Data Management & Sharing for MRSEC

Meghan Lafferty
Chemistry & CEMS Librarian
mlaffert@umn.edu

Wanda Marsolek
Engineering Liaison & Data Curation Librarian
Libraries' Research Data Services Team Member
mars0215@umn.edu

Lisa Wissbaum
Managing Director
Materials Research Science & Engineering Center (MRSEC)
wissbaum@umn.edu

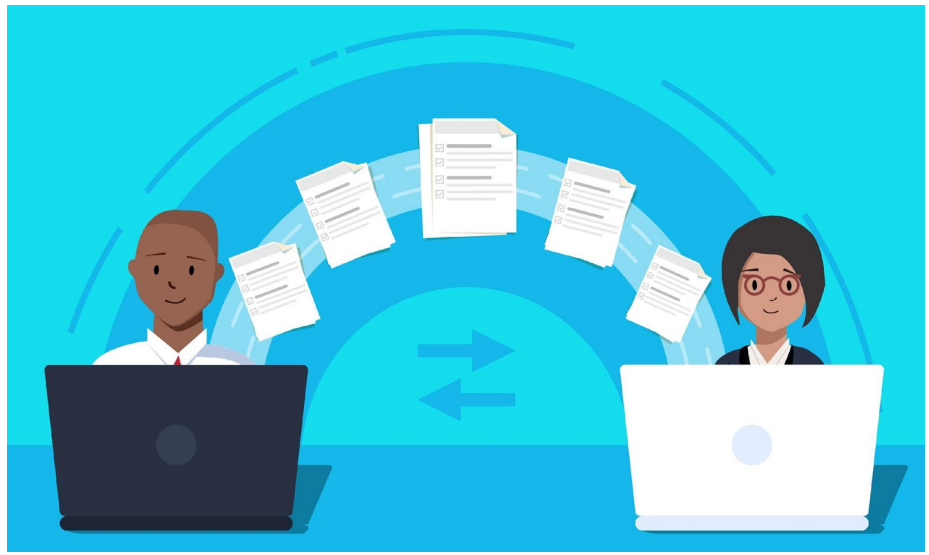
May 15, 2023

AGENDA

- Data management in MRSEC
- Why care about data management?
- Data management best practices
- MRSEC Data Management Process
- Q&A

Project priorities

1. Comply with FAIR data principles from the Data Curation Network , *i.e.*, emphasizing **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of data.
2. Meet requirements of the NSF grant
3. Protect data integrity
4. Manageable workload for participating researchers



MRSEC DATA MANAGEMENT PLAN

*"Data supporting primary and partially supported MRSEC publications can be published in a stable, citable format using the well-established data repository, **DRUM** (the Data Repository for the U of M, see: https://conservancy.umn.edu/drum)"**

*Excerpt from **UMN MRSEC data management plan** (NSF Award DMR-2011401, 9/1/20-8/31/26):

For every MRSEC primary or partially supported publication we need:

1. *a MRSEC acknowledgment to Award DMR-2011401**
2. *a publication author(s) to submit publication data sets to DRUM.* ***The expectation is that students or postdocs will submit data sets to DRUM.***
3. *DRUM curator(s) will work through this process with students/postdocs.*

*How to acknowledge the MRSEC:
http://mrsec.umn.edu/research/resources/#acknowledge

What MRSEC publications require data set archival?

Data supporting primarily and partially supported MRSEC publications are to be published in a stable, citable format using the well-established data repository, DRUM…

Primarily supported*

***Primarily** means that **50% or more** of the total financial support for the entire work reported was from MRSEC. Contact the MRSEC Managing Director or the MrSEC Direcctor if you are unsure of the correct categorization.

This is an acknowledgment example "This work was supported primarily by the National Science Foundation through the University of Minnesota MRSEC under Award Number DMR-2011401."

Partially supported

***Partially** means that **less than 50%** of the total financial support for the entire work reported was from MRSEC. Contact the MRSEC Managing Director or the MrSEC Direcctor if you are unsure of the correct categorization.

This is an acknowledgment example "This work was supported primarily by the National Science Foundation through the University of Minnesota MRSEC under Award Number DMR-2011401."

Objective, motivation, and partners

Objective: To archive data associated with MRSEC publications in a publicly accessible data repository and integrate data access with the manuscript publication process

Motivation: To improve data sharing and reuse with FAIR principles in mind, *i.e.,* emphasizing Findability, Accessibility, Interoperability, and Reuse of data.

Why Is Data Management and Sharing Important?

Respond in chat

Documenting and Describing Your Data

Label Data Thoroughly

Table 1. Ecological data with no metadata.

