Management of Scientific Data - Prüfung

Hatte die wirtschaftliche Stärke eines EU-Landes Einfluss auf ihren Covid-19 Pandemie Verlauf?

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA Fakultät für Mathematik
und Informatik

25.07.2024 | Fabian Hupfeld

Forschungsfrage

Datensatz: Covid-19 Data from the European Centre for Disease Prevention and Control Data

- Covid-19 Daten von Frühling 2020 bis Winter 2023
- 12648 Einträge (26 EU-Länder + EU mit wöchentliche Todesfälle und Testfälle)

Hatte die wirtschaftliche Stärke eines EU-Landes Einfluss auf ihren Covid-19 Pandemie Verlauf?

Zusätzlicher Datensatz benötigt:

- Wirtschaftliche Stärke aller EU-Länder von 2020-2023 benötigt

Plan

Plan - Collect - Assure - Describe - Preserve - Discover - Integrate - Analyse

Erstellen eines Data Management Plan

- Template von Horizon Europe (V1.1 2022)
- Einfache Struktur mit allen wesentlichen Management-Aspekten
- Übersichtliches kleines Projekt:
  - Keine Kosten, nur eine verantwortliche Person
  - keine persönlichen Daten gespeichert
  - Viele Fragen direkt beantwortet => Kein wirkliches “Living Document”

Projekt mit Github erstellt

- Readme und License erstellen + DMP hochladen
- Projekt entlang des Data Lifecycles bearbeiten

Collect

Plan - **Collect** - Assure - Describe - Preserve - Discover - Integrate - Analyse

Covid-Datensatz bereits gesammelt, wirtschaftlicher Datensatz fehlt

Daten vergangener Werte können nicht selber generierbar/aufgezeichnet werden

- Bereits existierende Datensätze verwenden
- Vertrauenswürdige Quellen mit strukturierten quantitativen Datensätzen
  - Ähnliche Organisation wie ECDC (Covid- Datensatz) für wirtschaftlichen Datensatz?
  - Ansonsten Web-Scraping um passende Datensätze zu finden
- Bei gefundenen Datensätzen Quelle überprüfen + Metadaten sammeln

Wirtschaftlicher Datensatz von EuroStat gefunden

- Kaufkraftstärke (KKS=Lebenskosten angepasster BIP) der EU-Länder von 2012-2023

Assure

Plan - Collect - **Assure** - Describe - Preserve - Discover - Integrate - Analyse

Datensätze auf Qualitätsmerkmale überprüfen (manuell und automatisch)

- **Completeness:** Covid Datensatz 89%, KKS Datensatz 100%
- **Uniqueness:** Kombination aus Land und Datum (+ Covid-Fallart) sind einzigartiger Eintrag
- **Timeliness:** Datensatz nach Ende der Erfassungszeitraums veröffentlicht (+ aktualisiert)
- **Validity:** Einträge stimmen mit den beschriebenen Eintragstypen überein (Metadaten)
- **Accuracy:** Nicht nachvollziehbar, aber Quelle vertrauenswürdig
- **Consistency:** Relativ Konsistent, aber Spaltennamen weichen leicht ab
  - Preprocessing der Datensätze nötig -> Spalten umbenennen + teilen
  - Script für Preprocessing schreiben

Describe

Plan - Collect - Assure - **Describe** - Preserve - Discover - Integrate - Analyse

Arbeitsschritte und generierte Daten dokumentieren

Workflow dokumentieren und ausführliche Readme erstellen

- Markdown Format bietet schnelle Möglichkeit der Dokumentation

Metadaten zur Datensatzbeschreibung erstellen

- Für wiederverwendeten Datensätzen Metadaten übernehmen oder neu erstellen
  - Covid Datensatz hat Metadaten als PDF mit grundlegenden Informationen
  - KKS Datensatz hat umfangreiche machinenlesbare Metadaten in xlsx-Format
  - Zusätzlich Provenance_Information.md um auf Source-URLs zu verweisen
- maschinenlesbare Metadaten für neue Datensätze (dublincoregenerator.com)
- Insbesondere Spalten der CSV-Dateien beschreiben

Preserve

Plan - Collect - Assure - Describe - **Preserve** - Discover - Integrate - Analyse

Github Repository zum Speichern der Daten/Code

- Automatische Backups und Zugriff per Internet
- Keine laufenden Kosten und lange archiviert
- MIT-Lizenze für Code und CC0-Lizenz für Daten angeben
- Daten zusätzlich lokal speichern
- Für die kleine Projektgröße akzeptable

Bei größeren Projekten

- Daten auf Zenodo speichern(Erleichtert das Finden des Datensatzes durch Keywords etc)
- Unique Identifier für Datensätze(DOI/URN) und für Autoren(ORCID/ResearcherID)

Discover

Plan - Collect - Assure - Describe - Preserve - **Discover** - Integrate - Analyse

Keine weiteren Datensätze notwendig für oberflächliche Analyse

Weitere Datensätze erst notwendig um die Forschungsfrage genauer zu beantworten

Zenodo hat viele Datensätze über Covid und Wirtschaft

- Oft regionale/lokale Daten, weniger globale Daten auf Länderebene

Insbesondere EuroStat empfehlenswert für Daten über Europa

- Sehr gut dokumentiert mit machinenlesbaren Metadaten
- Globale vollständige und aktuelle Datensätze mit Re-Use License

Integrate

Plan - Collect - Assure - Describe - Preserve - Discover - **Integrate** - Analyse

