Themenstellung

Datum: 15.10.2023

für ein/eine

- (●) SWE-Projekt (ca. 5 Studierende, 10-12 Wochen Arbeitszeit)
- ( ) Bachelor-Arbeit (1-2 Studierende, 6-9 Wochen + Einarbeitungszeit)
- ( ) Master-Arbeit (1 Studierender, ca. 22 Wochen + Einarbeitungszeit)

Betreuende Firma/Institut
(Name/Anschrift/Telefon)

JSC
Forschungszentrum Jülich

Ansprechpartner/Betreuer
(bei SWE-Projekten auch 'Auftraggeber')

Name: Sabine Schröder

Telefon: +49 2461 616937

Email: s.schroeder@fz-juelich.de

Studierende, die das Projekt/das Thema auf jeden Fall bearbeiten sollen:

Titel der Themenstellung

Erweiterte Such- und Filterfunktionen für eine Terabyte-Datenbank zur globalen Luftverschmutzung

Inhaltliche Beschreibung

Das JSC betreibt mit der PostgreSQL-Datenbank des Tropospheric Ozone Assessment Reports (TOAR) eine der weltweit größten Sammlungen globaler Luftqualitätsdaten auf der Terabyte-Skala. Über 400.000 Zeitreihen an über 20.000 Messstationen sind in der TOAR-Datenbank mit vielen Metadaten gespeichert, und die Suche nach bestimmten Stationen und Zeitreihen erfordert ein ausgefeiltes Konzept. Die Aufgabe dieses Software-Projektes ist es, gezielte Verbesserungen der Suchfunktionen der TOAR-API (siehe https://toar-data.fz-juelich.de) zu entwickeln und zu testen. Insbesondere soll eine Ähnlichkeitssuche für textbasierte Informationen (z.B. Stationsnamen) eingeführt werden, und es soll eine flexiblere Verknüpfung von Abfragen ermöglicht werden, d.h. Einführung von logischen Operatoren (AND, OR, NOT).

