Principles of Data Management and Stewardship

Ken Baclawski
Ontolog Forum

7 December 2022

Principles to be (briefly) presented

- GDPR
- FAIR
- TRUST
- Metadata 2020
- Ten Commandments of Ethical Medical AI

General Data Protection Regulation (GDPR)

- Provisions and requirements related to the processing of personal data of individuals
- Regulation not a directive
- Adopted in 2016
- Became enforceable in 2018

FAIR

- Guidelines for digital assets to have:
    - Findability
    - Accessibility (under well defined conditions)
    - Interoperability
    - Reuse
- FAIR is not necessarily “fair.”
- https://www.go-fair.org/fair-principles/ in *Scientific Data* (2016)
- Intended for scientific data but could be adapted for other data sources and possibly more than just data

FAIR Principles

F1: (Meta) data are assigned globally unique and persistent identifiers
F2: **Data are described with rich metadata**
F3: Metadata clearly and explicitly include the identifier of the data they describe
F4: (Meta)data are registered or indexed in a searchable resource

A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
A1.1: The protocol is open, free and universally implementable
A1.2: The protocol allows for an authentication and authorisation procedure where necessary
A2: **Metadata should be accessible even when the data is no longer available**

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2: **(Meta)data use vocabularies that follow the FAIR principles**
I3: (Meta)data include qualified references to other (meta)data

R1: (Meta)data are richly described with a plurality of accurate and relevant attributes
R1.1: (Meta)data are released with a clear and accessible data usage license
R1.2: (Meta)data are associated with detailed provenance
R1.3: (Meta)data meet domain-relevant community standards

FAIR is not the same as Open

- FAIR allows for legitimate reasons to shield data
- FAIR explicitly and deliberately does not address moral and ethical issues pertaining to the openness of data
- FAIR only requires:
  - A process for accessing discovered data
  - An open and rich description of the context within which data were generated, to enable evaluation of its utility
  - Explicitly defining the conditions under which data may be reused
  - Providing clear instructions on how data should be cited when reused
  - Clarity and transparency around the conditions governing access and reuse
- FAIR data need not be Open, and Open data need not be FAIR.

Adherence to FAIR by Ontologies

- A review of disaster related ontologies found very little adherence to FAIR principles
- In https://mdpi.com/2220-9964/10/5/324 it was reported:
    - Only 1.4% of all retrieved ontologies are published in semantic repositories.
    - 84.1% are not published at all.

TRUST

- TRUST describes the characteristics of data repositories that are responsible for storing data over a long period of time.
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7224370/
- *Scientific Data* Vol. 7, Article 144 (2020)
- Concerned with **entrusting**: To give over (something) to another for care, protection, or performance.
- Does not guarantee trust but is intended to help earn trust.

TRUST Characteristics

- Transparency
  - Verifiable by publicly accessible evidence.
  - Explicitly declare terms of use, both for the repository and the data holdings
  - Specify minimum digital preservation timeframe for the data holdings
  - Declare any pertinent additional features or services, such as the capacity to responsibly steward sensitive data.
- Responsibility
  - Ensure the authenticity and integrity of the data holdings
  - Ensure the reliability and persistence of the services
  - Adhere to the designated community's metadata and curation standards
  - Manage the intellectual property rights of data producers
  - Protect sensitive information resources
  - Secure the system and its content

TRUST Characteristics

- User Focus
  - Ensure that the expectations of their target user communities are met
  - May implement relevant data metrics and make these available to users
  - May provide or contribute to community catalogues to facilitate data discovery
  - May monitor and identify evolving community expectations and respond as required
- Sustainability
  - Should sustain services and preserve data holdings for the long term
  - Plan for risk mitigation, business continuity, disaster recovery, and succession
  - Secure funding to enable ongoing usage
  - Maintain the desirable properties of the data resources
  - Provide governance for necessary long-term preservation of data so that data resources remain FAIR
- Technology
  - Provide infrastructure to ensure the first four characteristics

