
iqual


PUZZLE ITC


Adfinis


APP Unternehmensberatung AG


GONICUS
PIONEERS OF OPEN SOURCE


BEKB | BCBE


HASLER
STIFTUNG

Stadt Zürich
Open Data

Automatisierung und Scraper

DINAcon 2020
23.10.2020
Stefan Oderbolz

1 Scraper - Was ist das?

Ausgangslage

– Es ist Mitte März, **aktuelle, maschinenlesbare** Daten zu SARS-CoV-2 sind rar

– Im openZH/covid_19 GitHub Repository entstehen CSVs mit **händisch zusammengetragenen** Zahlen

– Kantone publizieren meist direkt über ihre **Webseiten**

Was ist ein Scraper?

– Engl. *to scrape*: kratzen/schaben

– Idee: Daten von Webseiten oder PDFs zu *kratzen*, und diese dann strukturiert ablegen

– **Ein Scraper ist ein Programm, das Daten aus semi-strukturierten Quellen** holt und zur Verfügung stellt

2 Realisierung der Scraper

Wie sind die Scraper implementiert?

- Das Projekt wurde als Hackathon initiiert
  - Interessierte haben eigene Scraper entwickelt
  - Verschiedene Programmiersprachen (Shell, Python, Ruby)
  - Verschiedene Infrastruktur (eigene Server, morph.io)

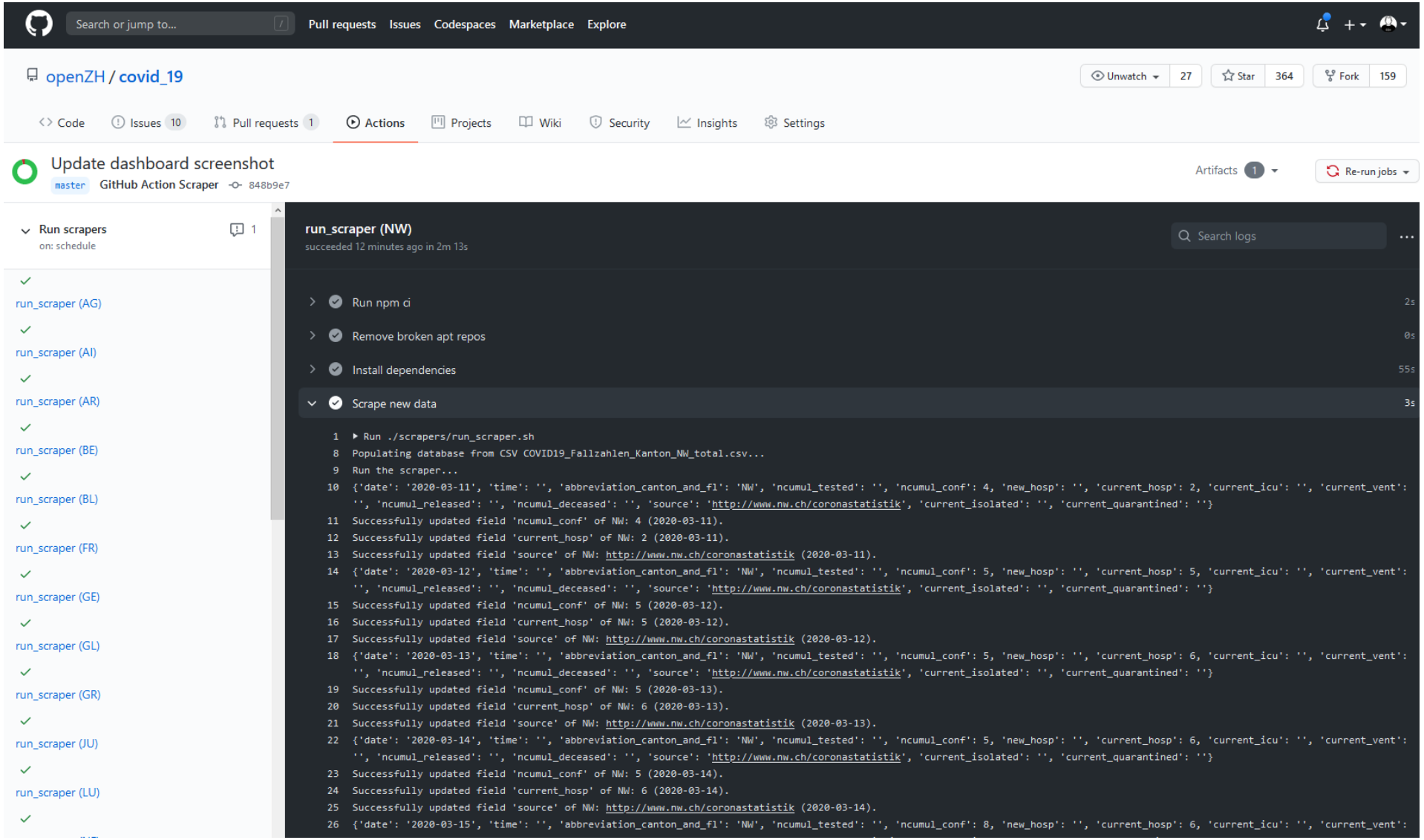
- Scraper basieren auf Unix-Philosophie
  - **«Write programs that do one thing and do it well»**
  - Der Output eines Programms ist der Input des nächsten (Pipes)
  - Modularer Aufbau, so dass einzelne Komponenten ausgetauscht bzw. in verschiedenen Programmiersprachen implementiert werden können