VO	5/30/2002	1	*AVFAT*	4.25	3.19	0.01
VO	5/30/2002	1	*BRHOR*	5.33	3.19	0.01
VO	5/30/2002	1	*CALUT*	3.33	3.19	0.01

Useless!

Table 2. Ecological data with a limited amount of metadata.

Site	Date	Plot	Sp	Bm	P	N
VO	5/30/2002	1	*AVFAT*	4.25	3.19	0.01
VO	5/30/2002	1	*BRHOR*	5.33	3.19	0.01
VO	5/30/2002	1	*CALUT*	3.33	3.19	0.01
VO	5/30/2002	2	*AVFAT*	20.82	11.91	0

Improved but still somewhat cryptic

Codes used in data table (Table 2) are given below:

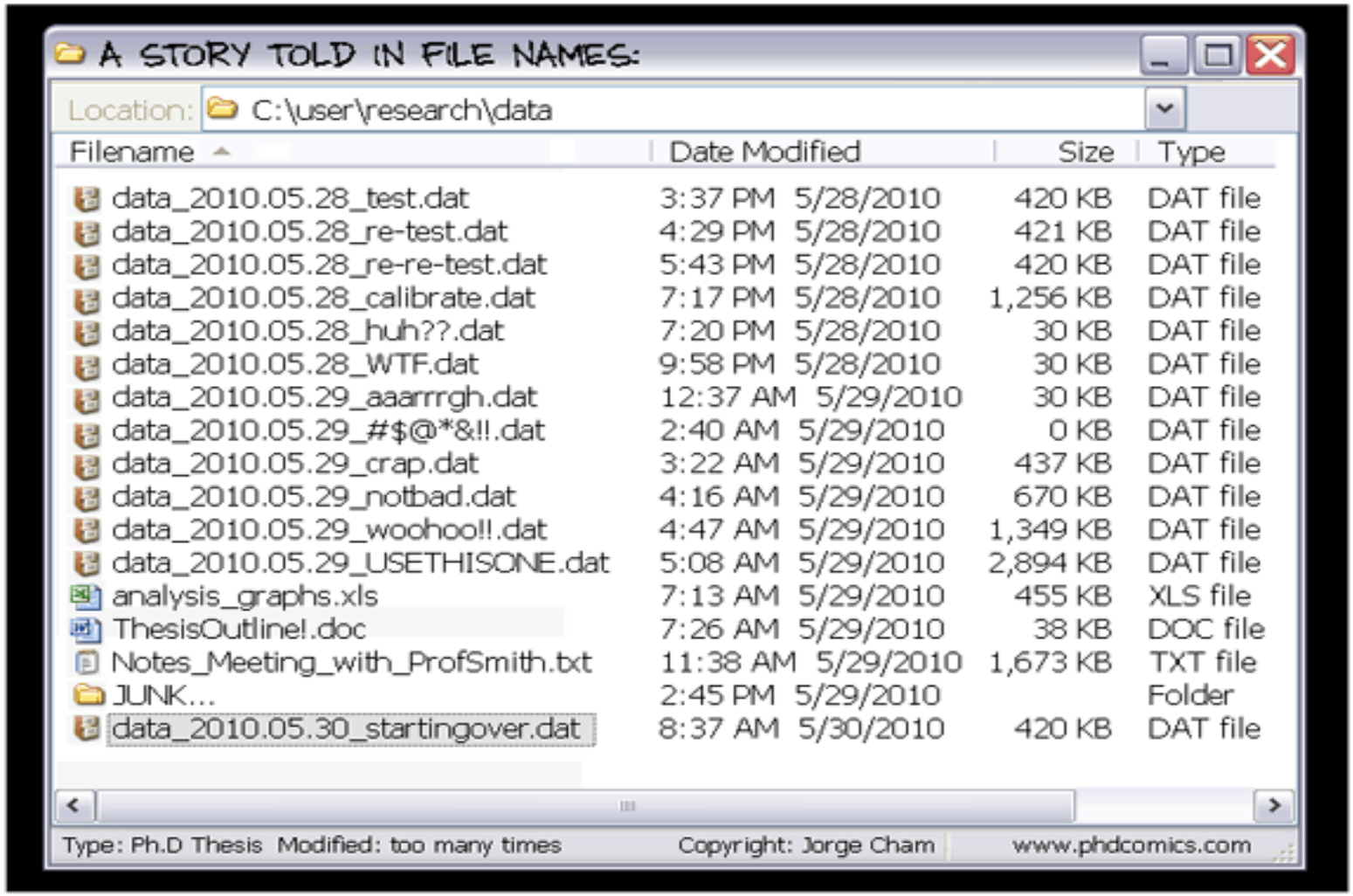
Site: Site at which data were collected – VO = Valley Oaks Reserve; Date: Date data were collected, mm/dd/yy format; Plot: Randomly assigned number of plot; Sp: Species code for each species found in plots.

