Wie mache ich Daten FAIR?

Im Jahr 2016 wurden die "FAIR Guiding Principles for scientific data management and stewardship" in *Scientific Data* veröffentlicht (Wilkinson et al 2016). Die Autoren beabsichtigten, Leitlinien zur Verbesserung der Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit digitaler Objekte zu erstellen.

F wie FINDABLE = AUFFINDBAR
Um Daten auffindbar zu machen, sollten die Daten mit umfassenden **Metadaten** versehen sein, die sowohl von Menschen als auch Maschinen gelesen werden können. Indem die Daten zusätzlich auf einem Repository hinterlegt werden, welches **persistente Identifikatoren** (= persistent identifiers) vergibt (z.B. DOI), können Sie garantieren, dass Ihre Daten auch bei wechselnden URLs immer gefunden werden. Verlinken Sie zudem Ihre verschiedenen digitalen Objekte (Publikationen, Daten, etc.) mit Hilfe dieser persistenten Identifikatoren miteinander.

A wie ACCESSIBLE = ZUGÄNGLICH
Machen Sie Ihre Daten zugänglich, indem Sie in einem «data access statement» festhalten, wer unter welchen Bedingungen Ihre Daten nutzen kann: Kann jeder die Daten einfach einsehen, der Internetzugang hat? Müssen Interessierte sich registrieren, um die Daten nutzen zu können? Auch wenn die Daten selbst aus ethischen oder legalen Gründen nicht offen sind, stellen Sie immer sicher, dass die **Metadaten** zugänglich bleiben.

I wie INTEROPERABLE = KOMPATIBEL
Datensätze werden oft mit anderen Datensätzen verbunden oder müssen in unterschiedlichen Computerprogrammen geöffnet werden können. Indem Sie Ihre Daten in einem **offenen Format** mit klaren Angaben zur Herkunft zur Verfügung stellen, garantieren Sie, dass Ihre Daten für die weitere Analyse, Verarbeitung und Archivierung mit diversen Software-Umgebungen kompatibel bleiben. Wenn Sie zudem Ihre (Meta)Daten und Variablen anhand eines **kontrollierten Vokabulars** benennen, können Ihre Daten von Anderen besser verstanden werden.

R wie REUSABLE = WIEDERVERWENDBAR
Letztendlich ist es das erklärte Ziel von FAIRen Daten, dass sie wiedergenutzt werden. Durch die Wahl einer **Lizenz** (z.B. Creative Commons) und detaillierter Beschreibung Ihrer Daten (**Metadaten**) erkennen Forschende die Nachnutzbarkeit und Nützlichkeit Ihrer Daten für ihre eigenen Forschungszwecke.

100% FAIRness ist fast unmöglich zu erreichen; man kann jedoch sein Augenmerk auf wichtige Aspekte von FAIR richten, nämlich: Metadaten, persistente Identifikatoren, offene Formate, kontrollierte Vokabulare und Lizenzen.

Metadaten

Im einfachsten Sinne sind Metadaten Informationen über Daten und beschreiben grundlegende Merkmale der Daten, wie z. B.

- **wer** die Daten erstellt hat
- **was** die Datendatei enthält
- **wann** die Daten erstellt wurden
- **wo** die Daten erstellt wurden
- **warum** die Daten erstellt wurden
- **wie** die Daten erstellt wurden

Metadaten erleichtern es Ihnen und anderen, Daten zu einem späteren Zeitpunkt korrekt zu identifizieren und wiederzuverwenden. Datendokumentation und Metadaten sind daher für die Reproduzierbarkeit und Wiederverwendbarkeit von Forschungsergebnissen unerlässlich.

Tipp: Erstellen Sie Ihre Metadaten sobald Sie beginnen, Ihre Daten zu sammeln. Das erspart Ihnen später viel Zeit und Mühe.