Automatisierung mit GitHub Actions

- GitHub Actions ermöglicht beliebige Programme laufen zu lassen
  - Zeitsteuerung (cron)
  - Integriert ins GitHub Ökosystem

- Scraper laufen alle 20 Minuten
  - Ein Scraper für jeden Kanton + FL
  - Neue Daten werden direkt ins Repository commitet
  - Fehlermeldungen landen in einem Slack-Channel
  - Die Daten werden validiert und auf Ausreisser geprüft

Automatisierung mit GitHub Actions

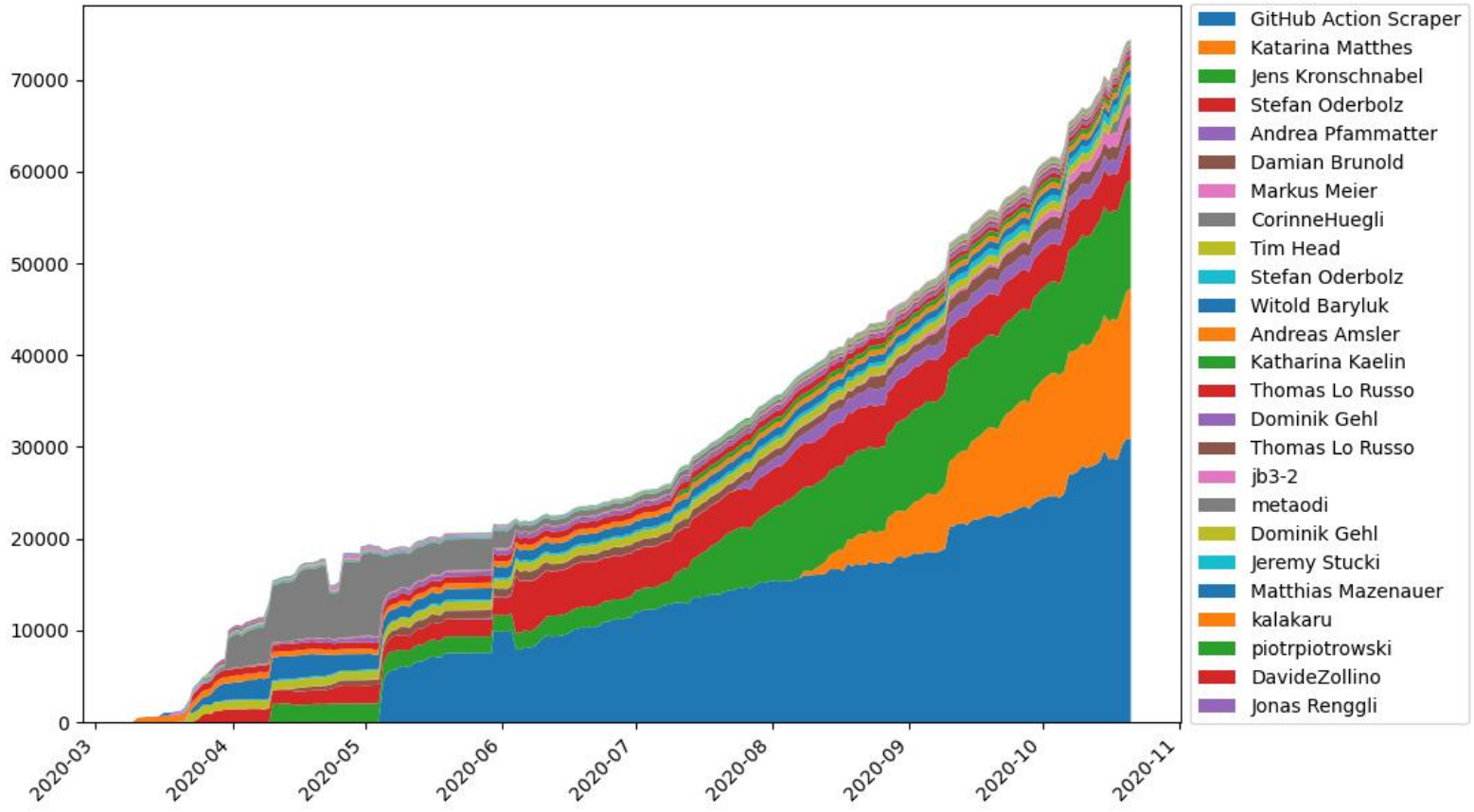


3 Hinter den Kulissen

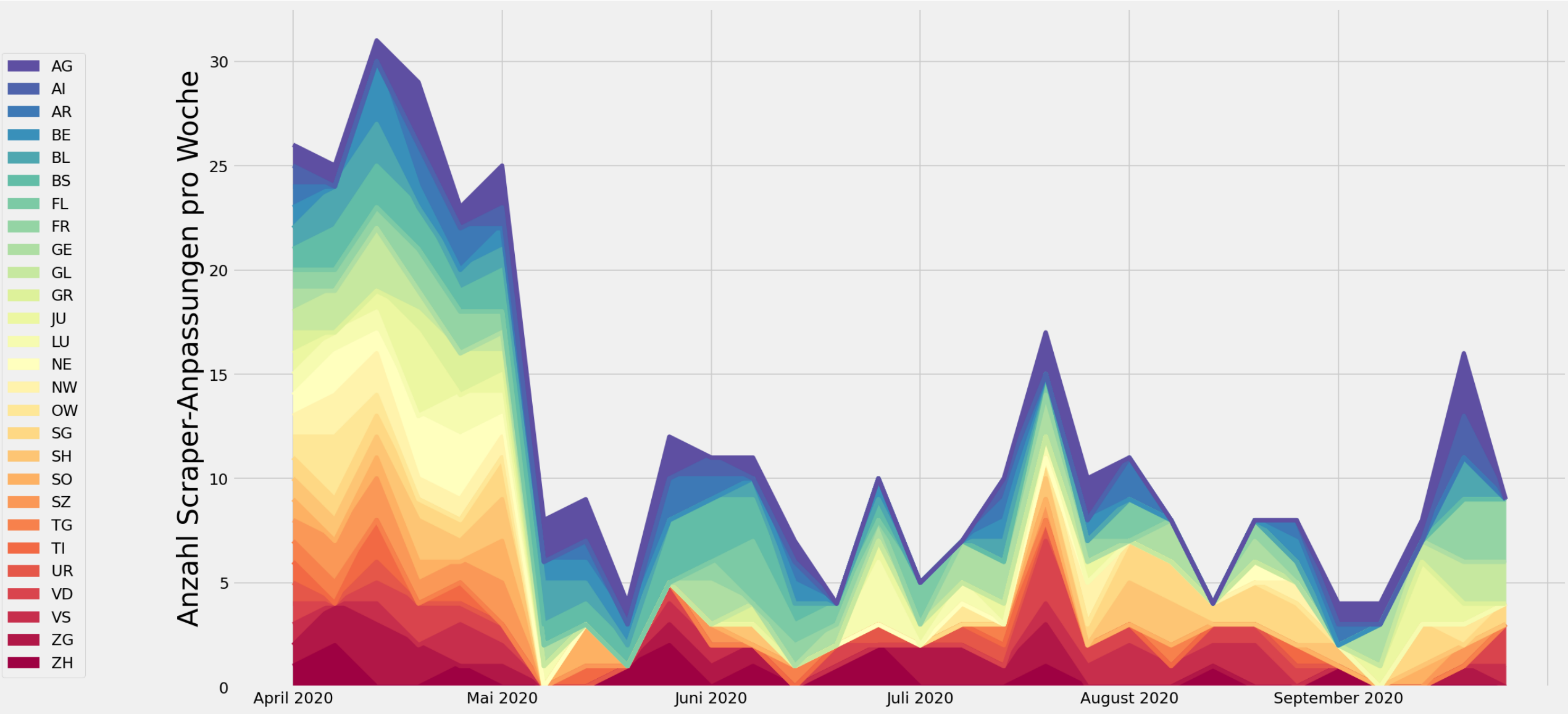
Repository-Statistik

Stand 22.10.2020

- Insgesamt über
