



FAIR Data for AI  
University of Florida



**Parent Project title:** GatorSTAR: A New MARC U\*STAR Program at the University of Florida

**Working title:** Adding a FAIR Data Practices Curriculum to UF's Practicum AI AI/ML training workshops

**PI:** David Julian

**Additional team members:** Matt Gitzendanner, Hao Ye, Yulia Strekalova

# Primary Audience

---

- Undergraduate students participating in biomedical research
- Small discussion groups
- Experiential learning



# Three 1-hour modules

1. Data organization in spreadsheets
  - Challenges with spreadsheets
2. Data availability in repositories
  - Journal article data accessibility, metadata, and ontologies
3. Data repositories
  - Finding shared data

<https://practicumai.org/courses/FAIR/>

# Student Activities

1. Data organization in spreadsheets
  - Examine sample data in spreadsheet files
  - Discuss issues and share possible solutions
2. Data availability in repositories
  - Look for data shared alongside publications in PubMed
  - Check for DOI associated with the data
3. Data repositories
  - Look for data in a specialist data repository
  - Discuss ease of finding shared data to reuse in a secondary analysis in particular, metadata that would be effective

# Module 1

*Background:* A professor was interested in measuring the effect of exercise on heart rate. For this project, students were assigned to record their resting pulse rate and then either run for one minute or sit for one minute and then record their pulse rate again.

...

This experiment was performed over three years and the data files for each year are available to you here:

[1993\\_data.xlsx](#)

[1996 Exercise Data.xlsx](#)

[Run\\_Sit\\_data\\_95.xlsx](#)

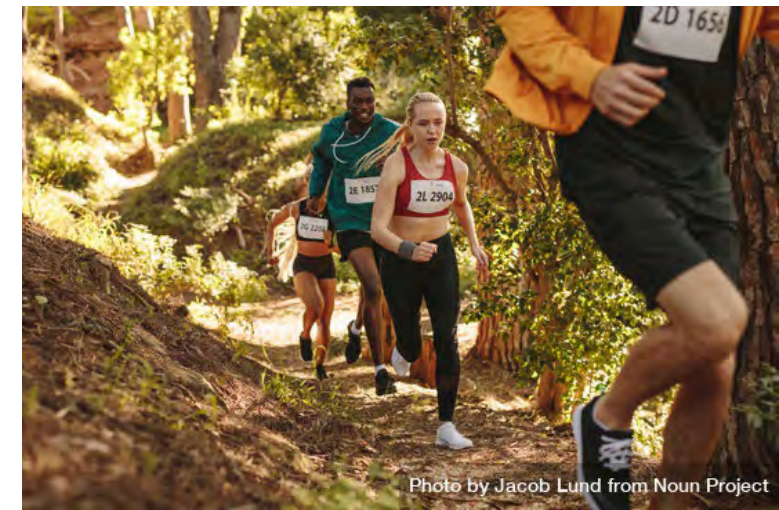


Photo by Jacob Lund from Noun Project

# FAIR Data in AI/ML: Exercise 2

## Student Instructions

After the last exercise where you worked to compile data on the effects of running on heart rate, you have started to think that you might want to do more research on the overall effects of exercise on fitness. Perhaps, this research will lead you to launch your company's new line of fitness trackers!!

As such, you decide to start looking through the existing data on the effects of different exercises on fitness. There is already a fair bit of data published on this, so why not make use of those data?

**Searching the literature for published datasets**



# Module 2

## FAIR concepts

**F**indable **A**ccessible **I**nteroperable **R**eusable

**practicum**  
building ai knowledge

**FAIR Principles**

“There is an original scientific experiment that is irrefutably supporting the concept of scientific value.”

“Therefore, there are your scientific data, their context, the way they were generated, and the way they are being used, all of which are important to understanding the value of the data, its usability for other researchers, and its potential for reuse.”

**FAIR** **SCIENTIFIC DATA**

FAIR Principles for scientific data: Findable, Accessible, Interoperable, and Reusable

**F**indable

- The first step is (including 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.

**Metadata**  
Data about data

**Metadata**

Information about information:

- Location
- Who collected the data
- How the data was collected
- Date and time
- Units, scales, resolution
- Relationships among multiple data files

**Findable**

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

**F**indable

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

**A**ccessible

Once the user finds the required data, s/he then needs to know how they can be accessed, possibly involving authentication and authorization.

- FAIR is not the same as Open Data
- Rather, they both should provide the exact conditions under which the data are accessible, access when necessary, and how the data can be used.