Metadata 2020

- Advocates richer, connected, and reusable, open metadata for all research outputs
- Principles based on existing best practices
- Complement and support FAIR data principles
- Inclusive of both data and metadata
- https://metadata2020.org/

Metadata 2020 Principles

- COMPATIBLE
  - Provide a guide to content for machines and people
  - Metadata must be as open, interoperable, parsable, machine actionable, human readable as possible.
- COMPLETE
  - Reflect the content, components and relationships as published
  - Metadata must be as complete and comprehensive as possible.
- CREDIBLE
  - Enable content discoverability and longevity
  - Metadata must be of clear provenance, trustworthy and accurate.
- CURATED
  - Reflect updates and new elements
  - Metadata must be maintained over time.

Metadata 2020 Collaboration Outputs

- Guidance
  - Metadata Principles (See previous slide)
  - Metadata Personas: Creators, Custodians, Curators, Consumers
  - Metadata Practices
- Understanding
  - Metadata Best Practices
  - Metadata Use Cases
  - Metadata Attitudes and Understandings
  - Metadata Literature Review

Ethical Medical AI

- Ten Commandments published in 2021
- https://ieeexplore.ieee.org/document/9473208
- Practical guidelines for those applying artificial intelligence
- Stating the guidelines as “commandments” is awkward compared with FAIR principles which allow for more nuanced conformance.

The First Five Commandments

1 It must be recognizable that and which part of a **decision** or action is taken and carried out by AI.

2 It must be recognizable which part of the **communication** is performed by an AI agent.

3 The **responsibility** for an AI decision, action, or communicative process must be taken by a competent physical or legal person.

4 AI decisions, actions, and communicative processes must be transparent and **explainable**.

5 An AI decision must be comprehensible and **repeatable**. [emphasis added]

The Second Five Commandments

6 An explanation of an AI decision must be based on state-of-the-art (**scientific**) theories.

7 An AI decision, action, or communication **must not be manipulative** by pretending accuracy.

8 An AI decision, action, or communication **must not violate any applicable law** and must not lead to human harm.

9 An AI decision, action, or communication **shall not be discriminatory**. This applies in particular to the training of algorithms.

10 The target setting, control, and monitoring of AI decisions, actions, and communications **shall not be performed by algorithms**. [emphasis added]

Other Commandments of Data

- Using biblical language for manifestos and commandments has a long history for computer data.
- The Object Oriented Database System Manifesto (1989) https://bit.ly/3FkOPwL is one of the earliest.
  - It uses the biblical language of the Ten Commandments.
  - It was rebutted by a subsequent manifesto.
- The next slide has some examples of “Ten Commandments” for data that are easily found by a web search.

More Ten Commandments

- Data Science: The 10 Commandments for Performing a Data Science Project https://bit.ly/3UqXwKr
- The Ten Commandments Of Data Visualization https://bit.ly/3Fl0eNd
- The Ten Commandments of Data Science Project Execution https://bit.ly/3iqbvCN
- The Ten Commandments for Divine Data Quality https://bit.ly/3XU5KgP
- The Ten Commandments of Data Collection https://bit.ly/3EYRKty
- Ten Commandments of Data Usage https://bit.ly/3VLoVaY
- The 10 Commandments of Data Security and Data Management https://bit.ly/3EYRPxm
- 10++ Commandments of Data Science Modeling https://bit.ly/3FlzkVk

Research Opportunities

- Adapt the FAIR Principles for scientific simulations
  - Simulations are increasingly common and important but are often unavailable and unrepeatable.
- Adapt the Ten Commandments for Ethical Medical AI for other domains (not necessarily using the “Ten Commandments format”) and recommend improvements.
- Survey major ontologies for adherence to FAIR principles, TRUST characteristics and Metadata 2020 principles and recommend improvements.