

The background of the slide features a complex molecular structure. It consists of numerous atoms represented by spheres in white, red, blue, and yellow, connected by bonds. The structure is layered, with some parts appearing more prominent than others, creating a sense of depth. The overall color palette is dominated by the colors of the spheres and the dark grey background.

Wang Materials Group

Group Meeting  
May 22, 2023

Data Archiving and Curation

# Agenda

- Why bother archiving and curating data
- Sharing strategies for archiving
- Guidelines for archiving
- Example case for a journal publication

# Why bother archiving data

- The big data analytics market is set to reach \$103 billion by 2023.
- Poor data quality costs the US economy up to \$3.1 trillion yearly.
- In 2020, every person generated 1.7 megabytes in just a second.
- The world will produce slightly over 180 zettabytes ( $10^{21}$ ) of data by 2025.
- 4<sup>th</sup> paradigm of materials science: machine learning, AI  
(<https://doi.org/10.1063/1.4946894>)

## Challenges:

- Sheer quantity of data
- Heterogeneity of data
- Infrastructure lacking
- Broad adoption and implementation

# Why bother archiving data

FAIR principles for data provenance

What are some examples in our field that illustrate FAIR principles?

## Box 2 | The FAIR Guiding Principles

### To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

### To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

### To be Interoperable:

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

### To be Reusable:

- 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

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016).  
<https://doi-org.ezproxy.lib.utexas.edu/10.1038/sdata.2016.18>

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

# Sharing archiving/curation strategies






# Guidelines for archiving/curating

- Redundancy is key
  - Backup your data and your files on your personal compute (can automate this- at least weekly)
  - Backup your data on supercomputers on regular basis (can automate this- at least monthly)
- What is the minimum amount of data needed to reproduce this study?
  - Input/output files, structure files intermediate files
  - File count, memory footprint
- A form of tracking meta-data and workflow (e.g., README file, yaml file, json file)
  - The order of data generation and post-processing → workflow
- Tools related to archiving/curating (with relatively differing extents of flexibility and purposes)
  - QRESP
  - Zenodo, Figshare
  - Github

# Minimum working example for journal paper

Group Box folder: Wang-Mat-Group-shared-files →  
examples-workflow →  
data-curation-example-10.1021/acs.chemmater.9b05047

🏠 > Wang-Mat-Group-shared-files > examples-workflow > data-curation-example-10.1021-ac.chemmater.9b05047

NAME ↑	UPDATED	SIZE
 Data	May 22, 2023 by Wennie Wang	363 Files
 Docs	May 22, 2023 by Wennie Wang	2 Files
 Figures_Tables	May 22, 2023 by Wennie Wang	68 Files
 Scripts	May 22, 2023 by Wennie Wang	12 Files
 main-notebook.ipynb	May 22, 2023 by Wennie Wang	328 KB

Use of Jupyter notebooks as main file for tracking workflow and metadata related to figures, tables  
Figures and tables are reproducible and reusable with provided scripts, codes, and data (raw and processed)