Datensätze im csv Format leicht zu integrieren

Data Preprocessing mit Jupyter Notebook:

- Format für den Zeitpunkt und ISO-Länderbezeichner anpassen
- Entfernen von nicht benötigten Spalten
- Aufteilen in Covid-Death und Covid-Cases csv file
- Zusätzlich bei dieser Visualisierung : Datenpunkte mit NaN-Einträgen entfernen
- Zusammenfügen der vorverarbeiteten Covid und KKS Datensätze

Weitere Datensätze können leicht in Jupyter Notebook integriert und vorverarbeitet werden

Analyse

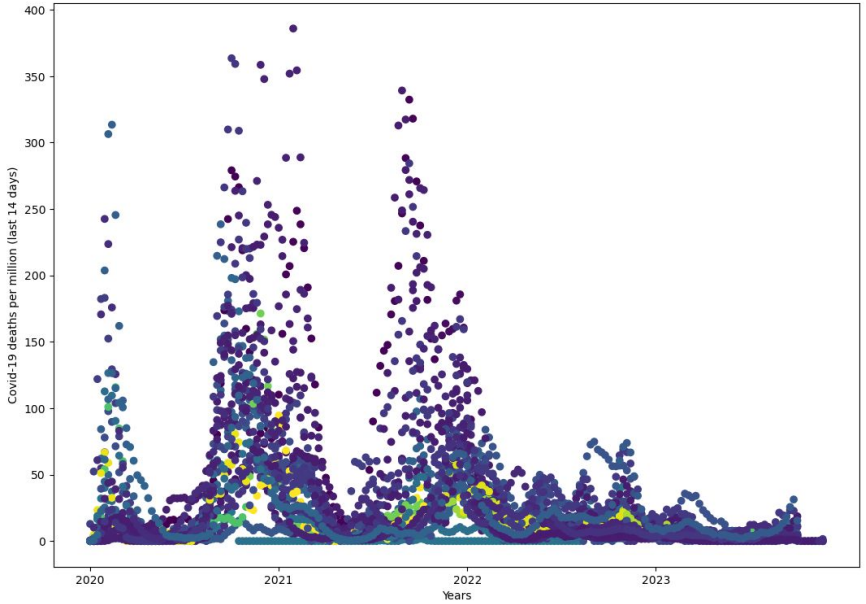
Plan - Collect - Assure - Describe - Preserve - Discover - Integrate - **Analyse**

Scatterplot-Visualisierung mit Jupyter-Notebook und Matplotlib

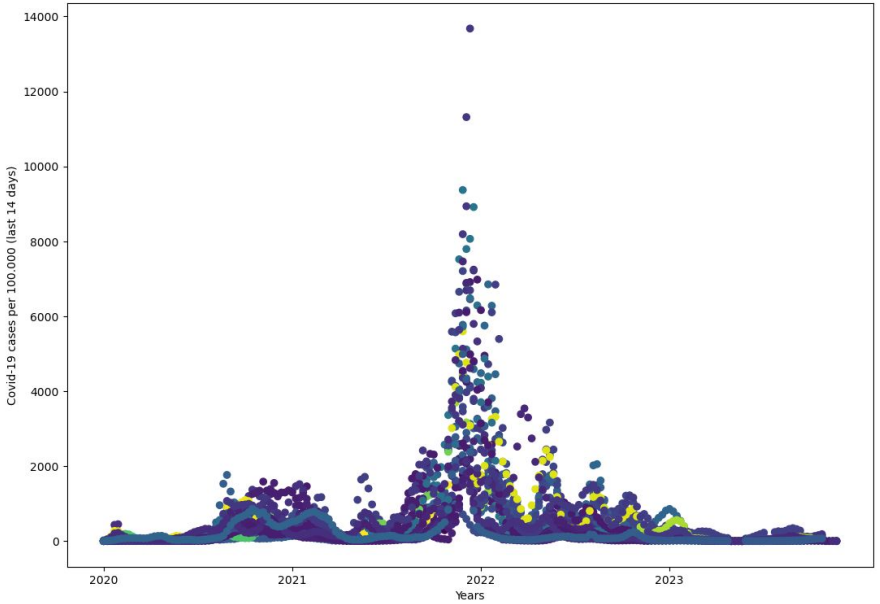
- Vorverarbeitete und zusammengeführte Datensätze laden
- Jeweils Plots für Covid-Deaths und Covid-Cases generieren und speichern
- Visuelle Überprüfung der Forschungsthese
  - Für eine genauere Aussage müssten statistische Methoden verwendet werden

Die wirtschaftliche Stärke eines EU-Landes scheint mit der jeweiligen Anzahl an Covid-19 Todesfällen zu korrelieren.

Analyse



Covid-19 Tote pro Millionen



Covid-19 Fälle pro Hunderttausend

Fair-Assessment

ARDC Assessment:

- Findable **59**%
  - Nur URL als Identifier und ein allgemeines Repository
- Accessible **90**%
  - Metadata wird gelöscht wenn das Projekt (Github Repo) gelöscht wird
- Interoperable **88**%
  - Metadata verweist meistens mit URLs zu anderen Daten/Metadaten
- Reusable **86**%
  - Provenance Informationen überwiegen als nicht machinenlesbares Format gegeben

Die optimistische Bewertung von ARDC ergibt ein FAIR-Gesamtergebnis von **84**%

Vielen Dank
für die Aufmerksamkeit