CURATE(D): Scientific Images

A DCN Workshop

The Pixels team

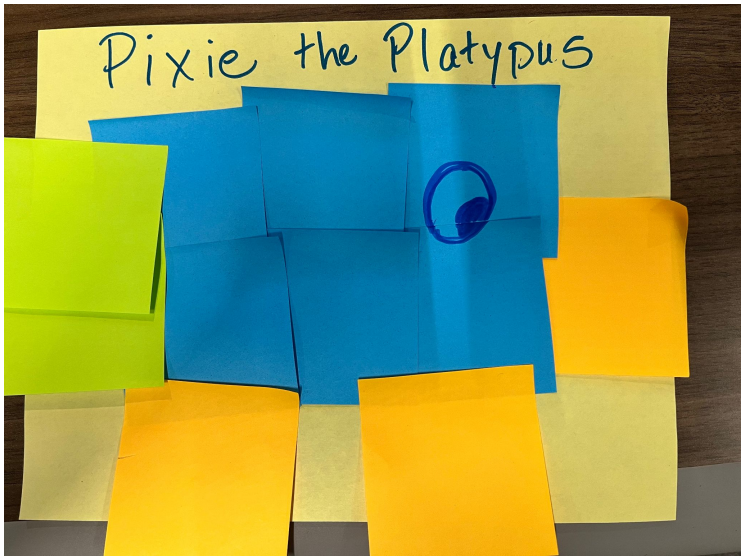
Paul Gignac, University of Arizona

Sarah J. Wright, Cornell University

Mariah Kenney, Carnegie Mellon University

Amy Schuler, Cary Institute of Ecosystem Studies

Mentor: Neggin Keshavarzian, Princeton University



Links!

- Workshop outline
- These slides
- Curator Log - make a copy
- Worksheets
- Software: ImageJ
- Dataset for exercises
    - Dataset for 'Multi-Step Crystallization of Self-Organized Spiral Eutectics'

Workshop Primary Aims:

Understand the process for curating a specialized dataset.

Apply the DCN CURATED workflow to real-world datasets containing scientific images.

Through reflection and discussion, evaluate your Curator Logs and curator experience.

Curator Log

- Keep notes on your CURATED activities throughout the workshop
- Your notes, your style!
- We do have some recommendations →

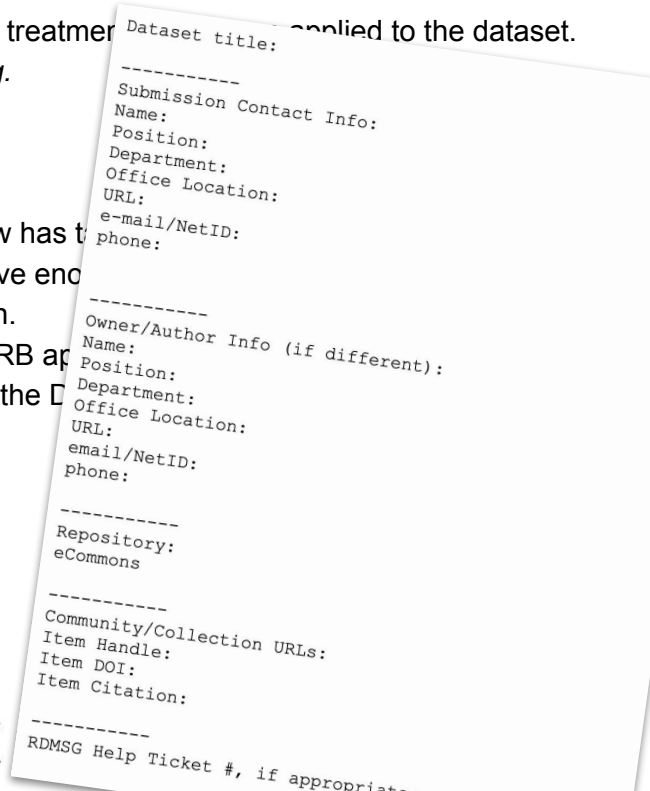
Curator Log prompt for scientific images workshop

There are many examples of curator logs, including this one from Cornell University Library. Use it as a template, or create your own.

The curator log is used to record significant treatmer applied to the dataset. This is for *your own archival record keeping.*

Key ethical considerations,

- Document that disclosure risk review has t data have been made, but do not give eno reverse-engineer any anonymization.
- Include consent (or waiver) and/or IRB ap documentation. A good resource is the D Consent primer.



Let’s Get Started

Define the term “scientific images.”

Discuss common software for viewing images.

ImageJ/FIJI

Introduce the steps to Check files and documentation.

What are “scientific images”?

White House Office of Science Technology Policy "Nelson" memorandum

For the purposes of this memorandum, "scientific data" include the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings. Such scientific data do not include ~~laboratory notebooks, preliminary analyses, case report forms, drafts of scientific papers, plans for future research, peer-reviews, communications with colleagues,~~ or ~~physical objects and materials, such as laboratory specimens, artifacts, or field notes.~~ The definition of "scientific data" is similar to, but broader than, the term "research data" defined by 2 CFR 200.315 (e) and 45 CFR 75.322 (e).

What are “scientific images”?

“A digital image is composed of a finite number of elements, each of which has a particular location and value”

-Gonzalez, Rafael (2018). Digital image processing.

