Report from FAIR Train WG

OLGA WODO (UNIVERSITY AT UNIVERSITY)

BRIAN SCHUSTER (UNIVERSITY AT TEXAS EL PASO)

ZACHARY TRAUTT (NIST)

LOGAN WARD (ARGONNE NATIONAL LAB)

KATHRYN KNIGHT (GOFAIR US, ORNL)

ARUN KUMAR MANNODI KANAKKITHODI (PURDUE UNIVERSITY)

ERIC TOBERER (COLORADO SCHOOL OF MINES)

https://github.com/marda-alliance/FAIRtrain

MaRDA 2024 annual meeting: 2/21/24

MaRDA | Materials Research Data Alliance

NIST

NSF

MaRCN

Where to find information?

Motivation and scope:

- FAIR principles are fundamental for exchanging scientific information – materials science community MGI 2021 document.
- To fuel the cultural changes, the FAIR principles must be integrated curriculum.
- WG aims to define a set of teaching modules

Mode of operation: 5 meetings (3 months)

https://github.com/marda-alliance/FAIRtrain

MaRDA Materials Research Data Alliance

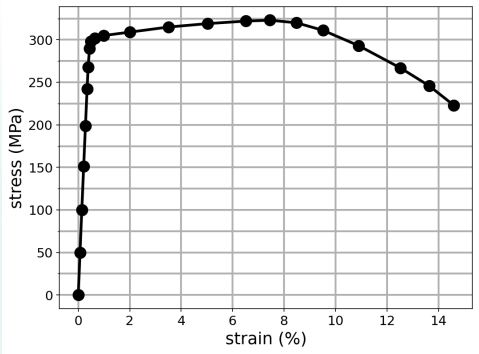
NIST

NSF

MaRCN

Motivation – fundamentals of FAIR using fundamental mechanical test

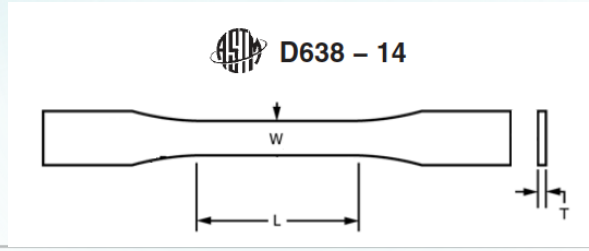
Goal: understand what it takes to make data FAIR

