werden. Dies wurde jedoch zu Projektbeginn gemeinsam mit dem FIZ Karlsruhe diskutiert und verworfen. Gemeinsam wurde beschlossen die gesamte Installation und Konfiguration auf dem Skriptserver des FIZ Karlsruhe durchzuführen. Auf diese Weise entfällt der letzte Meilenstein. Die verbleibenden Projektphasen werden nun im Detail erläutert.

2 Installation und Konfiguration des Suchservers

Das FIZ Karlsruhe hat für die Installation und Konfiguration von Elasticsearch den Skriptserver bereitgestellt der via VPN in einer virtuellen Maschine (No-Machine) verwendet werden konnte.

Auf dem Scriptserver des FIZ Karlsruhe wurde entsprechend Elasticsearch in der Version 6.5 installiert. Bezüglich der Konfiguration des Index wurden keine Änderungen an der Standardkonfiguration vorgenommen. Eine Anpassung der Konfiguration insbesondere bei erhöhter Last kann jedoch jederzeit vorgenommen werden.

3 Transformation der Daten

Für eine Verarbeitung der WoS-Daten in Elasticsearch müssen die aktuell in einer SQL-Datenbank vorliegenden Daten nach JSON konvertiert werden. Zu diesem Zweck sollte ein Migrationsskript entwickelt werden, das die Daten nach JSON konvertiert. Ursprünglich war für die Transformation der Daten ein verhältnismäßig geringer Aufwand veranschlagt (0,25 PM). Die Transformation der Daten stellte sich jedoch als wesentlich aufwändiger dar. Zu Projektbeginn wurde GESIS vom FIZ Karlsruhe ein Zugriff auf die Oracle-Datenbank bereitgestellt. Da sich das Erstellen der SQL-Queries zum Export der entsprechenden Records (Export-Queries) als zu aufwändig erwies, wurde beschlossen, die Transformation der Daten auf Basis der Original XML-Rohdaten durchzuführen.

3.1 Script für die Transformation

Zur Transformation der Daten wurde ein Python-Skript entwickelt. Da bei GESIS bereits in einem früheren Projekt XML-Daten des WoS verarbeitet wurden, konnte bei der Entwicklung auf bestehenden Lösungen aufgesetzt werden. Das Migrationsskript wurde den Anforderungen entsprechend angepasst. Das Migrationsskript liegt für alle Kooperationspartner frei zugänglich auf dem Script-Server in dem Verzeichnis `/data/elasticsearch/prog/migrate.py`.

Für die aktuelle Projektphase wurde eine reduzierte Menge der Metadaten in die Transformation aufgenommen. Hintergrund für diese Entscheidung ist, dass der Konvertierungsprozess eine nicht zu vernachlässigende Laufzeit umfasst und dies innerhalb der Entwicklungsphase einen unnötigen Mehraufwand bedeutet.

Die Liste der aktuell enthaltenen Metadaten ist wie folgt:

- Title
- Abstract
- Authors and their affiliations
- References
- Funding text
- Publication info (year, journal, keywords)

Bei der Entwicklung des Migrationsskripts wurde auf die Erweiterbarkeit der Metadaten geachtet. Dadurch lassen sich die zu konvertierenden Inhalte ohne großen Aufwand erweitern.

Aktuell umfassen die transformierten Daten 12.319.703 WoS-Records. Die Konvertierungszeit für diese Datenmenge beträgt rund 24 Stunden. Es ist daher ratsam sich innerhalb des Konsortiums auf die relevanten Metadaten zu einigen um auf diese Weise unnötige Transformationsprozesse zu vermeiden.

3.2 Indexierung der Daten in Elasticsearch

Für die Indexierung der Daten in Elasticsearch wurde ebenfalls ein Script entwickelt. Das Script übermittelt die nach JSON konvertierten Records via CURL an den Suchindex. Das Skript findet sich ebenfalls auf dem Skriptserver unter /data/elasticsearch/prog/index.sh. Die Indexierung der Daten läuft verhältnismäßig schnell (circa 1 Stunde für 12 Millionen Records).