Visual Representations of Scientific Data

- wide and varied field of applications
- almost no technical field that does not involve digital images in some way

What are “scientific images”?

Scope for this workshop:

- Excluding images intended for use in GIS
- Excluding scientific figures and illustrations
- Excluding galleries, libraries, archives, and museums images

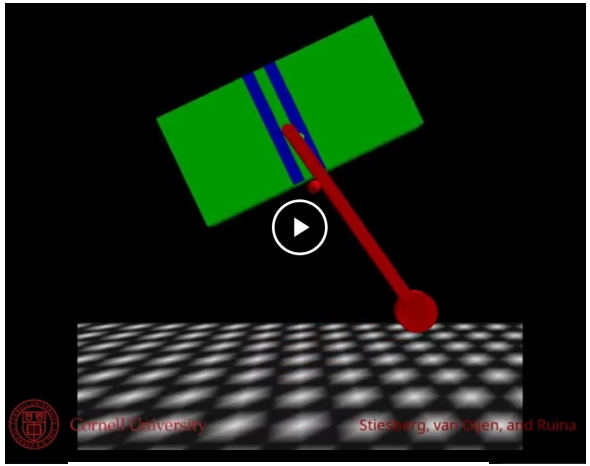
Visual representations of recorded factual material,

1) composed of a finite number of elements with a particular location and value,

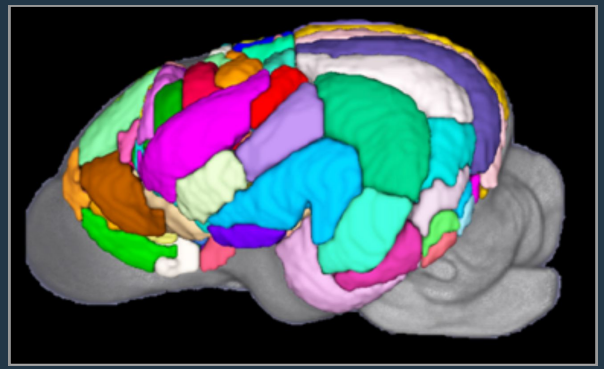
2) and commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings

Exemplar scientific images

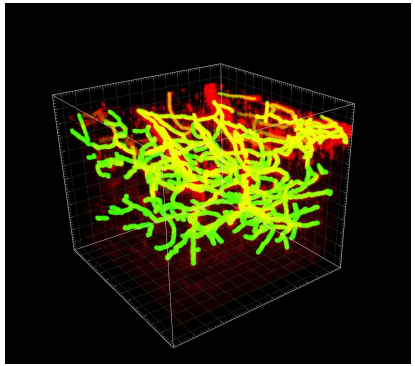
https://doi.org/10.7298/fxqt-zw38



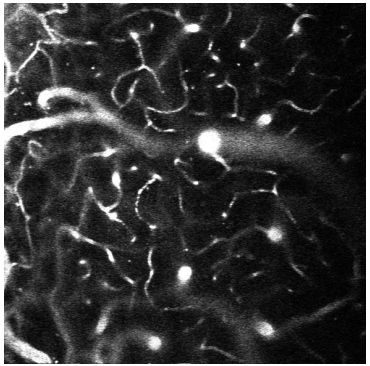
https://doi.org/10.7298/X4DZ0695



https://doi.org/10.7298/4t8z-aw34.2



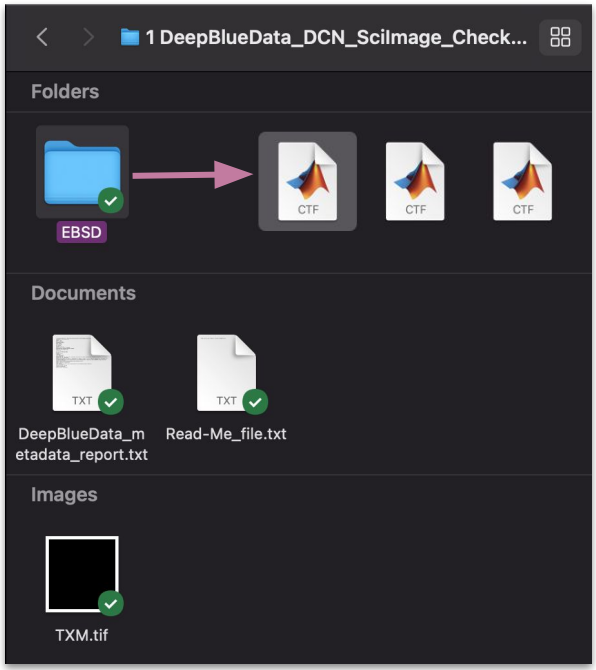
https://doi.org/10.7298/3fmv-rf23



https://doi.org/10.7298/X4FJ2F1D

Scientific Images Dataset for the first part of today's workshop

Dataset for 'Multi-Step Crystallization of Self-Organized Spiral Eutectics': https://deepblue.lib.umich.edu/data/concern/data_sets/h415p962w



Software Roundup

Imaging Tools for Processing & Review

- ImageJ (image manipulation)
- RStudio (graphing, data visualization)
- Matlab (graphing, data visualization)
- Omero (microscopic image visualization)
- Jena (3D molecular structure/Crystallographic image files)
- EBSD-Image (electron backscatter diffraction images)

