FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Management of Scientific Data - Prüfung

Zusammenhang zwischen COVID-19 Fällen und allg. Testabdeckung

12.07.2024

Überblick

- Szenario & Forschungsfrage
- Data Lifecycle
- FAIR
- Verwendete Werkzeuge & Live Demo

Das Szenario

- Datasets:
  - COVID-19 Fälle und Tode - ca. 12600 Einträge
  - COVID-19 Testhäufigkeit - ca. 6100 Einträge

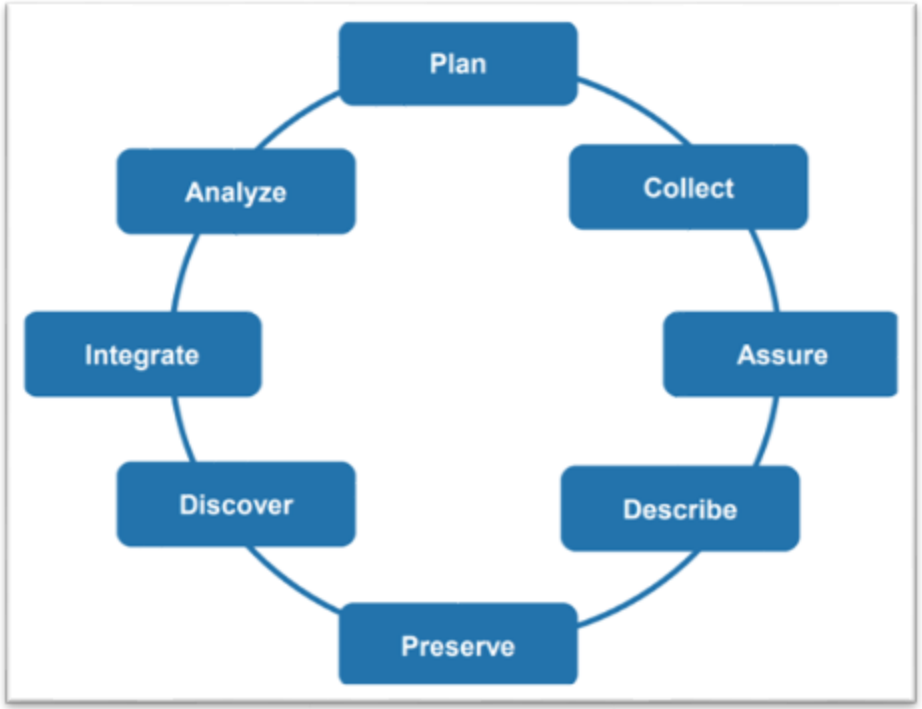
- Passt thematisch gut zu ReproHack Review im Semester
- Relativ aktuelles Thema
- Daten von Anfang 2020 – Ende 2023

Die Forschungsfrage

Gibt es eine Abhängigkeit zwischen Testhäufigkeit und gemeldeten COVID-19 Fällen?

- Relativ triviale Forschungsfrage
- Bietet trotzdem genug Möglichkeiten zur Datenverarbeitung und Analyse
- Fokus liegt auf dem Management der Daten, nicht auf der Analyse
- **Hypothese:** Es besteht eine Abhängigkeit zwischen den zwei Faktoren

Data Lifecycle



Plan

- **Data Management Plan**
  - Basierend auf Horizon Template (persönliche Präferenz)
  - Ermöglicht schnellen Projektstart und beinhaltet viele der zentralen Fragen
- **Dokumentation nach der Idee eines *"living document"***
  - Erstellung eines öffentlichen GitHub Repository mit README & MIT Lizenz
  - Erklärung der Ordnerstruktur und E-Mail Bereitstellung für weitere Fragen
  - Ein ständig aktualisiertes Dokument mit allen Informationen
  - Workflow: Stufe des DLC abarbeiten -> Informationen einfügen -> Nächste Stufe -> bei evtl. späteren Änderungen Dokumentation aktualisieren

Collect

- Die Daten sind quantitativ, strukturiert und glaubwürdig.
- In gängigen Formaten verfügbar (CSV, JSON, XML, …)
- Automatisches/Manuelles web-scraping der ECDC
- Datensätze heißen immer *"data.csv"* -> Keine Eindeutigkeit
- Nicht global repräsentativ -> Europa biased

Collect

- **Datenquellen**
  - Primär: European Surveillance System (TESSy)
  - Sekundär: Öffentliche Online-Quellen -> Datenqualität?

Assure

- **Completeness**
  - 7.63% aller Einträge des Deaths/Cases Datensatz haben NaN Werte
  - Bei Testing Datensatz: 18.86%
- **Uniqueness**
  - Sortierung nach Land und Datum stellt Einzigartigkeit sicher
- **Timeliness**
  - Relativ repräsentativ
  - Keine 100% Garantie das in einer Pandemie alles akkurat ist

Assure

- **Validity**
  - Spalten sind valide, konkret und selbsterklärend
  - Bei fehlenden Werten wird konstant NA angegeben
- **Accuracy**
  - Keine Duplikate
  - Alle Spalten enthalten vernünftige/erwartbare Werte
- **Consistency**
  - Gute Konsistenz
  - Kleinere Inkonsistenzen zwischen Datensätzen (Ländercodes: AUT/AT, ...)