Species name	Code
Avena fatua	AVFAT
Bromus hordeaceus	BRHOR
Calochortus lutens	CALUT

Bm: Biomass, measured in grams for each species; P: Phosphorus in soil, recorded in ppm (parts per million) per plot; N: Nitrogen in soil, recorded as a percentage per plot.

Documentation provides context for Future You & others

File Naming Best Practices



Cham, Jorge. (2010). “A story told in file names.” phdcomics.com. Available at http://phdcomics.com/comics/archive.php?comicid=1323

File naming best practices

Be descriptive & consistent

NMR_PEtAm_1H_and_13C.mnova

Figure S1 data.zip

File naming best practices

Put dates in YYYYMMDD format

20230515

File naming best practices

List versions alphanumerically

v01, v02, v03

MyThesisv01.docx NOT MyThesisUseTHISone.docx

Version control

- Name files based on anticipated number of versions (...01.csv, ...001.csv)
- Decide how many versions of a file to keep and when to and who will delete versions
- Create main/primary versions
  - Identify milestone versions to keep
  - Store them in a single location
- Assign responsibility of main/primary files to one team member

Good folder names:
Predictable & identify folder contents

Associated figure (e.g., Figure 1, Figure 2)

Type of data (e.g., NMR, SEC, TEM)

Substance (e.g., PLA, PGA)

Conditions (e.g., 5% strain, 10% strain)

README files

- **Who?** Who contributed to the project (authors, research assistants, etc.)?
- **What?** What kind(s) of data and analysis were used?
- **When?** When was the data collected? When was analysis performed? Any other pertinent dates?
- **Where?** Name of lab and institution or where data was collected
- **Why?** What is the impetus for the project? What questions are you trying to answer?

Readme.txt file

README template MRSEC 20200821 - Notepad

File Edit Format View Help

```
        ORCID:

 Associate or Co-investigator Contact Information
        Name:
        Institution:
        Address:
        Email:
        ORCID:

Add additional co-investigator contact information as needed.

Date of data collection (single date, range, approximate date):
 <suggested format YYYYMMDD>

Geographic location of data collection (where was data collected?):
University of Minnesota

Information about funding sources that supported the collection of the data:
This research was supported by a grant from the National Science Foundation throug

--------------------------
SHARING/ACCESS INFORMATION
--------------------------
1. Licenses/restrictions placed on the data:

2. Links to publications that cite or use the data:

3. Links to other publicly accessible locations of the data:

4. Links/relationships to ancillary data sets:

5. Was data derived from another source?

6. Recommended citation for the data:
LEAVE BLANK
```

```
---------------
File List
A. Filename:
   Short description[any information required to navigate and understand datasets]:

B. Filename:
   Short description:

C. Filename:
   Short description:

Relationship between files:

Additional related data collected that was not included in the current data package:

Are there multiple versions of the dataset?


-----------------------
THODOLOGICAL INFORMATION
-----------------------
Description of methods used for collection/generation of data [Copy the methods section from the paper (if th
), write up more detailed methods if necessary]:

Methods for processing the data:
scribe how the submitted data were generated from the raw or collected data>

Instrument- or software-specific information needed to interpret the data:

Standards and calibration information, if appropriate:

Environmental/experimental conditions:

Describe any quality-assurance procedures performed on the data:

People involved with sample collection, processing, analysis and/or submission:
st people and their role.

---------------------------------------
TA-SPECIFIC INFORMATION FOR: [FILENAME]
---------------------------------------
eate sections for each dataset included>

Number of variables:

Number of cases/rows:

Variable List
  A. Name: <variable name>
     Description: <description of the variable>
                  Value labels if appropriate
  B. Name: <variable name>
     Description: <description of the variable>
                  Value labels if appropriate

---------------------------------
ectory Structure
---------------------------------
opy directory structure.
mple: from the Windows command line: tree /a /f >listmyfiles.txt. Or copy the directory structure by hand.
```

Get credit for your data

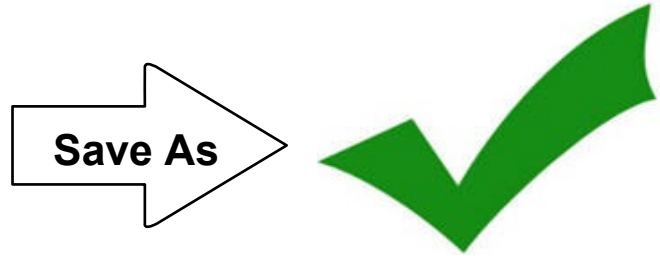
Data is not copyrightable

But you should be credited!