Advanced Tools

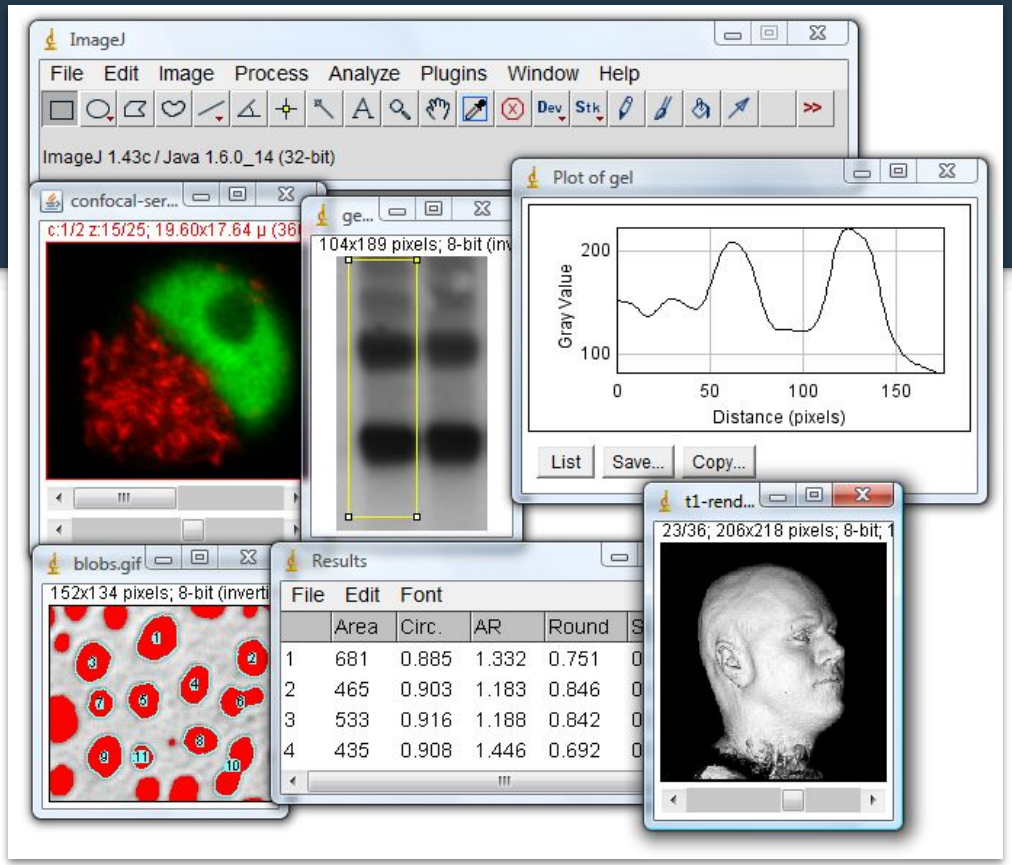
- Globus (large files)
- Python (file format conversion)
- MDEditor.org (metadata organization)
- MarvinView (chemical structures
- WebAIM Contrast Checker (accessibility tool)

Lots of proprietary & open source tools out there!

Software Roundup

1. Core Software Today:
   a. ImageJ →
   b. Text Editor/Spreadsheet

2. Supplemental Software
   a. Data Curation List Tools



ImageJ FYI for PC Users:

Caution: “Program Files” not recommended!

If you are installing ImageJ2 on Windows, we strongly recommend that you store your ImageJ2.app directory somewhere in your user space (e.g., `C:\Users\[your name]\ImageJ2.app`) rather than in `C:\Program Files` or other system-wide directory. If you move ImageJ2.app to such a directory, modern versions of Windows will deny ImageJ2 write permission to its own directory structure, preventing it from being able to update. See also imagej/imagej#72.

ImageJ FYI for Mac Users:

MacOS Arm64 Note: The default MacOS download should run on Arm64 via the Rosetta translator (https://en.wikipedia.org/wiki/Rosetta_(software)) which may come at some performance cost. Alternatively you can install the no-JRE version which defaults to the Mac Java and will limit some native library functionality that does not yet have Arm64 support (https://forum.image.sc/t/fiji-clij-etc-native-on-apple-silicon-arm64-m1/53627/25)
(^^ https://rb.gy/ko997)

Viewing Images with Fiji

1. Download ImageJ (Fiji) - https://imagej.net/software/fiji/
   - Fiji is an image processing package—a “batteries-included” distribution of ImageJ2, bundling a lot of plugins that facilitate scientific image analysis.
   - When opened, Fiji should look like the image above
2. To open an image, in the top toolbar select File -> Open
   - Select the file to open
3. Use your mouse to hover over the tools in the toolbar to get more information about how to use them
4. If your image is a z-stack, and you can view images by using the scroll bar at the bottom. The image within the stack you are currently viewing is in the top left corner.

CURATE Steps

C **Check** files and read documentation.

U **Understand** the data (or try to), if not…

R **Request** missing information or changes.

A **Augment** metadata for findability.

T **Transform** file formats for reuse.

E **Evaluate** for FAIRness.

