FAIR Re-use: Implications for AI-Readiness

Lydia Fletcher
Texas Advanced Computing Center
April 16, 2024

TACC

iHARP
NSF HDR INSTITUTE

Garbage In, Garbage Out

- The quality of output from a model is directly dependent on the quality of input
- High-quality, well-curated datasets are critical for AI-readiness even if the datasets aren't "AI-ready"

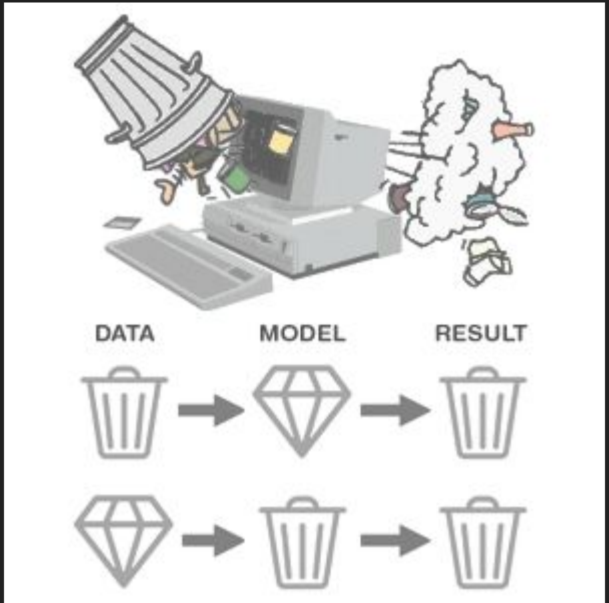

Image credit: https://thedailyomnivore.net/2015/12/02/garbage-in-garbage-out/

How can FAIR help with “Garbage In, Garbage Out”?

Findable → High quality datasets are easy to locate using GOFAIRUS, fairsharing.org or other resources.

Accessible → Datasets are available to download for training, benchmarking, and validation of models.

Interoperable → Well described provenance in the form of metadata and documentation, as well as availability in AI-ready formats.

Reusable → Following FAIR principles makes digital objects reusable and also act as guidelines for reusing data.

FAIR vs AI-ready

- FAIRified datasets are not inherently AI-ready and may not be fit for purpose for a particular AI/ML model or need additional processing, transformation, etc. to prepare them for use in modern computing environments such as Tensorflow or PyTorch
- But researchers can use the FAIR data principles and emerging FAIR for AI/ML principles as guidelines for choosing which data or models to reuse for their research.

A closer look at Reusability

> To achieve FAIR Reusability, these requirements must be met:
>
> 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

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

- Reusability emphasizes creating data in a format and structure that facilitates easy interpretation and reuse, even in contexts beyond the purpose for which it was originally collected.
- However, reusability is not just about making your research objects reusable, it's knowing how to responsibly reuse others' research objects.

What does Reuse really mean?

- Understanding Provenance
  - Where did it come from? How has it been cleaned or manipulated?
- Developing Search Skills
  - Researchers often default to using datasets they are already aware of or encouraged to used by collaborators.
  - Without learning data literacy skills such as how to locate datasets or how to evaluate them, researchers are still at risk of putting garbage into their models.
- Becoming Familiar With Disciplinary Repositories

How FAIR principles can be used as guidelines

- Understand Data Context
  - Understand the context in which the data was collected, including the methodology, sampling techniques, and any limitations or biases inherent in the dataset. Where did the data come from? Who collected it? How? What has happened to it since it was collected?
- Document Your Usage
  - Distinguish whether you used a dataset to train a model or as input for predictions or analysis
  - Clearly describe any processing/transformations you do to the data
- Engage in Transparent Research
  - Be transparent about your research methods, analysis techniques, and data sources to promote reproducibility and accountability
- Contribute to the Open Data Ecosystem
  - Share findings or derived datasets with the community by contributing to data repositories or collaborative projects to increase the amount of open data available

Related work:
Improving Traceability Throughout the Data Lifecycle
https://doi.org/10.26153/tsw/49483

A key component to Reuse is licensing

- Acknowledge Sources
    - Properly acknowledge the source of the data by citing the original dataset and providing appropriate attribution to the data provider
- Respect Terms of Use and/or Restrictions
    - Lots of data is public domain, but not all of it
    - When using data from a private-public partnership it’s essential to understand restrictions on publication
- Respect Privacy and Confidentiality
    - Ensure compliance with privacy regulations and ethical standards when working with sensitive or personally identifiable information

Challenges

- Time
  - Everyone is busy all the time
- Data Literacy Education
  - "Trust" for a data source plays a big deal in people's assumptions that data is "good"
  - Much of this is based on reputation but there needs to be more education around dataset assessment
- Assessing Data Quality and Consistency
  - Verifying that datasets adhere to FAIR principles can be challenging and time-consuming
- Ensuring Data Integration and Interoperability
  - Integrating diverse datasets from different sources while maintaining interoperability can be a complex process

Opportunities

- Enhanced Collaboration and Innovation
  - Acknowledging sources and contributing back increases data available and builds communities
- Improved Model Performance and Replicability
  - Access to high-quality, diverse datasets following FAIR principles can improve AI/ML model performance
- Increased Transparency and Accountability
  - More transparency creates more trust in AI/ML model outputs
- Extending FAIR Principles
  - Work is already being done to extend the FAIR principles to address AI/ML models and software – such as the FAIR4HEP initiative and the work of FARR

Contact info:

lfletcher@tacc.utexas.edu