We recommend CCZero

Archiving data for preservation and long-term access

Preservation file formats: Long-term access



Text	MS Word	PDF, TXT, HTML
Images	Photoshop	TIFF
Video/Media	Quicktime	MPEG4
Database	MS Access	DBF
Tabular Data	MS Excel	CSV
Presentations	MS Powerpoint	PDF
Sound/Music	Windows Media	WAV

Publishing/Sharing data

Sharing data with others for reuse

- Meet research community expectations (e.g., NSF)
- Data citation increases your impact
- Stimulate new research
- Facilitate reuse and higher impact with proper attribution & stable URL/DOI
- Data retention & preservation for long-term access

LIBRARIES
digital conservancy

Search | Browse | Help | Sign in

Data Repository for U of M

How to Upload

1. Prepare Data

Data should be free of identifying or sensitive information and include adequate documentation. Not sure? Contact us for help!

2. Upload

Have your files ready (up to 2GB each) and use the upload form to fill out metadata about your data

3. Curatorial Review

Our data experts will consult with you to ensure that your data is in a format and structure that best facilitates long-term access, discovery, and reuse.

DRUM Policies

Features

Flexible Access Options

Choose to make your data immediately accessible to everyone, or moderate access to your data upon request.

Meet Grant Requirements

Comply with federal mandates for data management planning (DMP) and sharing. Read more

Maximize Reuseability

Our data experts will consult with you to ensure that your data is in a format and structure that best facilitates long-term access, discovery and reuse

About DRUM

Our Services

Data Management Plan Assistance

We offer personalize assistance for drafting your next grant's Data Management Plan. Contact us for assistance during your planning process.

Metadata Consultation

Structure your data using technological best practices to ensure the best longevity of your data.

Training and Workshops

The library offers free drop-in workshops on data management best practices periodically throughout the year.