  16'500 Commits
  - **55** Contributors
  - **882** Pull Requests
  - **351** Issues
  - **14'229** Scraper-Durchläufe je Kanton

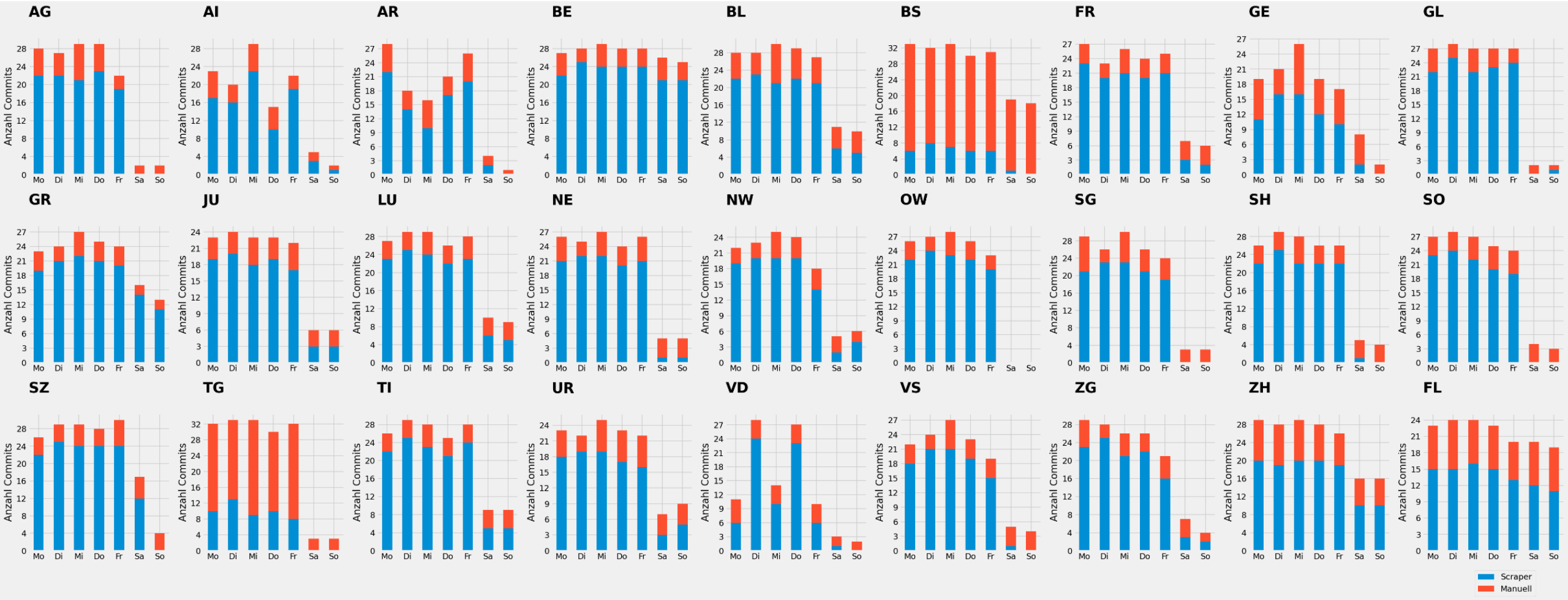


Scraper-Anpassungen



Daten von den Kantonen

Commits pro Wochentag und Kanton in den letzten 20 Wochen



Vielen Dank.