Metadatenstrukturen werden oft als "Schemas" bezeichnet. Das Schema enthält eine Reihe von Merkmalen zur Beschreibung der Daten. Die fertigen Metadaten können in einer maschinenlesbaren Sprache wie XML angegeben oder aber auch in einer von Menschen lesbaren Form einer ReadMe-Datei wiedergegeben werden.

Vorlagen für die Erstellung von Metadaten finden Sie in folgenden Sammlungen:
- https://rd-alliance.github.io/metadata-directory
- https://dataverse.harvard.edu
- https://www.dcc.ac.uk/guidance/standards/metadata
- Metadata Standards Catalogue: https://rdamsc.bath.ac.uk

ReadMe Dateien können Sie z.B. auch mit folgenden Vorlagen erstellen:
- https://www.library.gatech.edu/smartech-metadata
- https://data.research.cornell.edu/content/readme
- https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

Persistente Identifikatoren

Ein persistenter Identifikator ist ein permanenter und eindeutiger Verweis (Link) auf ein digitales Objekt, unabhängig von Änderungen des (Online-)Standorts dieses Objekts. Die Dienste, die einen solchen Verweis bereitstellen, werden "Resolver-Dienste" genannt, einer davon ist https://doi.org. Durch Anhängen eines DOI an diese URL (https://doi.org) wird eine weitere URL erstellt (z.B. https://doi.org/10.5281/zenodo.422135), die auf das digitale Objekt verweist. Persistente Identifikatoren können nicht nur für Datensätze und Publikationen verwendet werden, sondern auch um Personen eindeutig zu referenzieren (z.B. mit ORCiD ID: https://orcid.org). Zudem erleichtern sie auch die Zitierbarkeit von Datensätzen:

> **Beispiel**: Hanigan, Ivan (2012): Monthly drought data for Australia 1890-2008 using the Hutchinson Drought Index. The Australian National University Australian Data Archive. http://doi.org/10.4225/13/50BBFD7E6727A

Tipp: Publizieren Sie Ihren Datensatz auf einem Repositorium, welches DOIs oder andere persistente Identifikatoren vergibt.

Offene Formate

Ein Dateiformat ist eine Standardmethode zur Codierung von Daten für die Speicherung in einer Computerdatei. Dateiformate können entweder proprietär oder frei, und entweder unveröffentlicht oder offen sein. Die von Ihnen verwendeten Dateiformate wirken sich direkt auf Ihre Fähigkeit aus, diese Dateien zu einem späteren Zeitpunkt zu öffnen und auf die Fähigkeit anderer Personen, auf diese Daten zuzugreifen.

Bei der Auswahl von Dateiformaten für die Archivierung sollten die Formate idealerweise offen sein, d.h. *nicht proprietär* (Benutzer zahlen nichts für die Nutzung), *in der Forschungsgemeinschaft allgemein gebräuchlich*, interoperabel zwischen verschiedenen Plattformen und Anwendungen, vollständig veröffentlicht und lizenzgebührenfrei verfügbar, und unabhängig von mehreren Softwareanbietern auf mehreren Plattformen implementierbar sein.

Folgende Dateiformaterweiterungen für die Wiederverwendbarkeit werden empfohlen:

Art der Daten	**geeignet**	**akzeptierbar**	**nicht geeignet**
Tabularische Daten mit vielen Metadaten	.csv / .hdf5	.txt / .html / .tex / .por	
Tabularische Daten mit wenig Metadaten	.csv / .tab / .ods / SQL	.xml if appropriate DTD / .xlsx	.xls / .xlsb
Textdaten	.pdf / .txt / .odt / .odm / .tex / .md / .htm / .xml	.pptx / .pdf with embedded forms / .rtf	.doc / .ppt
Code	.m / .R / .py / .iypnb / .rstudio / .rmd / NetCDF	.sdd	.mat / .rdata
Digitale Bilddatei	.tif / .png / .svg / .jpeg	.jpg / .jp2 / .tiff / .pdf / .gif / .bmp	.indd / .ait / .psd
Digitale Audiodatei	.flac / .wav / .ogg	.mp3 / .mp4 / .aif	
Digitale Videodatei	.mp4 / .mj2 / .avi / .mkv	.ogm / .webm	.wmv / .mov
Geografische Daten	NetCDF, tabular GIS attribute data, .shp / .shx / .dbf / .prj / .sbx / .sbn / PostGIS / .tif / .tfw / GeoJSON	.mdb / .mif	
CAD / Vektor- und Rasterdaten	.x3d / .x3dv / .x3db / PDF3D .pdf	.dwg / .dxf	
Generische Daten	.xml / .json / .rdf		