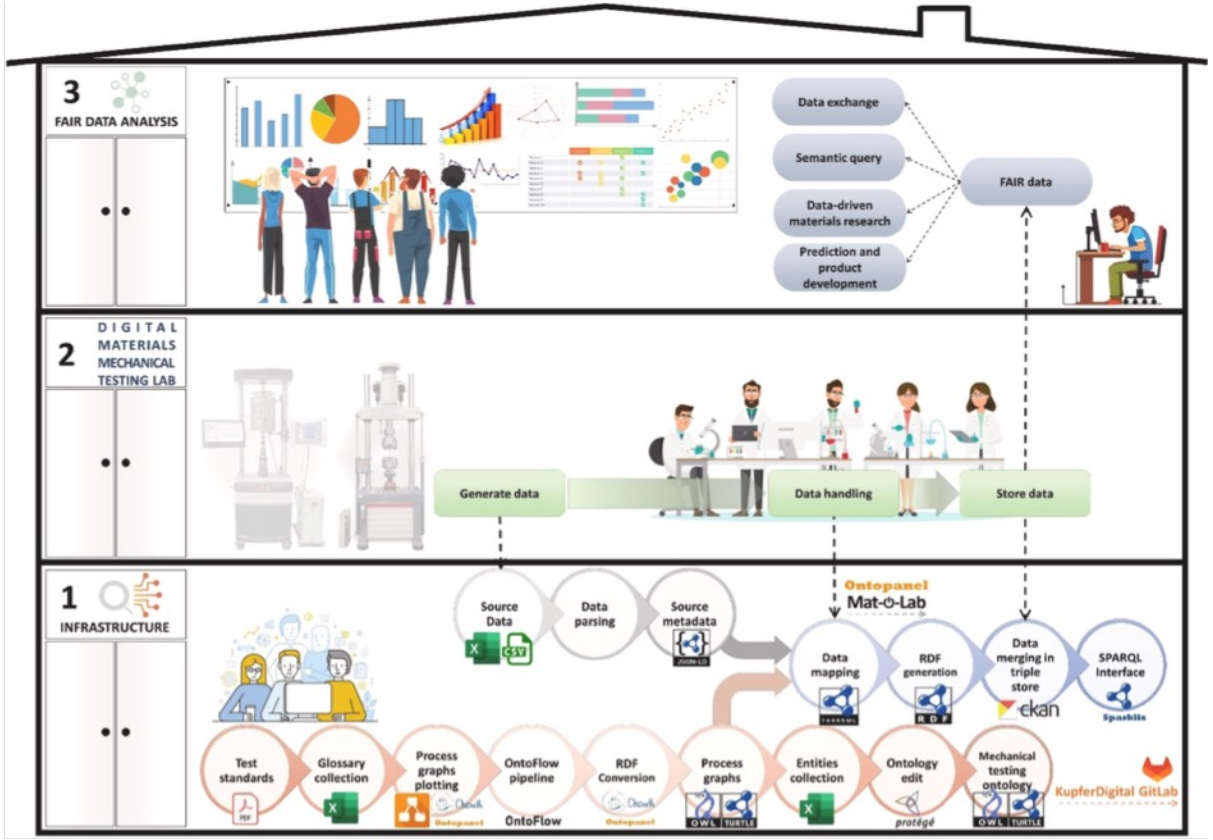


Source data (table)

Plot

Extract property using standard





Computers in Industry 153 (2023) 104016

MaRDA Materials Research Data Alliance

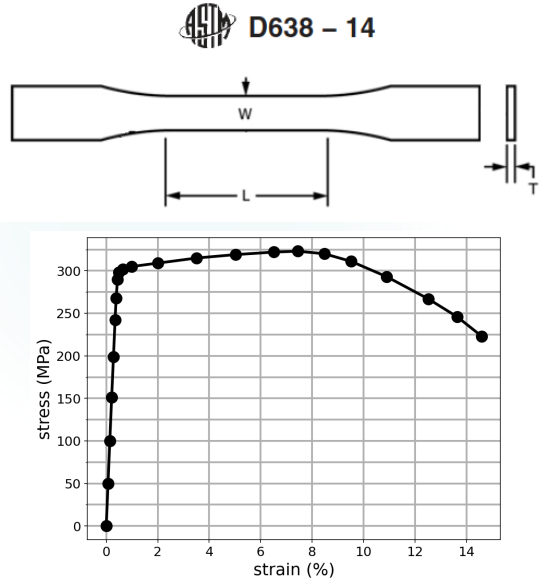
NIST

NSF

MaRCN

Example module – fundamentals of FAIR using fundamental mechanical test

- Load data from previous years, without any metadata
- Generate new data, publish data, share with other students, integrate
- Analyze data, publish the code to the repository
- Design the vocabulary and document both data and codes
- Annotate data using ontology
- Standards (ISO ASTM)



A2.23 *yield strength*—the stress at which a material exhibits a specified limiting deviation from the proportionality of stress to strain. Unless otherwise specified, this stress will be the stress at the yield point and when expressed in relation to the tensile strength shall be designated either tensile strength at yield or tensile stress at yield as required in A2.18 (Fig. A2.3). (See *offset yield strength.*)

MaRDA | Materials Research Data Alliance

NIST

NSF

MaRCN

FAIR train modules

M1: Data reuse

- L1: Work with data from prior work or last year: load data, and visualize using the script (python/matplotlib, gnuplot) – write notebook to capture the workflow
- L2: Use REST API (Materials Project, Citirination, AWFLOW)
- L3: Learn data scraping and data restructuring (e.g., Beautiful Soup for XML and HTML documents in Python, rvest and xml2 in R)