(D) **Document** your curation activities

C⇛ U⇛ R⇛ A⇛ T⇛ E⇛(D)

Check

CHECK Step - Overview (Check Worksheet)

1. Review and inventory the content of the data files (e.g., open the files)
2. Verify all metadata provided by the author and review the available documentation
3. Ensure content of dataset is in scope
4. Look for obvious ethical/sharing red flags

CHECK Step

Obvious ethical/sharing red flags

- Human subjects/faces
- Endangered species and fossil specimens with location information
- Licensed/copywritten materials (logos, etc.)
- Missing documentation — ReadMe or other metadata should be included with each dataset

CHECK Step

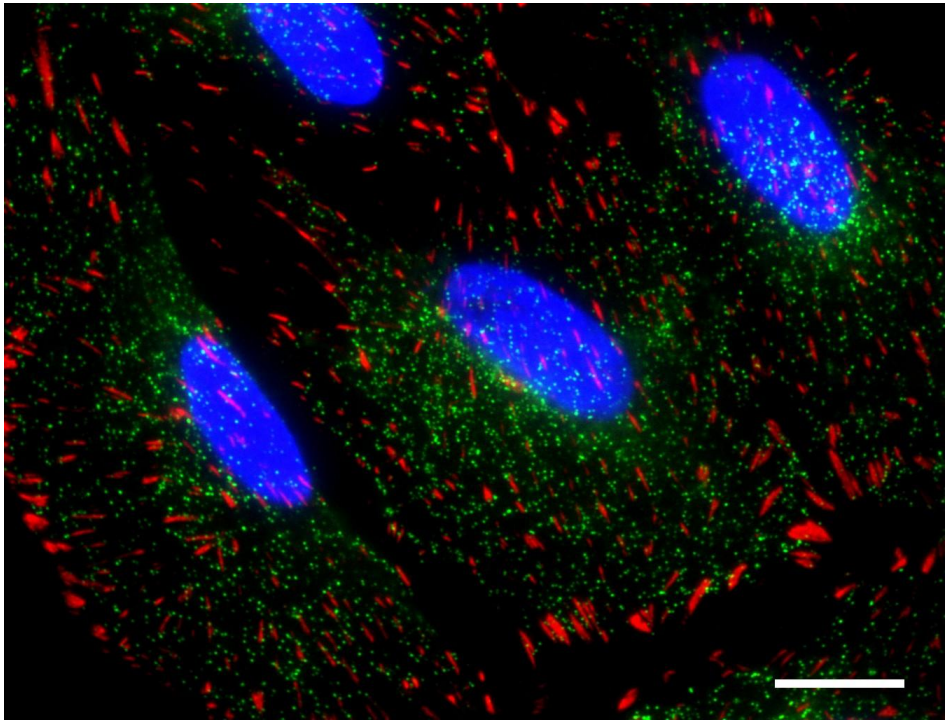
Is the dataset in scope?

- Who is intended user?
- Are there search, discovery, or access requirements?
- Are there privacy concerns or need for restricted access?

  → Would a disciplinary or image-specific repository provide better functionality?

Example: Cell Image Library

Marcel Mettlen, Sandra L. Schmid (2010) CIL:10105, Homo sapiens, epithelial cell, retinal pigmented epithelial cell. CIL. Dataset. https://doi.org/doi:10.7295/W9CIL10105



http://www.cellimagelibrary.org/images/10105

Where to find repositories?

re3data.org

https://www.re3data.org/

C⇛ **U**⇛ R⇛ A⇛ T⇛ E⇛ (D)

Understand

Learning Objectives -U-

Determine what information is missing from the practice dataset containing scientific images.

README Template

```
README TEMPLATE

Dataset Title:

Author(s):

Description:

Methods and Processing:

Funding Source:

Supplements:
```

```
<?xml version="1.0"?>

<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <dc:title>
    UKOLN
  </dc:title>
  <dc:description>
    UKOLN is a national focus of expertise in digital information
    management. It provides policy, research and awareness services
    to the UK library, information and cultural heritage communities.
    UKOLN is based at the University of Bath.
  </dc:description>
  <dc:publisher>
    UKOLN, University of Bath
  </dc:publisher>
  <dc:identifier>
    http://www.ukoln.ac.uk/
  </dc:identifier>

</metadata>
```

Note that the http://example.org/myapp/schema.xsd XML schema does not exist - this is a ficticious example.

Informal ReadMe → **Formal Schema**

Lower-Barrier	Lower-Potential	Higher-Potential	Higher-Barrier
Fast	Irregular	Standardized	Slow
Easy	Incomplete	Machine actionable	Skilled

Microscopy does not have a long history of data sharing, and most journals have no microscopy data deposition mandates. **The challenges this field faces are many and include, huge dataset sizes, diverse data output from different modalities, questions surrounding what counts as 'raw data', the need to store and save multiple versions of files due to data processing, optimal file formats, best practices for metadata recording, and cost.** However, groups like Quarep-LIMI, REMBI, Global BioImaging, Bioimaging North America and more are developing guidelines for data reporting and sharing that, should enable meaningful sharing and reuse of bioimaging data. And although not yet meeting the needs of all microscopists, image data resources and repositories such as the Image Data Resource and Bioimage Archive are growing and setting standards for the field.

"Data Sharing Is the Future." *Nature Methods* 20, no. 4 (April 2023): 471–471. https://doi.org/10.1038/s41592-023-01865-4.