Tipp: Verwenden Sie ein offenes Format für Ihre publizierten Daten/Codes, auch wenn Sie während des Arbeitsprozesses mit proprietärer Software arbeiten.

Kontrollierte Vokabulare (controlled vocabularies)

Kontrollierte Vokabulare sind Listen mit vordefinierten, autorisierten Begriffen, die in einer Disziplin gebräuchlich und anerkannt sind. Zusätzlich zur Verwendung eines Metadatenstandards möchten Sie vielleicht kontrollierte Vokabulare für die Benennung Ihrer Variablen verwenden. Kontrollierte Vokabulare finden Sie unter http://fairsharing.org/standards (suchen Sie nach Fachgebiet).

Lizenzen

Wenn Sie das Urheberrecht an Ihren Daten besitzen, können Sie mit einer Lizenz informieren, wie Ihre Daten nachgenutzt werden dürfen. Lizenzen sind ein Rechtsinstrument, das es den Nutzenden erlaubt, Dinge mit den Daten zu tun, die andernfalls Ihre Rechte als Urheberrechtsinhaber verletzen würden. Um eine Lizenz auf Ihre Arbeit anzuwenden, können Sie eine Standardlizenz für Datensätze wählen, wie z.B. die Creative-Commons-Lizenzen (https://creativecommons.org/choose; Achtung: Version 4.0 International verwenden für Datensätze) oder die Open-Data-Commons-Lizenzen (https://opendatacommons.org).

Tipp: Publizieren Sie Ihren Datensatz mit einer offenen Lizenz, wie z.B. CC0 oder CC-BY, und geben Sie an, wie Ihr Datensatz referenziert werden soll (falls dies nicht schon vom Repositorium automatisiert wird).

Falls das Repositorium, in dem Sie Ihre Daten hochladen, nicht bereits eine Vorauswahl von Lizenzen zur Verfügung stellt, können Sie die Lizenz zum Beispiel in der README Datei, prominent auf der Download-Seite oder in einer separaten «Lizenz.txt»-Datei zur Verfügung stellen. Wichtig: Der Verweis muss den Namen der Lizenz und den Link zum offiziellen Lizenztext enthalten.

> **Beispiel**: Dieser {DATEN(SATZ)-NAME} ist mit einer CC-BY 4.0 Internationalen Lizenz (https://creativecommons.org/licenses/by/4.0) lizenziert.

Wie FAIR sind meine Daten?

Benutzen Sie dafür die folgenden Online-Tools:

- DANS Tool : https://satifyd.dans.knaw.nl
- Australian National Data Service: https://www.ands-nectar-rds.org.au/fair-tool
- ARDC (2018) FAIR self-assessment tool, Australian Research Data Commons: https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/

Referenzen

Blumer et al. (2019). EPFL Library Research Data Management Fastguides. http://infoscience.epfl.ch/record/265349. Published with a CC-BY-NC-SA license

SNSF (2021). Explanation of the FAIR data principles. http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

Standford Libraries. Data best practices and case studies: https://guides.library.stanford.edu/data-best-practices/format-files

Top 10 FAIR Data and software things: https://librarycarpentry.org/Top-10-FAIR

Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18.