**A**ccessible

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

**I**nteroperable

- The data usually need to be integrated with other data.
- In addition, the data need to be interoperable with applications or software for analysis, storage, and processing.

**I**nteroperable

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

**Ontologies**

The Open Biological and Biomedical Ontology (OBO) Foundry

Cooperatively development of interoperable ontologies for the biological sciences

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

**Reusable**

Cardiovascular Disease Ontology (CDO) <http://www.ebi.ac.uk/ontology-lookup/>

```
[Name]
id: OBO:0000010
name: cardiovascular_disease
is_a: OBO:0000001 | pathologic | health | disease
is_a: OBO:0000002 | health | disease
is_a: OBO:0000003 | health | disease
is_a: OBO:0000004 | health | disease
is_a: OBO:0000005 | health | disease
is_a: OBO:0000006 | health | disease
is_a: OBO:0000007 | health | disease
is_a: OBO:0000008 | health | disease
is_a: OBO:0000009 | health | disease
is_a: OBO:0000010 | health | disease
is_a: OBO:0000011 | health | disease
is_a: OBO:0000012 | health | disease
is_a: OBO:0000013 | health | disease
is_a: OBO:0000014 | health | disease
is_a: OBO:0000015 | health | disease
is_a: OBO:0000016 | health | disease
is_a: OBO:0000017 | health | disease
is_a: OBO:0000018 | health | disease
is_a: OBO:0000019 | health | disease
is_a: OBO:0000020 | health | disease
is_a: OBO:0000021 | health | disease
is_a: OBO:0000022 | health | disease
is_a: OBO:0000023 | health | disease
is_a: OBO:0000024 | health | disease
is_a: OBO:0000025 | health | disease
is_a: OBO:0000026 | health | disease
is_a: OBO:0000027 | health | disease
is_a: OBO:0000028 | health | disease
is_a: OBO:0000029 | health | disease
is_a: OBO:0000030 | health | disease
is_a: OBO:0000031 | health | disease
is_a: OBO:0000032 | health | disease
is_a: OBO:0000033 | health | disease
is_a: OBO:0000034 | health | disease
is_a: OBO:0000035 | health | disease
is_a: OBO:0000036 | health | disease
is_a: OBO:0000037 | health | disease
is_a: OBO:0000038 | health | disease
is_a: OBO:0000039 | health | disease
is_a: OBO:0000040 | health | disease
is_a: OBO:0000041 | health | disease
is_a: OBO:0000042 | health | disease
is_a: OBO:0000043 | health | disease
is_a: OBO:0000044 | health | disease
is_a: OBO:0000045 | health | disease
is_a: OBO:0000046 | health | disease
is_a: OBO:0000047 | health | disease
is_a: OBO:0000048 | health | disease
is_a: OBO:0000049 | health | disease
is_a: OBO:0000050 | health | disease
is_a: OBO:0000051 | health | disease
is_a: OBO:0000052 | health | disease
is_a: OBO:0000053 | health | disease
is_a: OBO:0000054 | health | disease
is_a: OBO:0000055 | health | disease
is_a: OBO:0000056 | health | disease
is_a: OBO:0000057 | health | disease
is_a: OBO:0000058 | health | disease
is_a: OBO:0000059 | health | disease
is_a: OBO:0000060 | health | disease
is_a: OBO:0000061 | health | disease
is_a: OBO:0000062 | health | disease
is_a: OBO:0000063 | health | disease
is_a: OBO:0000064 | health | disease
is_a: OBO:0000065 | health | disease
is_a: OBO:0000066 | health | disease
is_a: OBO:0000067 | health | disease
is_a: OBO:0000068 | health | disease
is_a: OBO:0000069 | health | disease
is_a: OBO:0000070 | health | disease
is_a: OBO:0000071 | health | disease
is_a: OBO:0000072 | health | disease
is_a: OBO:0000073 | health | disease
is_a: OBO:0000074 | health | disease
is_a: OBO:0000075 | health | disease
is_a: OBO:0000076 | health | disease
is_a: OBO:0000077 | health | disease
is_a: OBO:0000078 | health | disease
is_a: OBO:0000079 | health | disease
is_a: OBO:0000080 | health | disease
is_a: OBO:0000081 | health | disease
is_a: OBO:0000082 | health | disease
is_a: OBO:0000083 | health | disease
is_a: OBO:0000084 | health | disease
is_a: OBO:0000085 | health | disease
is_a: OBO:0000086 | health | disease
is_a: OBO:0000087 | health | disease
is_a: OBO:0000088 | health | disease
is_a: OBO:0000089 | health | disease
is_a: OBO:0000090 | health | disease
is_a: OBO:0000091 | health | disease
is_a: OBO:0000092 | health | disease
is_a: OBO:0000093 | health | disease
is_a: OBO:0000094 | health | disease
is_a: OBO:0000095 | health | disease
is_a: OBO:0000096 | health | disease
is_a: OBO:0000097 | health | disease
is_a: OBO:0000098 | health | disease
is_a: OBO:0000099 | health | disease
is_a: OBO:0000100 | health | disease
```