UNDERSTAND

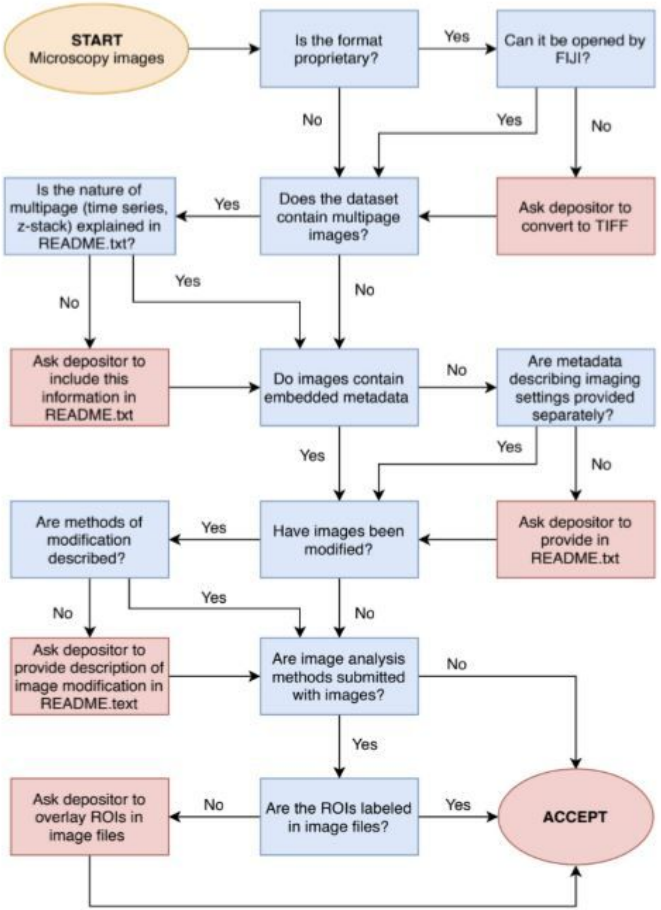
In this step, examine the dataset closely to understand what it is, how the files interrelate, and what information is needed for reuse.

Look for:

- Methods used to collect images
- Methods used to process the images
- Instrument- or software-specific information
- Standards and calibration information, if appropriate
- Environmental/experimental conditions

Confocal Microscopy Images Primer

Ivey, Susan; Koshoffer, Amy; Sneff, Gretchen; Wang, Huajin. (2019). Confocal Microscopy Images Data Curation Primer. Data Curation Network GitHub Repository.



UNDERSTAND:

DCN
Primers for Image formats

- ISO Images Primer
- Confocal Microscopy Image Primer
- GeoTIFF Primer
- netCDF Primer and Tutorial using NCAR dataset
- Neuroimagining and NICOM NIfTI Primer

Your Mission (if you choose to accept it)

Imagine you are a data curator in the DCN. We just got a new dataset that contains scientific images, and it is assigned to YOU.

Curator Log

- Log questions for follow-up during the *Request* step

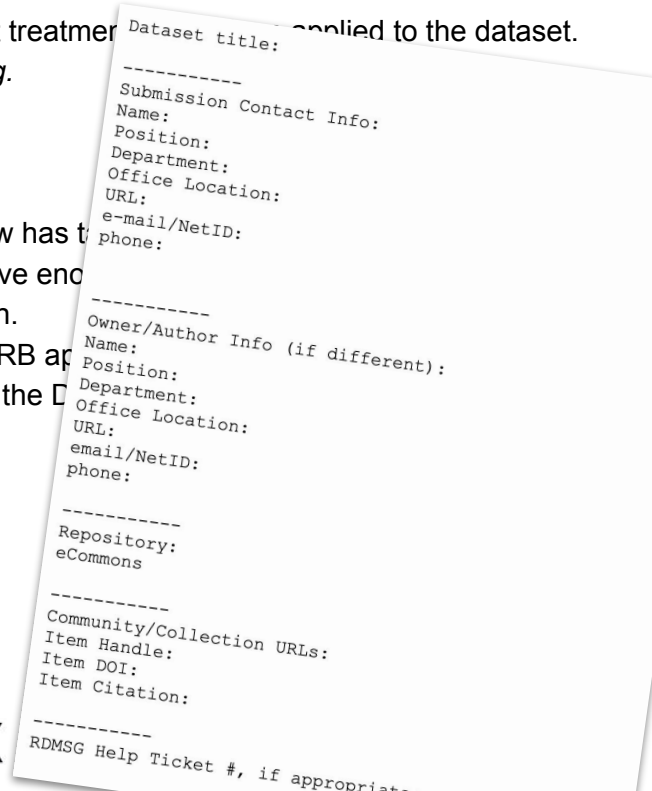
Curator Log prompt for scientific images workshop

There are many examples of curator logs, including this one from Cornell University Library. Use it as a template, or create your own.

The curator log is used to record significant treatmer applied to the dataset. This is for *your own archival record keeping.*

Key ethical considerations,

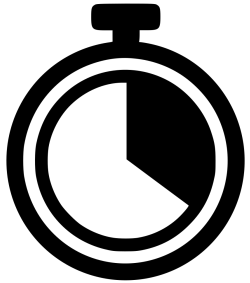
- Document that disclosure risk review has t data have been made, but do not give eno reverse-engineer any anonymization.
- Include consent (or waiver) and/or IRB ap documentation. A good resource is the D Consent primer.