Describe

- Website bietet für Deaths/Cases Dataset wenig Informationen
- Testing Volume Dataset enthielt deutlich mehr Metadaten
- GitHub Repository enthält keine Metadaten -> ausführliche README oder Dokumentation wäre hilfreich
- **Aber:** Daten sind meist selbsterklärend, selbst für Menschen ohne medizinischen Hintergrund -> Arbeit mit Daten ist gut möglich

Preserve

- Daten redundant auf Website & GitHub gespeichert -> gut
- Zusätzlicher Upload auf Zenodo o.ä. Wünschenswert
- Keine Verbindung zu einem Artikel & keine Quality Features angegeben
- DOI oder andere PID fehlen
- Keine Autoren, aber Accounts bei GitHub auffindbar

Preserve

- Metadata ist teilweise vorhanden
- Öffentlicher Zugriff auf Daten
- Keine direkte Lizenz, aber Verweis auf ECDC Copyright (CC BY 4.0)
- Kein Überblick auf die Daten/Struktur von der Website aus
- Archive von früheren Zeitpunkt vorhanden (Juni 2022)
- Website wurde indiziert und ist gut bei Suchmaschinen zu finden

Discover

- Viele COVID-19 Datensätze verfügbar auf Zenodo o.ä.
  - Öffentliche Datensätze schränken die Anzahl stark ein
  - Regionale Probleme -> Viele Daten sind nur für spezifische Regionen
- Daten unseres ReproHack-Projekt könnten genutzt werden
- Mehr Informationen dann in Live Demo

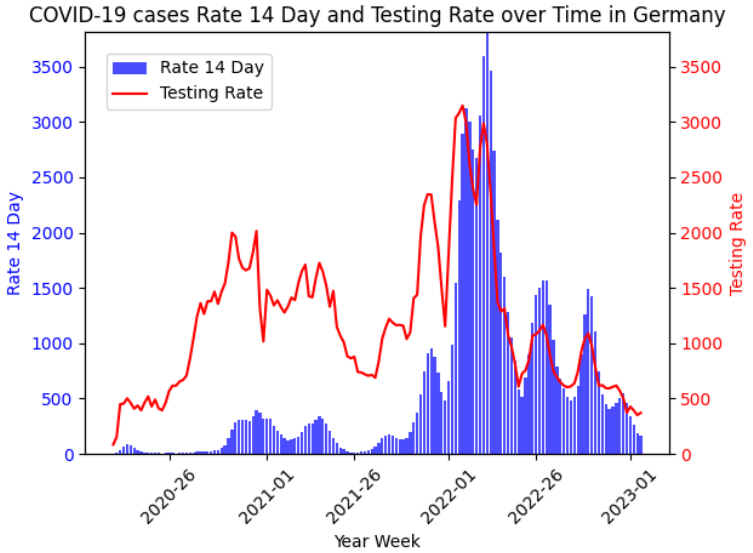
Integrate

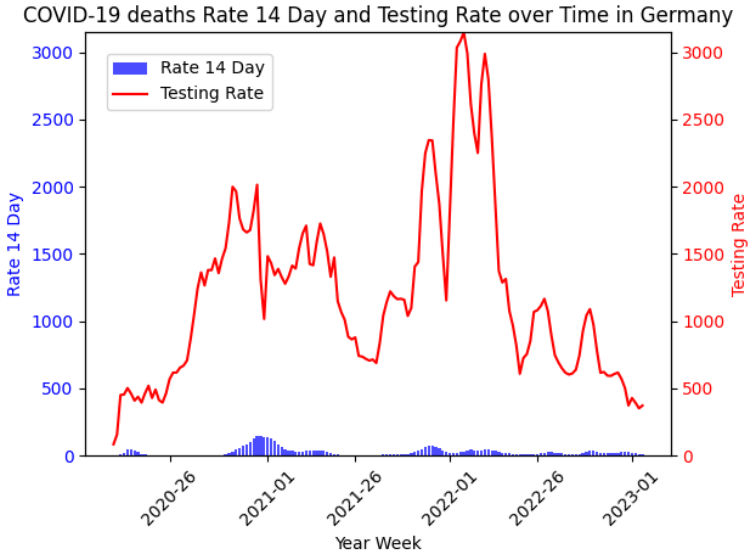
- Datensätze waren gut zu kombinieren
- Vorverarbeitung:
  - Anpassung des Datum-Formats
  - Zusammenführung der Datensätze
  - Entfernung nicht nötiger und redundanten Spalten
  - Entfernung aller Spalten mit NaN Werten
  - Export nach Land
- Hilfe von AI bei Entwicklung hatte positiven Einfluss

