JOINT INNOVATION LAB
energizing eGov!

Daten – Wie müssen sie aussehen?

Datenqualität

Wie muss die Datenqualität für wissenschaftliche Analysen und gesellschaftliche Teilhabe aussehen?

Die „**FAIR Guiding Principles for scientific data management and stewardship**" von Wilkinson et al. wurden im Jahr 2016 veröffentlicht. FAIR steht für:

Auffindbarkeit (engl. **F**indability)	Der erste Schritt bei der (Wieder-)Verwendung von Daten besteht darin, sie zu finden. (Meta-)daten sollten sowohl für Menschen als auch für Computer leicht zu finden sein.
Interoperabilität (engl. **I**nteroperability)	Sobald der Nutzende die gewünschten Daten gefunden hat, muss er wissen, wie er auf sie zugreifen kann, möglicherweise einschließlich Authentifizierung und Autorisierung.
Erreichbarkeit (engl. **A**ccessibility)	Für Anwendungen oder Arbeitsabläufe für die Analyse, Speicherung und Verarbeitung müssen die Daten mit anderen Daten zusammengeführt werden.
Wiederverwendung (engl. **R**euse)	Das Ziel von FAIR ist es, die Wiederverwendung von Daten zu optimieren. Um dies zu erreichen, sollten Metadaten und Daten gut beschrieben sein.

Appelfeller und Feldmann (2018) definieren **sechs Merkmale** für eine **hochwertige Datenqualität**:

Relevanz	Die Inhalte der einzelnen Felder sind für den Anwendenden von Nutzen.
Eindeutigkeit	Die einzelnen Felder des Datensatzes identifizieren eindeutig ein Objekt der Realität.
Richtigkeit	Die Inhalte der einzelnen Felder eines Datensatzes müssen richtig sein.
Konsistenz	In jedem System müssen alle Felder des Datensatzes die gleiche Ausprägung haben.
Aktualität	Die Inhalte der einzelnen Felder eines Datensatzes müssen aktuell sein.
Vollständigkeit	Alle für einen Datensatz definierten Felder müssen gepflegt sein.

Bsp.: im Kontext Haushaltsdaten

Welche Beispiele gibt es im Kontext Haushaltsdaten für KI-Systeme?

Bereitstellung auf Portalen:
Nutzung von APIs ermöglichen, um Auffindbarkeit zu erhöhen z.B. auf *opendata.schleswig-holstein.de*

Format:
Maschinenlesbarkeit gewährleisten
z.B. über .csv oder .json

Rohdaten:
Trotz aufbereiteten Visualisierungen und interaktiven Angeboten Rohdaten bereitstellen

Zusätzliche Daten:
z.B. Daten zu tatsächlichen Auszahlungen bereitstellen

Nutzung von Meta-Daten:
Was verbirgt sich hinter welchem Titel?
Sind Änderungen in Titeln dokumentiert?

Aussagekräftige Beschriftungen/Erklärungen, z.B. Spaltenbezeichnung „Betrag" nicht aussagekräftig; Sind die Deckungskreise nachvollziehbar?

Teilweise defekte .csv-Dateien auf öffentlich zugänglichen Seiten

Sich ändernde Titelbezeichnungen sind problematisch, selbst wenn sich nur ein Punkt oder Komma ändert, kann schon keine Zeitreihe erstellt werden.

Hier bestand das Problem, dass nur aktuelle Daten vorlagen. Für KI-Systemen sind Daten über einen längeren Zeitraum notwendig.

Lernen Sie mehr über Daten in unserem eGov-Campus Kurs!

Appelfeller, W. & Feldmann, C. (2023). Die digitale Transformation des Unternehmens. In *Springer eBooks*. https://doi.org/10.1007/978-3-66265413-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. [...] Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18