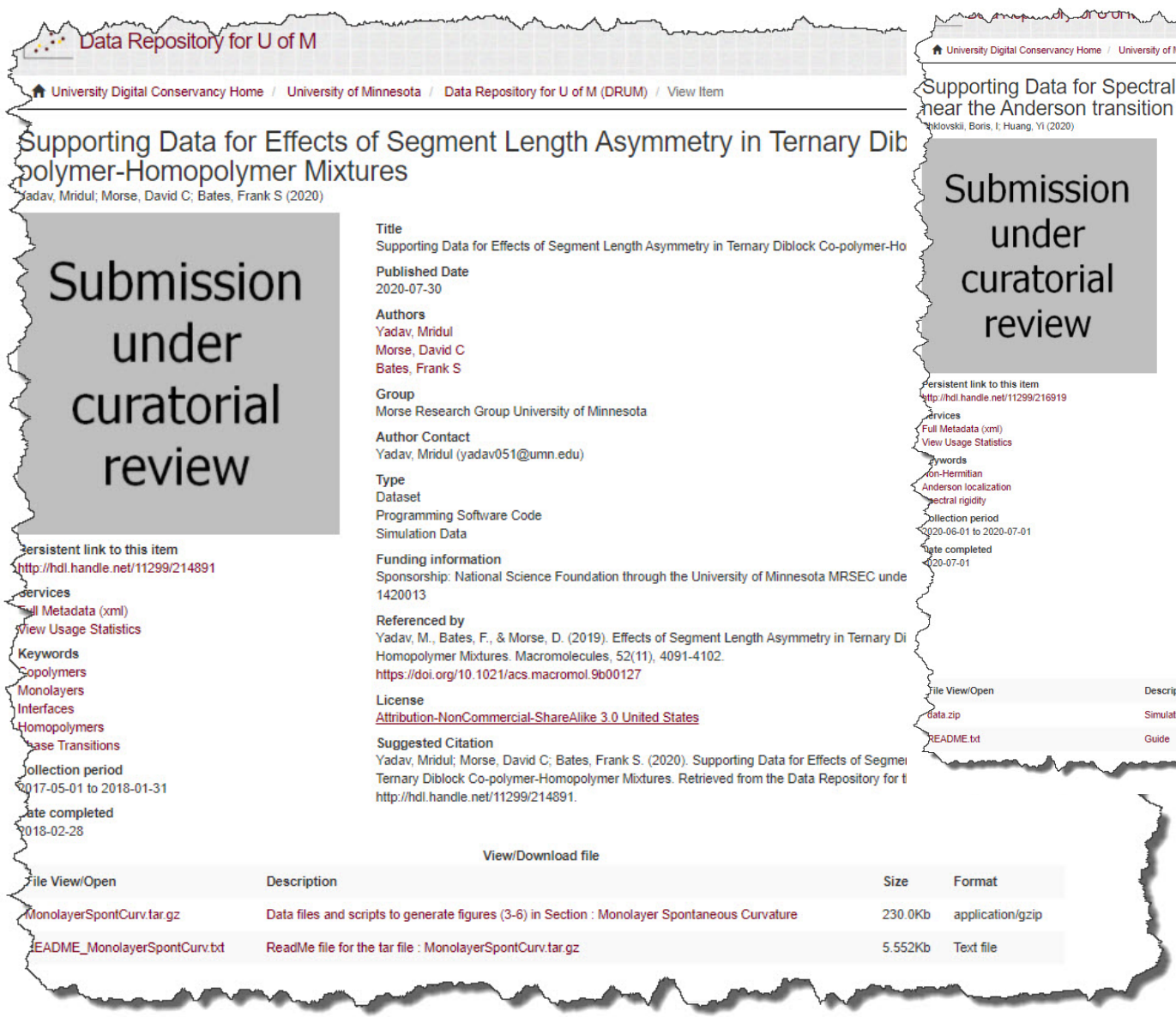
Library Data Services

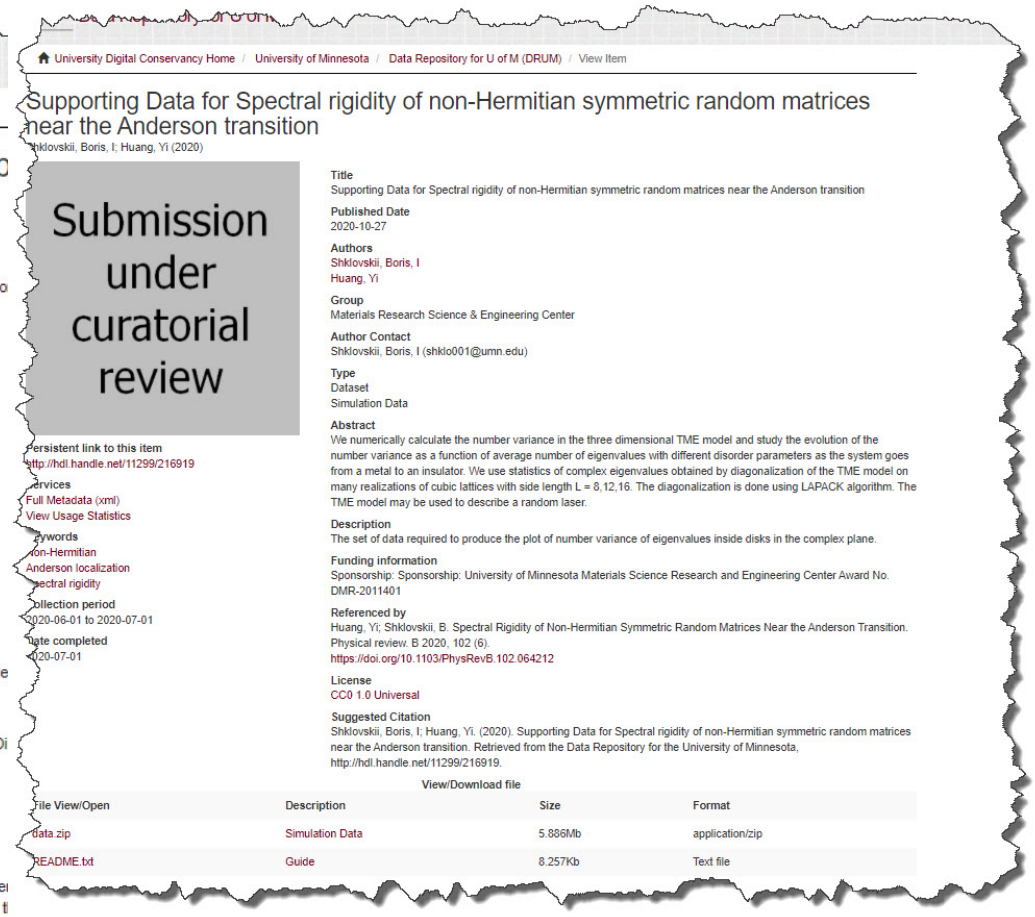
http://z.umn.edu/drum

- Free for UMN-TC students, staff, faculty
- All data types & topics
- Files up to 3GB each or 50GB for dataset
- Curated by disciplinary experts
- Public data sharing
- Database is crawled by Google