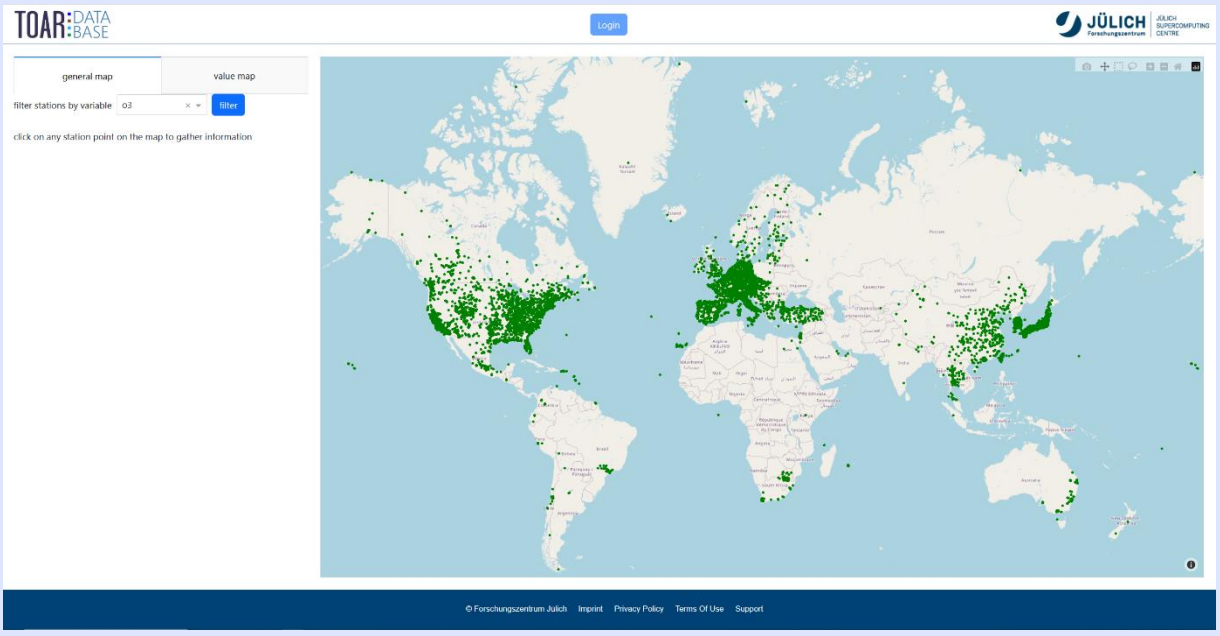

Abbildung 1: Webbasierte GUI der TOAR-Datenbank; die verbesserte Suchfunktion soll auch hier über weitere Filter verfügbar gemacht werden.

Themenstellung

Datum:

Inhaltliche Beschreibung – Fortsetzung

Spezifische Aufgaben des Projektes sind:
1. Erstellen einer Testdatenbank basierend auf der operationellen TOAR DB (Software und Arbeitsabläufe hierfür sind verfügbar)
2. Implementierung einer Ähnlichkeitssuche in dem PostgreSQL-Backend und FastAPI Frontend der TOAR DB (siehe z.B. https://stackoverflow.com/questions/11249635/finding-similar-strings-with-postgresql-quickly)
3. Implementierung von logischen Operatoren in Suchabfragen (Backend und Frontend)
4. Entwicklung von automatisierten und reproduzierbaren Tests für Punkte 2 und 3
5. Portierung der Entwicklungen in die operationelle TOAR-Datenbank und finale Tests

Hintergrund-Informationen:
Der Tropospheric Ozone Assessment Report ist ein Zusammenschluss von über 250 Wissenschaftlern aus etwa 30 Ländern mit dem Ziel, die globale Verteilung und zeitlichen Trends des Luftschadstoffs Ozon zu analysieren, dokumentieren und interpretieren. Eine wesentliche Grundlage von TOAR ist die TOAR-Datenbank, die am Jülich Supercomputing Center entwickelt und betrieben wird.
Um den beteiligten Wissenschaftlern effiziente Datenanalysen zu ermöglichen, müssen Daten und Metadaten aus der TOAR Datenbank leicht und effizient abfragbar sein. Dementsprechend wurde eine generische Suchfunktion (SEARCH Endpoint der TOAR DB API) entwickelt. Diese ist in ihrem Funktionsumfang jedoch aktuell noch beschränkt. Zum einen müssen Stationsnamen bekannt sein, um gefunden zu werden, was vor allem im mehrsprachigen Kontext zu Problemen führen kann (Beispiel: Dueren statt Düren). Ähnliches gilt für andere Textfelder in der Datenbank, deren Abfrage robuster auf weitere mögliche Schreibweisen oder Tippfehler reagieren sollte.
Zum zweiten ist aktuell nur eine UND Verknüpfung von Suchbegriffen möglich (bzw. nur eine ODER Verknüpfung mehrerer Terme desselben Schlüsselwortes). Demgegenüber ist es oftmals gewünscht, komplexere Abfragen wie z.B. "(Stationshöhe > 1000 m UND Bevölkerungsdichte < 800 km-2) ODER (Stationshöhe <= 1000 m UND Bevölkerungsdichte < 500 km-2)" durchführen zu können.
Die TOAR-Datenbank ist eine PostgreSQL (https://www.postgresql.org/)-Datenbank und die REST API wurde mit Python unter Nutzung von FastAPI (https://fastapi.tiangolo.com/), SQL Alchemy (https://www.sqlalchemy.org/) und Pydantic (https://docs.pydantic.dev/) entwickelt.
Als Ausgangspunkt für diese Arbeiten steht der Quellcode der REST-API im Repository https://gitlab.jsc.fz-juelich.de/esde/toar-data/toardb_fastapi zur Verfügung. Da ein Fokus unserer Arbeitsgruppe auf "Open Science" liegt und wir in diesem Bereich bereits mehrere Auszeichnungen erhalten haben, ist darüber hinaus beim Einsatz von bereits existierenden Software-Paketen darauf zu achten, dass der Code weiterhin den Kriterien für "Open Source" und "Open Access" genügt.