FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Management of Scientific Data – Prüfung

Korrelation zwischen Testhäufigkeit und Anzahl der COVID-19 Fällen

19.07.2025

Inhalt

- Szenario & Forschungsfrage
- Data Lifecycle
- FAIR Analyse
- Live Demo

Szenario

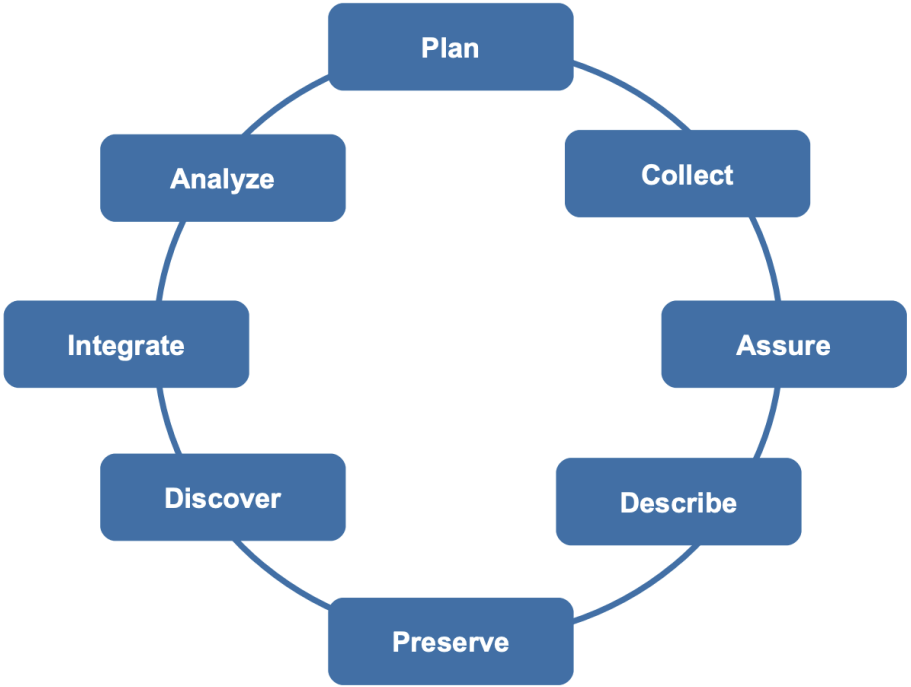
- Datensätze:
  - COVID-19 Fälle und Tode – ca. 12600 Einträge
  - COVID-19 Testhäufigkeit – ca. 6100 Einträge
- Thema betrifft nach wie vor viele Menschen
- Relativ aktuelle Daten (Anfang 2020 - Ende 2023)

Forschungsfrage

Korreliert die Testhäufigkeit mit der Anzahl gemeldeter COVID-19-Fälle?

- Relativ triviale Forschungsfrage
- Bietet trotzdem genug Möglichkeiten der Datenverarbeitung und Analyse
- Fokus dieser Ausarbeitung liegt auf Management der Daten, nicht auf Analyse
- **Hypothese:** Die Anzahl der COVID-19-Fälle korreliert mit der Testhäufigkeit

Data Lifecycle



Plan

- **Data Management Plan**
  - Horizon Template als Vorlage
  - Schneller und strukturierter Projektstart
- **Dokumentation nach der Idee eines „living document“**
  - GitHub Repository mit README & Open Source Lizenz – MIT Lizenz
  - Workflow: Stufe des Data Life Cycle abarbeiten -> Informationen einfügen -> Nächste Stufe -> bei evtl. späteren Änderungen Dokumentation aktualisieren

Collect

- Die Daten sind strukturiert, offiziell und quantitative
- Daten sind in gängigen Formaten verfügbar (CSV, JSON, XML, XLSX)
- Automatisches web-scraping der ECDC
- Eindeutigkeit geht beim Herunterladen verloren -> Bezeichnung immer ”data.csv
- Auf Europa beschränkt, daher nicht unbedingt repräsentativ