CHECK Step - Exercise

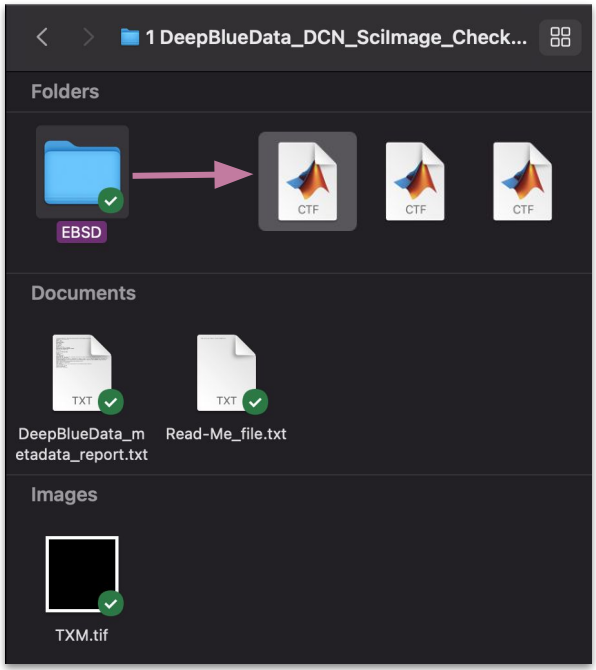
Check Sample Image Dataset with Check Worksheet (15 mins)

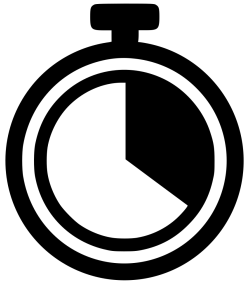
1. Open curator log and make a copy to take notes and track curator decisions
2. Download and open dataset
3. Use the Check Worksheet to review image dataset, taking notes in curator log
4. Discuss your dataset in your group and document your findings. Summarize to share out!



Share out: What did your group find?

Dataset for 'Multi-Step Crystallization of Self-Organized Spiral Eutectics': https://deepblue.lib.umich.edu/data/concern/data_sets/h415p962w





Some issues with **CHECK**

- Can’t open the files/code
- Can open the files, but don’t know what I’m looking for
- Not enough documentation

UNDERSTAND Step - Digging Deeper

1. Now, dig deeper into the data. Use the **UNDERSTAND** worksheet from the CURATE steps.
2. Keep using the curator log to take notes and track decisions.
3. Discuss your dataset in your group and document your findings.
4. Summarize findings and be prepared to share out key points.

Some issues with **UNDERSTAND**

- Not enough experience to evaluate these data
- Documentation is over my head
- Still can’t tell what’s typical in this field
- Required software is too hard to find/use/justify $$$

C⇛ U⇛ **R**⇛ A⇛ T⇛ E⇛ (D)

Request missing information

Learning Objectives -R-

Identify highest priority issues that affect FAIR-ness of the scientific images.

Write an email requesting missing information and/or any changes needed to the practice dataset containing scientific images.

The “Request” Step

What information do you need to address the gaps that you have identified?

How will you get this information?

Advice: Make it Easy

Image from Staples website: https://www.staples.com/Staples-Easy-Button/product_606396

- Limit to 4 asks (triage/prioritize)
- Be specific, but try to keep it short
- Keep it as simple as possible (Ideally, response could be “yes or no”)
- Provide resources where useful
- If possible, do the work yourself and ask for approval

Request Step - Key Ethical Considerations

- Consider asking researchers if their participants will be notified that their data are being shared
- If you feel uncomfortable about sharing the data in its current state and/or it does not meet your institution's requirements, reserve the right not to publish
- Consider asking researcher(s) if there are limitations to how data could/should be used. Include any limitations in the documentation

Request info

- Email template from University of Michigan:

Dear [name of the person Identified as the contact for the data set as stated in the DBD metadata],

Thank you for depositing your data set, [title of the data set] to the library's Deep Blue Data repository.

After we receive a data set, we review it to ensure that the data sets we host are as complete, accessible and understandable as possible. We have reviewed your data set and have the following recommendations for you:

- Recommendation #1
- Recommendation #2
- Recommendation #3
- Recommendation #4

We look forward to hearing your response to our questions and requests for additional information.

Please do let us know if you have any questions about or recommendations. We would be happy to talk with you over the phone or meet with you in person to discuss our review of your data should you wish to do so.

Sincerely,
[Name of Liaison]

Thank you

What you need from them, and why

How they should get you the info

Offer to help

Exercise on “Request” step

1. As a group, come up with the 3-4 points you’re going to ask more information on. You can use the Request worksheet if it’s helpful.
2. Pair off, and assign pairs a bullet or two (depending on size of group) and role-play curator/researcher to explain, rebut and defend that bullet point.
3. Swap roles.
4. Report back to larger group one argument that stood out. Might be to larger group, or just to dataset group, depending on time.