Analyze

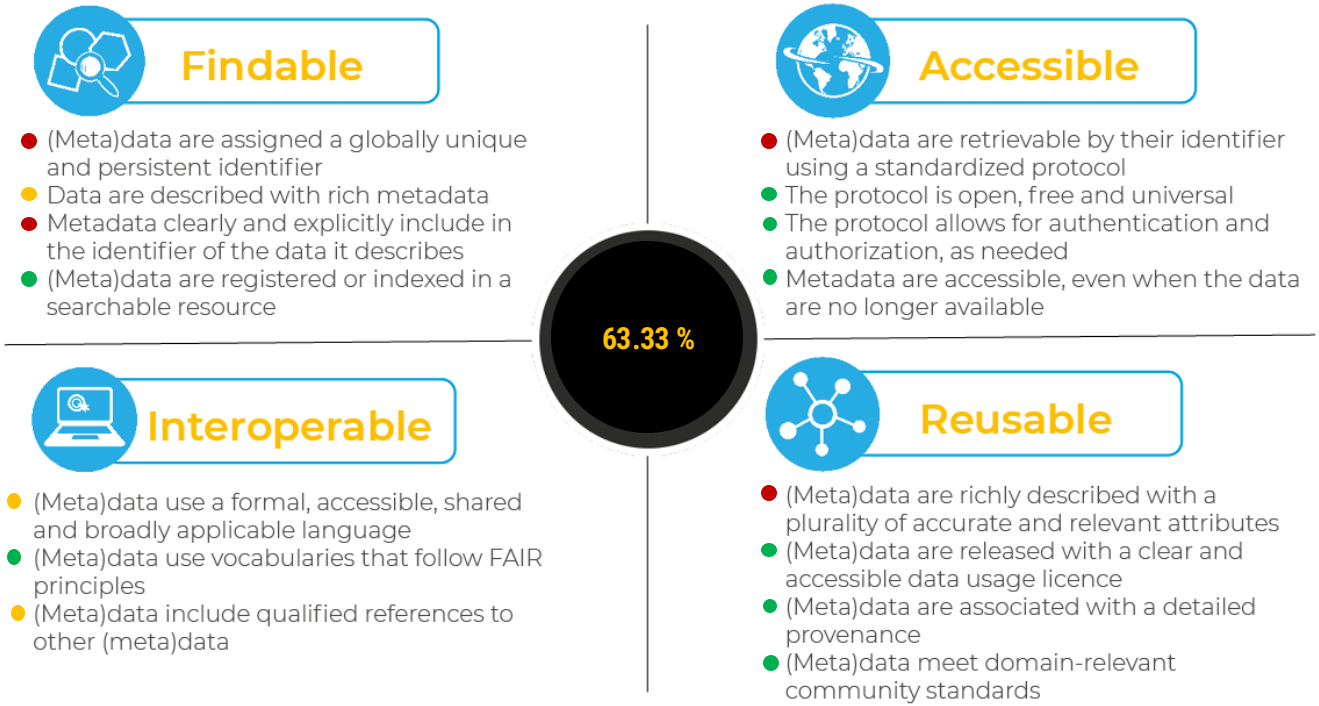
- Arbeitsschritte:
  - Iteration über alle vorverarbeiteten Länder-Daten
  - Aufteilung in "Cases" und "Deaths"
  - Generation der Plots in Kombination mit der Test-Rate
  - Überprüfung ob Abhängigkeit besteht

Analyze


COVID-19 cases Rate 14 Day and Testing Rate over Time in Germany
Rate 14 Day
Testing Rate
Rate 14 Day
Testing Rate
Year Week
0
500
1000
1500
2000
2500
3000
3500
2020-26
2021-01
2021-26
2022-01
2022-26
2023-01


COVID-19 deaths Rate 14 Day and Testing Rate over Time in Germany
Rate 14 Day
Testing Rate
Rate 14 Day
Testing Rate
Year Week
0
500
1000
1500
2000
2500
3000
2020-26
2021-01
2021-26
2022-01
2022-26
2023-01

FAIR


Findable
(Meta)data are assigned a globally unique and persistent identifier
Data are described with rich metadata
Metadata clearly and explicitly include in the identifier of the data it describes
(Meta)data are registered or indexed in a searchable resource
Accessible
(Meta)data are retrievable by their identifier using a standardized protocol
The protocol is open, free and universal
The protocol allows for authentication and authorization, as needed
Metadata are accessible, even when the data are no longer available
63.33 %
Interoperable
(Meta)data use a formal, accessible, shared and broadly applicable language
(Meta)data use vocabularies that follow FAIR principles
(Meta)data include qualified references to other (meta)data
Reusable
(Meta)data are richly described with a plurality of accurate and relevant attributes
(Meta)data are released with a clear and accessible data usage licence
(Meta)data are associated with a detailed provenance
(Meta)data meet domain-relevant community standards

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Vielen Dank für Ihre Aufmerksamkeit!