Collect

- **Datenquellen**
  - Primär: European Surveillance System (TESSy)
  - Sekundär: Öffentliche online Quellen -> Kein Hinweis zur 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**
  - Pro Land repräsentativ
  - Während Pandemie schwierig eine 100% Garantie zu geben

Assure

- **Validity**
  - Spalten sind valide, konkret und selbsterklärend
  - Bei fehlenden Werten wird konstant NA
- **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 Zeitpunkten vorhanden (Juni 2022)
- Website wurde indiziert und ist gut durch 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

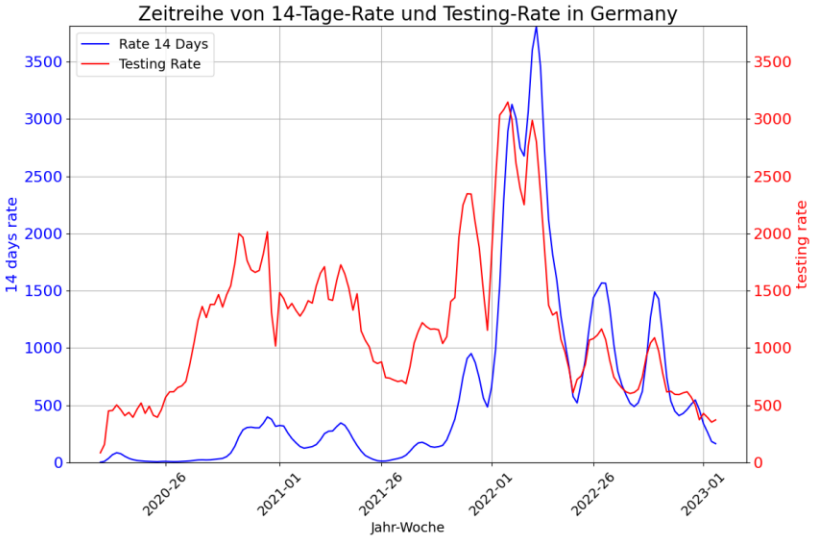
Integrate

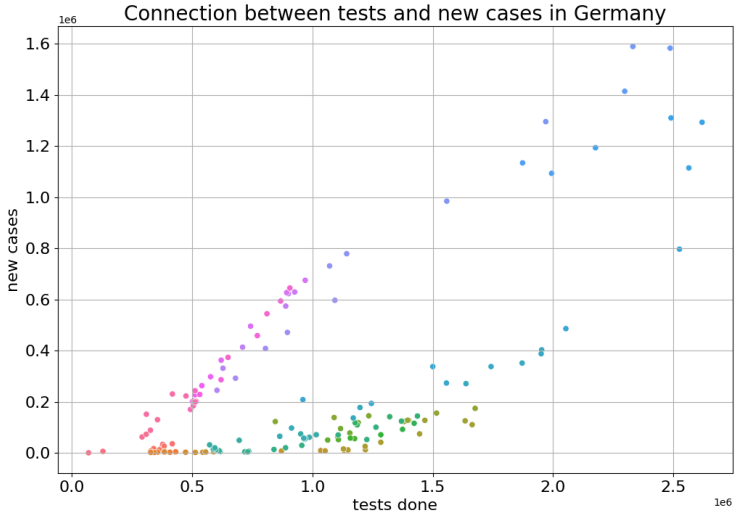
- Datensätze waren gut zu kombinieren
- Vorverarbeitung:
  - Anpassung des Datum-Formats
  - Zusammenführung der Datensätze
  - Entfernung nicht nötiger und redundanten Spalten
  - Entfernung aller Spalten mit NaN Werten
  - Export nach Land
- Hilfe von AI bei Entwicklung hatte positiven Einfluss