AI Exercise on “Request” step

1. As a group, come up with the 3-4 points you’re going to ask more information on. You can use the Request worksheet if it’s helpful.
2. Pair off, and use the bullet points to role-play as curators. You will request additional information and explain why (or rebut and defend the request as needed) from Dr. Roe Bot, an AI-based simulated researcher.
3. Report back to larger group about how the interaction went. Were you able to get the requested information?

Request information

“I'm impressed with how well you seem to understand the data and how thorough your review was!”

- Researcher response to Data Repository for the U of Minnesota (DRUM) curator, Feb 2017

C⇛ U⇛ R⇛ **A**⇛ T⇛ E⇛ (D)

Augment the submission

Learning Objectives -A-

Determine how you would enhance metadata to best facilitate discoverability, such as by ensuring images have a persistent identifier and/or links to related content.

Discuss how you would implement enhancements to metadata or documentation in the practice dataset containing scientific images.

What is Metadata?

Metadata is data about the data. What counts as necessary metadata is often repository specific and discipline specific.

Curators should request that authors provide all metadata files or parameters needed to:

1) Characterize how the images represent research findings (which often means accurately representing the physical world);
2) Repeat the study or studies described in the research paper.

Enhancing Metadata

Typically, metadata includes:

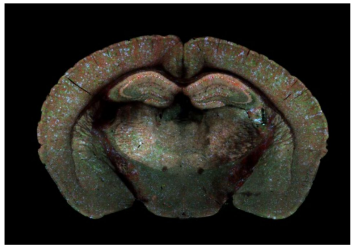
- The make, model, and settings of equipment used to capture the image(s)
- Parameters that tie the images to the physical world:
  - Spatial dimensions of pixels or voxels
  - Absolute values of brightness
  - Specific color values

Can you think of other important kinds of metadata?

Some metadata schemas

Open Microscopy Schema (OME, June 2016)

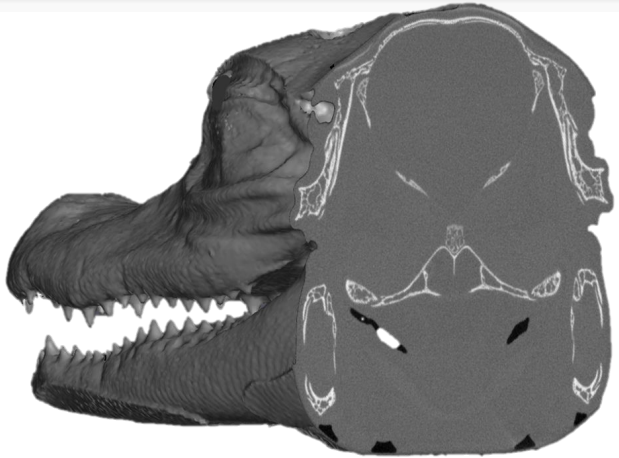
- Element:Microscope
  - Attribute: Type -microscope used to capture the image, e.g. Two Photon
- Element: Detector
  - Attribute: Model -manufacturer and model of the detector
- Element: Image
  - Attribute: Pixels -describes the actual image and its metadata, physical size of pixels are microns[µm].



Example transgenic whole mouse brain image from Image Data Resource, IDR

Some metadata schemas

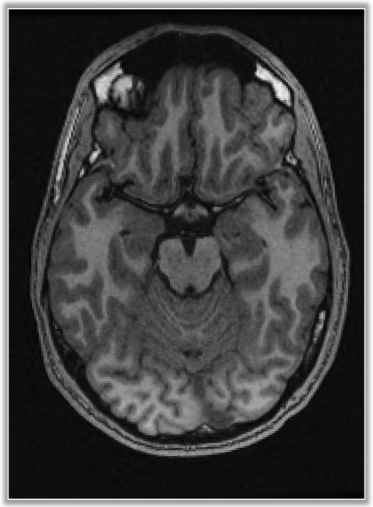
- Computed Tomography (CT) Schema
  - Element: CT Scanner
    - Attribute: Manufacturer and model of the scanner
- Hardware configuration and parameters used to capture the images
  - Element: Detector
    - Attribute: Panel size (in X & Y sensor units)
    - Attribute: Exposure timing (in seconds)
    - Attribute: Frame averaging (e.g., 2, 4, 8)
  - Element: Target
    - Attribute: Metal type (e.g., Tungsten, Molybdenum)
    - Attribute: Power (kV, Amperage, Watts)
  - Element: Magnification
    - Attribute: Voxel size (in mm x mm x mm)



Example 3D representation of an alligator head from a micro-CT scan.

Some metadata schemas

- Magnetic Resonance Imaging (MRI) Schema
  - Element: MR Imager
    - Attribute: Manufacturer and model of the scanner
- Hardware configuration and parameters used to capture the images
  - Element: Power
    - Attribute: Magnetic field strength (in Tesla)
  - Element: Coil Configuration
    - Attribute: Receive coil information (e.g., name, active elements)
  - Element: Sequence Specifics
    - Attribute: Pulse sequence (e.g., Spoiled gradient recalled echo)
    - Attribute: Scanning sequence (e.g, T1-FLAIR)
    - Attribute: Scan options (e.g., Maximum gradient amplitude, gradient slew rate adjustments)
  - Element: Resolution
    - Attribute: Voxel size (in mm x mm x mm)



Example MRI section of a human brain.

