Decision making procedures in data management and data stewardship for Open Science

Connie Clare, PhD
Community Development Manager
Research Data Alliance
connie.clare@rda-foundation.org

EOSC Symposium 2022
16 November 2022
doi: 10.5281/zenodo.7326280

An overview

Research Data Management*

**Standard/best practices for accurate data/code collection, processing, documentation, analysis, storage & preservation as a prerequisite for open science (FAIR ≠ Open).*

- **What decisions do researchers make to achieve 'FAIR' data management?**
- **Data-centric AI and data stewardship challenges**

What decisions do researchers make to achieve 'FAIR' data management?

Who is responsible for data management?

How will new data be collected or produced and/or how will existing data be reused?

What data types, formats, and volumes will be collected or produced?
Structured data, formats: JSON, XML, CSV

What metadata and documentation will accompany data?

What resources will be dedicated to data management and ensuring that data will be FAIR?

How, when and which will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Are there disciplinary standards and vocabularies that should be used?

What data quality control measures will be used?

If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

What methods or software tools will be needed to access and use the data?

Will the application of a unique persistent identifier (e.g., DOI) be assigned to the dataset?

Is an ethical review (HREC, ERB) required?

Is informed consent required?

Does data need to be anonymised or pseudonymised?

Which license will be suitable to specify data modification, redistribution and reuse?

How will data for preservation be selected? Where will data be preserved long-term?

How will data and metadata be stored and backed up during the research process?

How will other legal issues, such as IPR and ownership, be managed? What legislation is applicable?

How will possible ethical issues be taken into account, and codes of conduct followed?

FAIR principles focus on machine-actionability for automated discoverability of data…

In 2016, the '**FAIR Guiding Principles for scientific data management and stewardship'** were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

A practical "how to" guidance to go FAIR can be found in the **Three-point FAIRification Framework**.

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

https://www.go-fair.org/fair-principles/

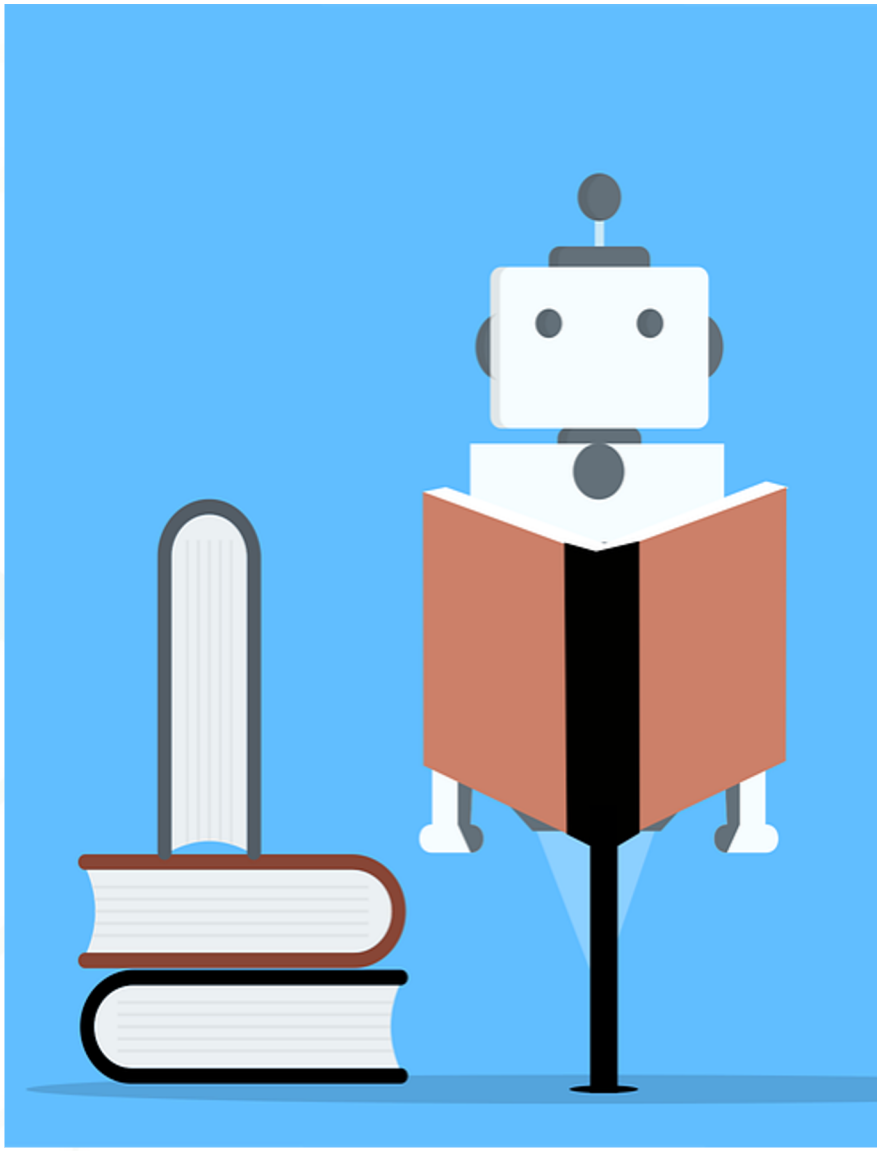
Data-centric AI

Automated decision making using data.

Data is fundamental for training and deploying AI models.

Data management and/or curation is a crucial step to feed into AI model.

'Machine learning models are only as good as the data they're trained on' - https://fairmlbook.org/datasets.html (Chapter 8)



Data stewardship challenges & AI ethics

Black box AI - Model inputs and operations remain a mystery. Unknown input data provenance and quality. Automated data retrieval lead to inconsistent results.

AI bias due to generalisation (insufficient representative input data), or unsuitable data collection, processing (cleaning), quality, mislabelling and model design. Synthetic (output) data generated inherits and propagates bias affecting scientific validity.

Data misuse - Using data as input for an AI model that causes harm.

Lack of standards, tools and mechanisms to evaluate data quality and assess whether datasets are fit for purpose.

Thank you for your attention

Acknowledgements

- Nicolas Dintzner, Data Steward, TU Delft
- Santosh Ilamparuthi, Data Steward, TU Delft
- Shalini Kurapati, Co-Founder & CEO, Clearbox AI https://www.clearbox.ai/about