**R**eusable

- The ultimate goal of FAIR is to optimize the reuse of data.
- To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

**R**eusable

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

**R**eusable

1. Use data identifiers (e.g., URIs) to uniquely identify data.

2. Use persistent identifiers (e.g., DOIs) to ensure that the identifiers remain valid over time.

3. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.

4. Use machine-readable identifiers (e.g., URIs) to enable automatic discovery of data.



# FAIR Data in AI/ML: Exercise 3

In [exercise 1](#), we explored common issues with how data are organized in spreadsheets. We also provided a [handout](#) and the [Broman and Woo 2018 paper](#) with some best practices in organizing data in spreadsheets.



Then, in [exercise 2](#), we explored the challenges with finding data associated with published literature and introduced the [FAIR principles](#).

Now we turn to the process of making data available in data repositories.

## Data repositories

### About Biomedical Data Repositories and Knowledgebases



Accessible, well-maintained, and efficiently operated data resources are critical enablers of modern biomedical research. Data resources, through good data management practices, are the key to data and knowledge discovery, integration, and data reuse, as outlined by the [FAIR Data Principles](#)<sup>®</sup>. To better support such a modern data resource

ecosystem, NIH makes a distinction between data repositories and knowledgebases. While each activity is important for advancing biomedical research, data repositories and knowledgebases can have unique functions, metrics for success and sustainability needs.



# Module 3

## FAIR, RDM, Open

**F**indable **A**ccessible **I**nteroperable **R**eusable

**Plan** → **Collect** → **Process** → **Analyze** → **Preserve** → **Share** → **Reuse** → **Plan**

**FAIR, RDM, and Open**

**practicum**  
building ai knowledge

Not all or nothing

FAIR = Open

**Research Data Management (RDM)**

RDM "can be defined as a set of practices to handle information collected and created during research"

"Data management in contrast to data handling" "It is seen as mental grunt work that people doing the data do but do not particularly value or engage in"

"RDM is the backbone of data that has not been properly created and managed during the early stages of research. It will be very difficult to clean that up or open"

Higman, Bangert and Jones, 2019

**FAIR**

Emphasis on both human and machine accessibility

**F**indable **A**ccessible **I**nteroperable **R**eusable

**Open Data and Open Science**

"...open data has increasingly become an expectation of funders and participating, often framed by the notion of 'as open as possible, as closed as necessary'. Open data can be defined on a continuum, for instance ... a minimum requirement of open data is to have an open license (such as Creative Commons CC), but to achieve greater openness and reuse potential, data should also be machine-readable, in a non-proprietary format, use open standards and link to other data to provide context"

Higman, Bangert and Jones, 2019

**Misconceptions highlighted by Higman, Bangert and Jones, 2019**

1. FAIR data has to be open
2. Open data is more useful than FAIR data
3. All FAIR and open data is of good quality
4. FAIR is limited to the EU and the life sciences - why should I care?

**Relationship between RDM, FAIR and open**

(From Higman, Bangert and Jones, 2019)

FAIR	Open	FAIR + Open
<ul style="list-style-type: none"> <li>Findable: Metadata, Persistent ID, Searchable</li> <li>Accessible: Accessible, Machine-readable</li> <li>Interoperable: Standardized, Linked</li> <li>Reusable: Attribution, License, Versioning</li> </ul>	<ul style="list-style-type: none"> <li>Open License</li> <li>Machine-readable</li> <li>Non-proprietary format</li> <li>Open standards</li> <li>Link to other data</li> </ul>	<ul style="list-style-type: none"> <li>Findable: Metadata, Persistent ID, Searchable</li> <li>Accessible: Accessible, Machine-readable</li> <li>Interoperable: Standardized, Linked</li> <li>Reusable: Attribution, License, Versioning</li> <li>Open License</li> <li>Machine-readable</li> <li>Non-proprietary format</li> <li>Open standards</li> <li>Link to other data</li> </ul>