http://z.umn.edu/drum

DRUM example: MRSEC under curatorial review





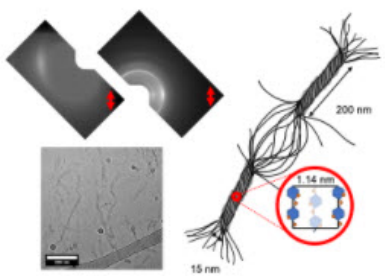
DRUM example: MRSEC Finalized

Data Repository for U of M

University Digital Conservancy Home / University of Minnesota / Data Repository for U of M (DRUM) / View Item

Supporting Data for Internal Structure of Methylcellulose Fibrils

Schmidt, Peter W; Morozova, Svetlana; Ertem, S. Piril; Coughlin, McKenzie L; Davidovich, Irina; Talmon, Yeshayahu; Reineke, Theresa M; Bates, Frank S; Lodge, Timothy P (2020)



Persistent link to this item
https://doi.org/10.13020/4gms-4q97
http://hdl.handle.net/11299/215250

Services
Full Metadata (xml)
View Usage Statistics

Keywords
nanofibers
X-ray scattering
scattering
chemical structure
methylcellulose

Collection period
2014-09-10 to 2018-06-29

Date completed
2020-01-02

Title
Supporting Data for Internal Structure of Methylcellulose Fibrils

Published Date
2020-08-26

Authors
Schmidt, Peter W
Morozova, Svetlana
Ertem, S. Piril
Coughlin, McKenzie L
Davidovich, Irina
Talmon, Yeshayahu
Reineke, Theresa M
Bates, Frank S
Lodge, Timothy P

Group
Materials Research Science & Engin

Author Contact
Lodge, Timothy P (lodge@umn.edu)

Type
Dataset
Experimental Data

Abstract
Data files used to generate all figures
Fibrils" and its supplementary inform

Funding information

Collection period
2014-09-10 to 2018-06-29

Date completed
2020-01-02

Data files used to generate all figures in the manuscript "Internal Structure of Methylcellulose Fibrils" and its supplementary information.

Funding information
Sponsorship: The collection of this data was supported primarily by the National Science Foundation through University of Minnesota MRSEC under award number DMR-1420013.

Referenced by
"Internal Structure of Methylcellulose Fibrils," P. W. Schmidt, S. Morozova, S. P. Ertem, M. L. Coughlin, I. Davidovich, Y. Talmon, T. M. Reineke, F. S. Bates, T. P. Lodge, Macromolecules 2020, 53, 398-405.
https://doi.org/10.1021/acs.macromol.9b01773

License
CC0 1.0 Universal

Suggested Citation
Schmidt, Peter W; Morozova, Svetlana; Ertem, S. Piril; Coughlin, McKenzie L; Davidovich, Irina; Talmon, Yeshayahu; Reineke, Theresa M; Bates, Frank S; Lodge, Timothy P. (2020). Supporting Data for Internal Structure of Methylcellulose Fibrils. Retrieved from the Data Repository for the University of Minnesota, https://doi.org/10.13020/4gms-4q97.

View/Download file

File View/Open	Description	Size	Format
Readme_Internal_Structure_MC_fibrils.txt	Readme	31.74Kb	Text file
Data_files_Internal_Structure_MC_fibrils.zip	Data to generate figures in manuscript and SI (131 files: txt, tiff, png)	43.58Mb	application/zip

By using these files, users agree to the Terms of Use. Content distributed via the University of Minnesota's Digital Conservancy may be subject to additional license and use restrictions applied by the

https://doi.org/10.13020/4gms -4q97

What not to include in DRUM

- Others' data
- Proprietary data
- Confidential materials
- Communications (e.g., emails)

MRSEC Data Management Process

FAIR data standards

Findable

✓ Metadata exceeds author/title/date
✓ Unique DOI
✓ Discoverable via web search engines

Accessible

✓ Retrievable via a standard protocol
✓ Free, open (e.g., download link)

Interoperable

✓ Metadata formatted in a standard schema
✓ Metadata provided in machine-readable format

Reusable

✓ Data include sufficient metadata about the data characteristics to reuse
✓ Contact information displayed if author assistance needed
✓ Clear indicators of who created, owns, and stewards the data
✓ Data are released with clear data usage terms (e.g., a CC license)