4 Kibana Frontend

4.1 Installation und Konfiguration des Webfrontends

Um die in Elasticsearch vorhandenen Daten nutzbar zu machen, war für diese Projektphase die Verwendung eines Elasticsearch Frontends angedacht. Die zu Projektbeginn angedachte Lösung über DeJaVu erwies sich jedoch als nicht praktikabel, da bereits die geringe Menge von 12 Millionen Records zu einer Überlastung der Komponente geführt hat. Aus diesem Grund wurde als Frontend das von Elasticsearch vertriebene *Kibana*¹ verwendet. Kibana als Frontend erlaubt es, eigene Suchanfrage an den Index zu stellen, die Daten zu visualisieren und als JSON zu exportieren.

4.2 Zugriff auf den Index

Für einen Zugriff auf das Frontend wird eine VPN-Verbindung zum Scriptserver via NoMachine benötigt. In Abbildung 1 ist der Startbildschirm von Kibana dargestellt. Die Verbindung wird dabei über einen Browser zu <http://localhost:5601/app/kibana> hergestellt.

4.3 Abfragen ausführen und speichern

Für eine Suchanfrage an den Index wird in Kibana der Eintrag *Discover* aus dem Menü aufgerufen. In Abbildung 2 ist eine Beispielanfrage dargestellt.

Die Query in dem dargestellten Beispiel ist nachfolgend separat aufgelistet.

model and year:(2016 TO 2018) and NOT (docType:"Editorial Material or docType:"Meeting Abstract")

In dem Beispiel erfolgt eine Suche nach Records, die den Begriff "model" in einem beliebigen Felder enthalten, zwischen 2016 und 2018 veröffentlicht wurden und deren Dokumenttyp weder "Editorial Material" noch "Meeting Abstract"