Analyze

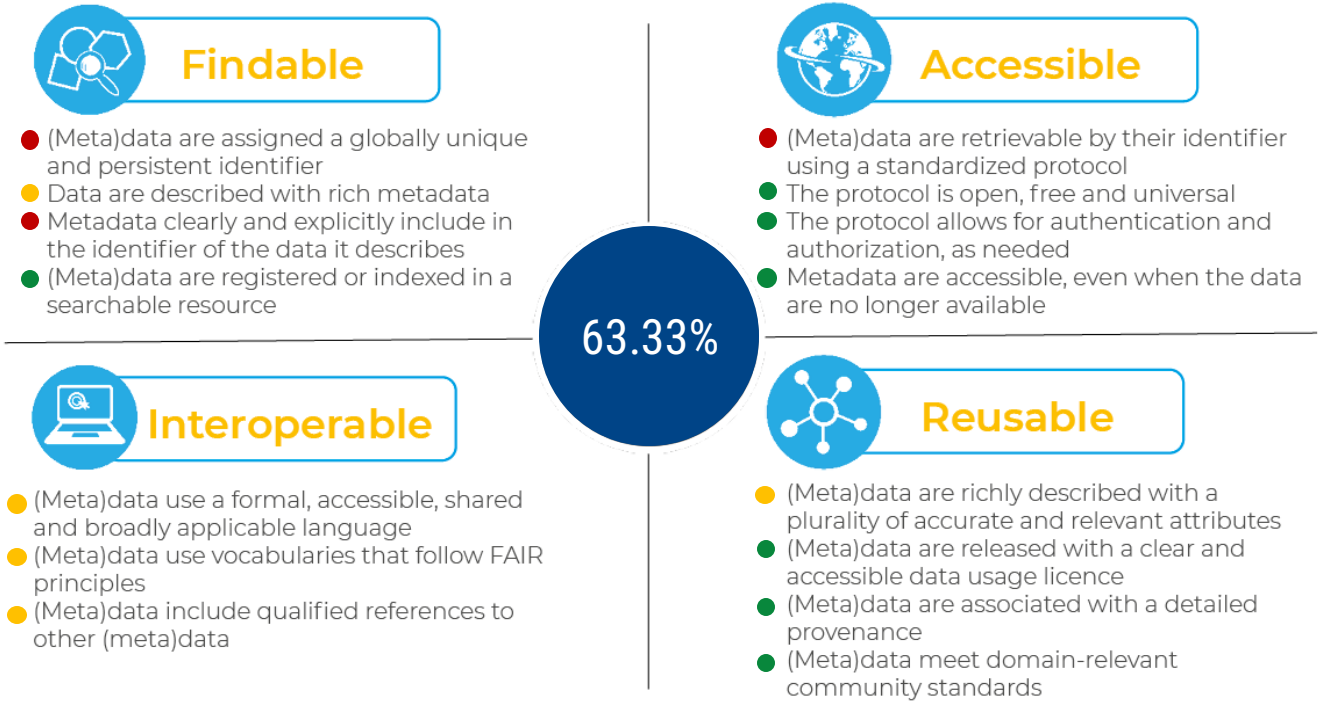
- Arbeitsschritte:
  - Iteration über alle vorverarbeiteten Länder-Daten
  - Aufteilung in "Cases" und "Deaths"
  - Generation der Plots in Kombination mit der Test-Rate
  - Überprüfung ob Abhängigkeit besteht

Analyze


Zeitreihe von 14-Tage-Rate und Testing-Rate in Germany
Rate 14 Days
Testing Rate
14 days rate
testing rate
Jahr-Woche
2020-26
2021-01
2021-26
2022-01
2022-26
2023-01


Connection between tests and new cases in Germany
1e6
new cases
tests done
1e6

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

“The only relevant test of the validity of a hypothesis is comparison of prediction with experience.”

- Milton Friedman -

Vielen Dank für Ihre Aufmerksamkeit!

Justin Bergmann