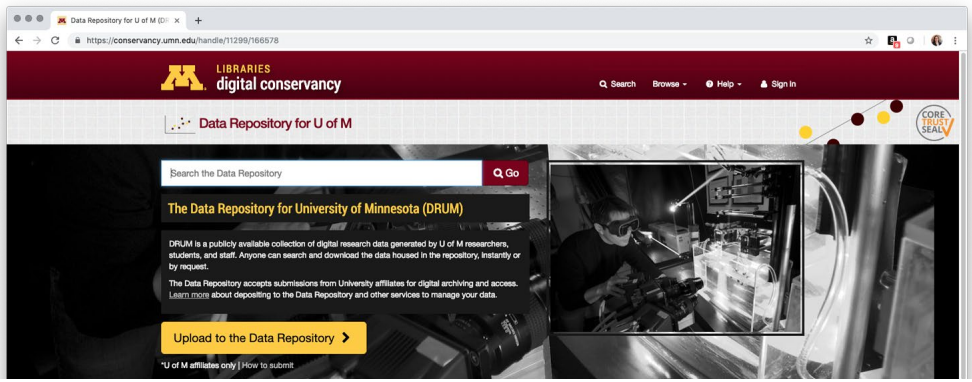
Three step process

(1) Prepare the data

- During your project use best practices for data management
- Data scope is limited to the manuscript
- Label data clearly and logically with units and specify if specific software required to open
- Organize files into a logical directory structure

(2) Upload to the data repository

- Follow specific directions in the workflow
  Ex: “Peer Reviewer Access Only”
- Write Readme.txt file
- Recommended license: Creative Commons (“No rights reserved”)



Three step process

(3) Give the journal access to the data

- Include DOI URL in the manuscript
- Notify journal editor of the process
- Revise data set as necessary
- Sync data DOI with the published manuscript

EXAMPLE MANUSCRIPT

Supporting Information

See Supporting information for Tables S1–S2, Figures S1–S13, 1H and 13C NMR spectra, HPLC results, and X-ray crystal structure data. This material is available free of charge via the internet at http://pubs.acs.org.

Data Access Statement

All primary data files are available free of charge at [Insert the URL for the DOI from A in step 1].