¹<https://www.elastic.co/products/kibana>

Figure 1: Kibana Startbildschirm

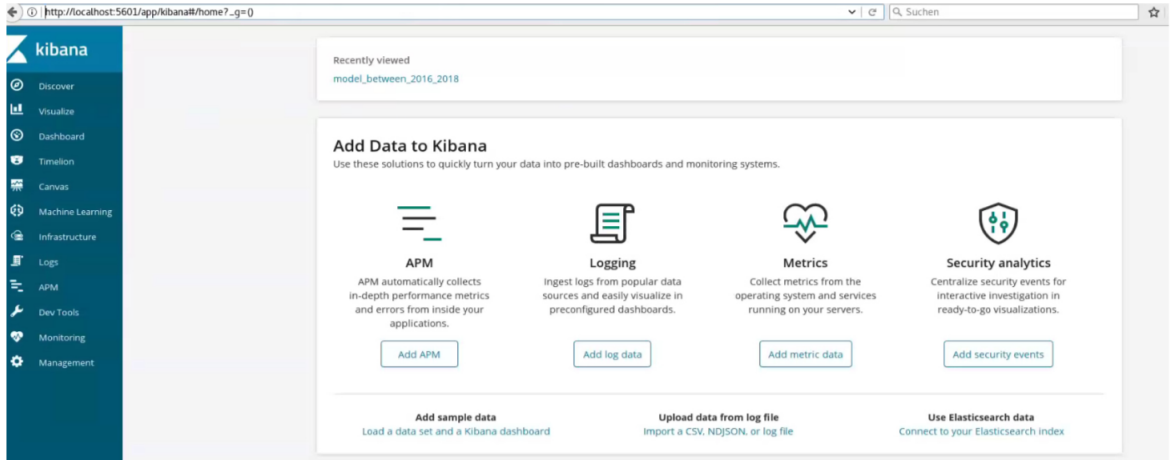


Figure 2: Abfrage an den Index

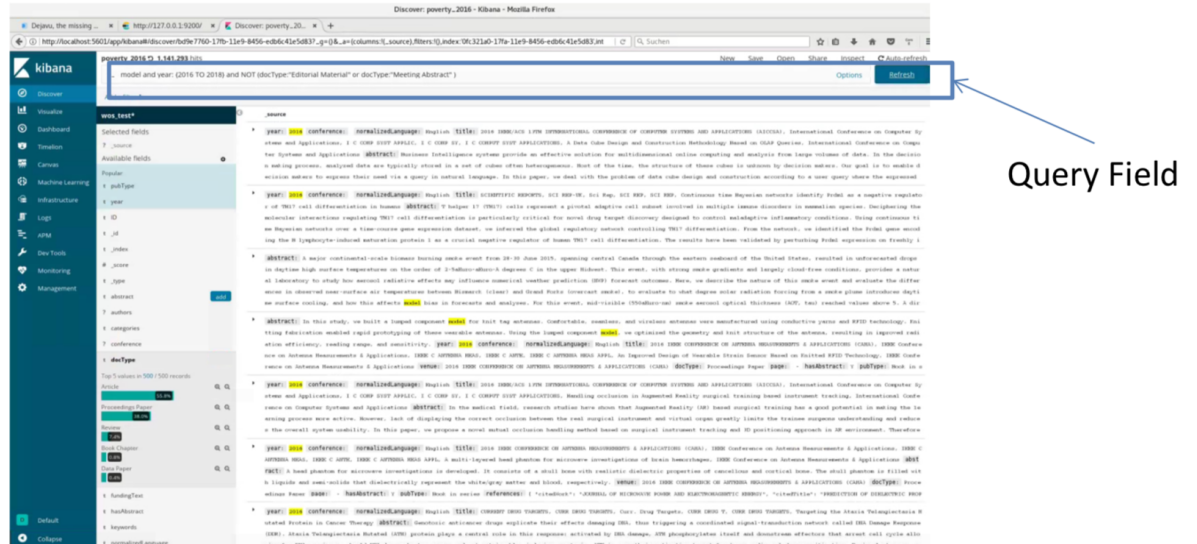
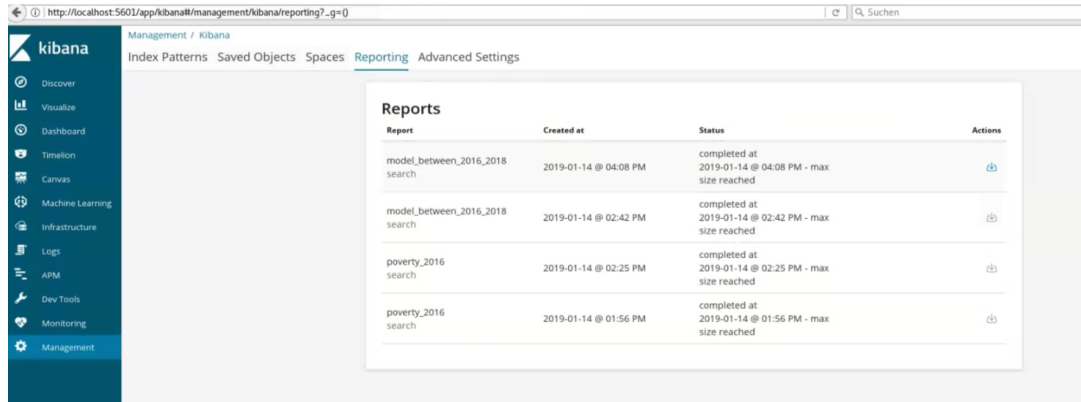


Figure 3: Export von Records



ist. Eine Abfrage kann innerhalb von Kibana gespeichert und wiederverwendet werden.

4.4 Export von Records

Kibana erlaubt den Export von Records in Form von JSON. Um einen Export zu starten muss die Abfrage gespeichert werden (siehe letzter Abschnitt). Der Export wird über den Menüeintrag "Share" aufgerufen. Da ein Export einige Zeit in Anspruch nehmen kann, werden die Inhalte im "Hintergrund" generiert und lassen sich über den Menüeintrag *Management* → *Reporting* aufrufen. Dies ist in Abbildung 3 dargestellt.

Aktuell erlaubt die Konfiguration von Kibana eine maximale Größe von 10MB pro Export. Diese Limitierung kann jedoch ohne Weiteres erhöht werden was sich jedoch auf die Performanz bzw. Exportdauer auswirkt.

5 Ausblick

In diesem Projekt wurde ein exemplarischer Abzug der WoS Rohdaten in einen Elasticsearch Index überführt. Für den Zugriff auf den erstellten Suchindex wurde Kibana als Frontend bereitgestellt.

Aktuell umfasst der Index 12 Millionen WoS-Records was nur einen geringen Anteil der WoS Daten ausmacht. Dies kann nach einer Evaluation des Systems jedoch ohne großen Aufwand erweitert werden. Das selbe gilt auch für den Umfang an indextierten Metadaten. Aktuell wurde nur ein Subset an Metadaten konvertiert. Da die Laufzeit der Konvertierung relativ hoch ist, sollten sich die Partner auf eine Liste von relevanten Metadaten festlegen, die dann in den Index überführt werden.