M2: Data integration

- L1: Data cleaning
- L2: Data integration: metadata standardization challenges (dealing with synonymous and homonymous terms, singular/plural word forms, lexical/dialectical variants...)
- L3: Knowledge representation: data + metadata + rules - adding semantics to structure and syntax: Resource Description Framework (RDF) and schema (RDFS), database and ontology

M3: FAIR data generation

- L1: Generate data with metadata
- L2: Generate data with metadata and annotate with existing ontology/vocabulary
- L3: Prepare FAIR publication (publication space: zenodo, MDF), or plan for effortless and reproducible work: describe a computational/experimental environment, provide notebook as an essay/journal article, plan for incremental work.

Potential resources:
https://www.writethedocs.org/
https://github.com/marda-alliance/FAIR_2023_Workshop/

tool OpenRefine
https://librarycarpentry.org/lc-open-refine/01-introduction.html

YAMZ https://www.yamz.net/
https://doi.org/10.1162/dint_a_00211

schema
https://pages.nist.gov/material-schema/
Manufacturing ontologies:
https://matportal.org/
https://terminology.nfdi4ing.de/ts/
Mechanical testing ontologies:
https://doi.org/10.1016/j.compind.2023.104016

FAIR train modules

M+: Software Carpentry and Data Management Related Tools:
programming, version control, unix shell for automating tasks, database tools for programmatic data access interaction and lab book for documenting science

M0: World beyond spreadsheets and metadata in filenames
L1: data formats for machine-readable data (CSV, JSON, XML for materials data)
L2: Introduction to knowledge representation: data + metadata - introducing structure and syntax into the data or how to make the data machine-actionable through self-describing key/value data structure (e.g., XML, JSON with semantically relevant key names)
L3: Prepare your first dataset and code repository (git or zenodo, jupyterlab vs notebook) and pay attention to documentation
L4: Verify if data is well structured for correctness - schema and schema validators

M1: Data reuse
- L1: Work with data from prior work
- L2: Use REST API
- L3: Learn data scraping and data restructuring

M2: Data integration
- L1: Data cleaning
- L2: Data integration
- L3: Knowledge representation

M3: FAIR data generation
- L1: Generate data with metadata
- L2: Generate data with metadata and annotate with existing ontology/vocabulary
- L3: Prepare FAIR publication

Potential resources:
https://www.writethedocs.org/
https://github.com/marda-alliance/FAIR_2023_Workshop/

tool OpenRefine
https://librarycarpentry.org/lc-open-refine/01-introduction.html

YAMZ https://www.yamz.net/
https://doi.org/10.1162/dint_a_00211

https://pages.nist.gov/material-schema/
Manufacturing ontologies:
https://matportal.org/
https://terminology.nfdi4ing.de/ts/
Mechanical testing ontologies:
https://doi.org/10.1016/j.compind.2023.104016

MaRDA | Materials Research Data Alliance
NIST
NSF
MaRCN

FAIR train – next steps

- Need to expand on materials science aspects of the modules
- FAIR Train Working Group – expand modules

- Training-the trainers workshop with special emphasis on the outreach to community colleges and MSI (workshop at UTEP summer 2025), looking for participants

How to contribute to WG?

just contact us (olgawodo@buffalo.edu or anyone at MaRDA)

https://github.com/marda-alliance/FAIRtrain

MaRDA Materials Research Data Alliance

NIST

NSF

MaRCN

We need your input

- Materials science FAIR resources:
  - It can be anything as large as workshop or as small as one class (e.g., https://github.com/marda-alliance/FAIR_2023_Workshop/)
  - …
- Materials science datasets that are FAIR
  - Ideally datasets with different FAIR maturity (for educational purposes)
  - …
- Materials science centric FAIR Tools
  - Sim2Ls: FAIR simulation workflows and data (Nanohub)
  - ..