References

DRUM workflow

MRSEC Collection includes 46 datasets as of today

- 15 datasets in last 12 months

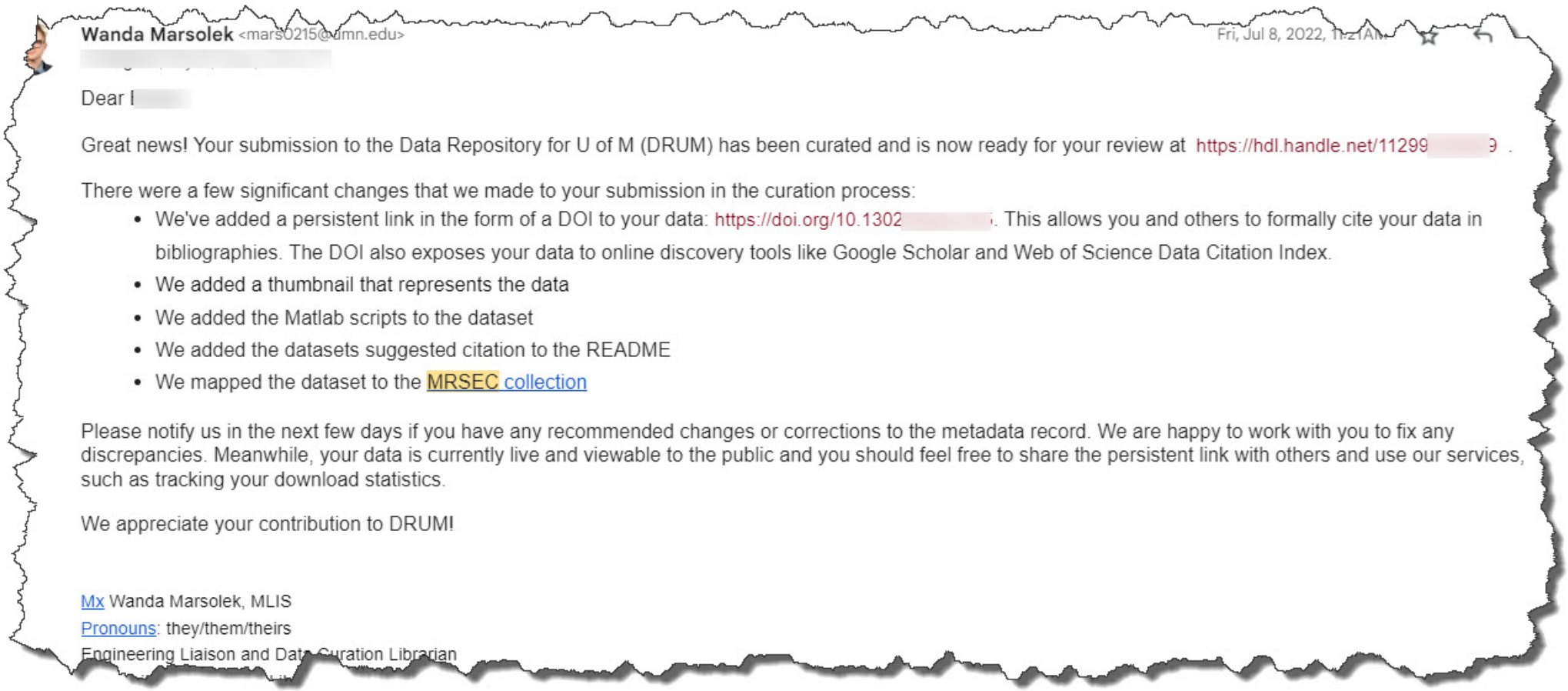
Workflow on DRUM side

- Author/submitter deposits dataset
  - DRUM coordinator looks over, accepts, assigns to data curator
- Curator runs through **C-U-R** steps of the **CURATE(D)** steps
  - **C**heck files/code & read documentation
  - **U**nderstand data (or try to)
  - **R**equest missing information or changes
- Curator emails author with questions and recommendations
  - curator waits to hear back from author before proceeding
  - this can go back and forth quickly or slowly
- Curator runs through **A-T-E-(D)** steps of the **CURATE(D)** steps
  - **A**ugment metadata
    - Curator makes changes, uploads changes, edits README
    - Curator mints DOI
  - **T**ransform file formats
  - **E**valuate for FAIRness
- Curator sends author “Great news!” email with list of changes
  - **MRSEC submitter should forward this email to Lisa W.**

****Workflow can take anywhere from 5 days to 18 months to complete****

Example of Great news! email

MRSEC submitter should forward this email to Lisa W.



Wanda Marsolek <mars0215@umn.edu> Fri, Jul 8, 2022, 11:21 AM

Dear [redacted]

Great news! Your submission to the Data Repository for U of M (DRUM) has been curated and is now ready for your review at https://hdl.handle.net/11299[redacted]9 .

There were a few significant changes that we made to your submission in the curation process:

- We've added a persistent link in the form of a DOI to your data: https://doi.org/10.1302[redacted]. This allows you and others to formally cite your data in bibliographies. The DOI also exposes your data to online discovery tools like Google Scholar and Web of Science Data Citation Index.
- We added a thumbnail that represents the data
- We added the Matlab scripts to the dataset
- We added the datasets suggested citation to the README
- We mapped the dataset to the MRSEC collection

Please notify us in the next few days if you have any recommended changes or corrections to the metadata record. We are happy to work with you to fix any discrepancies. Meanwhile, your data is currently live and viewable to the public and you should feel free to share the persistent link with others and use our services, such as tracking your download statistics.

We appreciate your contribution to DRUM!

Mx Wanda Marsolek, MLIS
Pronouns: they/them/theirs
Engineering Liaison and Data Curation Librarian

Common questions and recommendations

Zip files are encouraged (this is how you upload a folder)

- more than one folder/zip can be uploaded
    - 3-5 gb file size | 50gb total dataset

How to organize files? Consider how the files will be used

- by figure?
- by chemical?
- raw data/code/documentation

Spreadsheets

- consider how the spreadsheet is being used
    - Wanda's hot take - spreadsheets are overused
- rectangular data

Documentation

- READMEs
- help future reusers make sense of workflow
- audit of package

Code

- relative path over absolute path
- commented code and workflow included in README

Abstract and Description recommendations

- **The description works well to define the files and file types.**
  - i.e. “figure* directories include data files and MATLAB files for generating figures in the paper. *_test directories include LAMMPS input script and atomistic structure files to simulate uniaxial and bending tests. Detailed description is in the paper. *_new fix directories has .cpp and .h files. These files can be implemented in LAMMPS as a new fix command

- T**he abstract should describe the contents of the files and the value of the work.**
  - i.e. “Data includes LAMMPS input script for MoS2 test problems and MATLAB data for generating figures in the paper..”
  - i.e. “The data contains X-ray and electrical characterization of SrIrO3 films grown by solid-source metal-organic molecular beam epitaxy (SSMOMBE). It reveals that SSMOMBE can produce high-quality crystals and has numerous other advantages compared to conventional molecular beam epitaxy.”

Questions?