Additional Considerations

- Linkages
  - Articles & other outputs
  - Other sources or inputs
  - Related datasets
- Persistent Identifiers
  - DOIs, ORCIDs, RORs, RRIDs

Photo by Chris Leipelt on Unsplash

C⇛ U⇛ R⇛ A⇛ **T**⇛ E⇛ (D)

Transform file formats

Learning Objectives -T-

1. Explore examples of different file format transformations.
2. Identify pros and cons of transforming files using the practice dataset or our relevant experiences.
3. Determine whether you would recommend transformation for the practice dataset containing scientific images.
4. Discuss formats for preservation versus interoperability versus reusability.

Why transform an image file?

Image file transformation depends on the image type and the goal:

- Proprietary to open format
- Preservation of original files
- Storage size considerations
- Formats to retain metadata within the image file
- Support transparency
- Use in data analysis/pipelines

Example File Format Transformations

Native Software or Format	Suggested Formats or Transformations	Transformation Tools and Notes
Microscopy Images (CZI, ND2, LSM, LIF)	TIFF, JPG	Use “export”; Omero, Bioformats; WikiData tracks software and file formats for preservation
RAW	Raster file format	Can be opened using ImageJ/Fiji

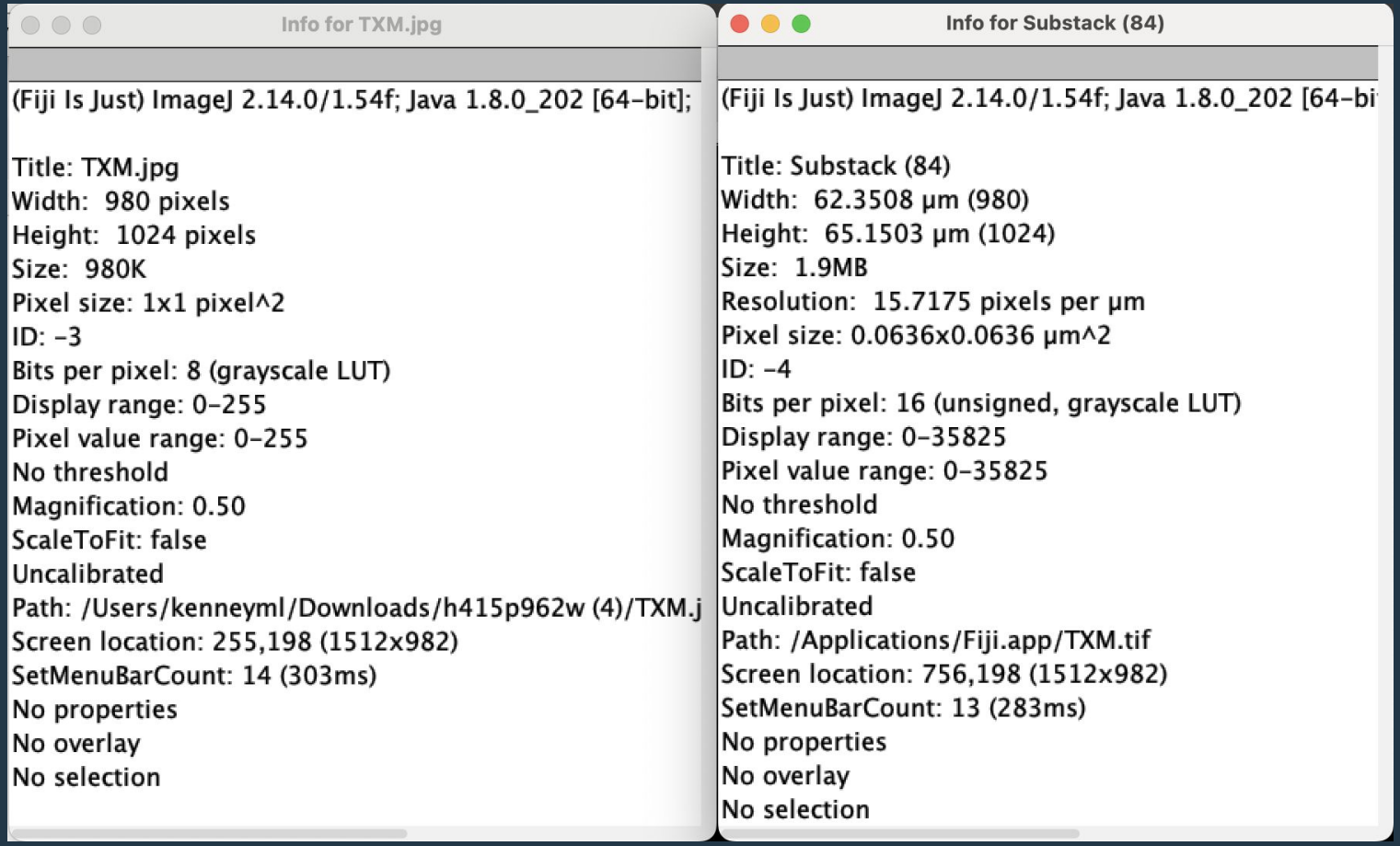
Common Formats	Format Notes
TIFF	Lossless, larger file size, raster image, supports transparency
JPEG	Lossy, raster image, common for web images
PNG	Lossless, common for web images, raster image, support transparency
SVG	Vector information, common with illustration

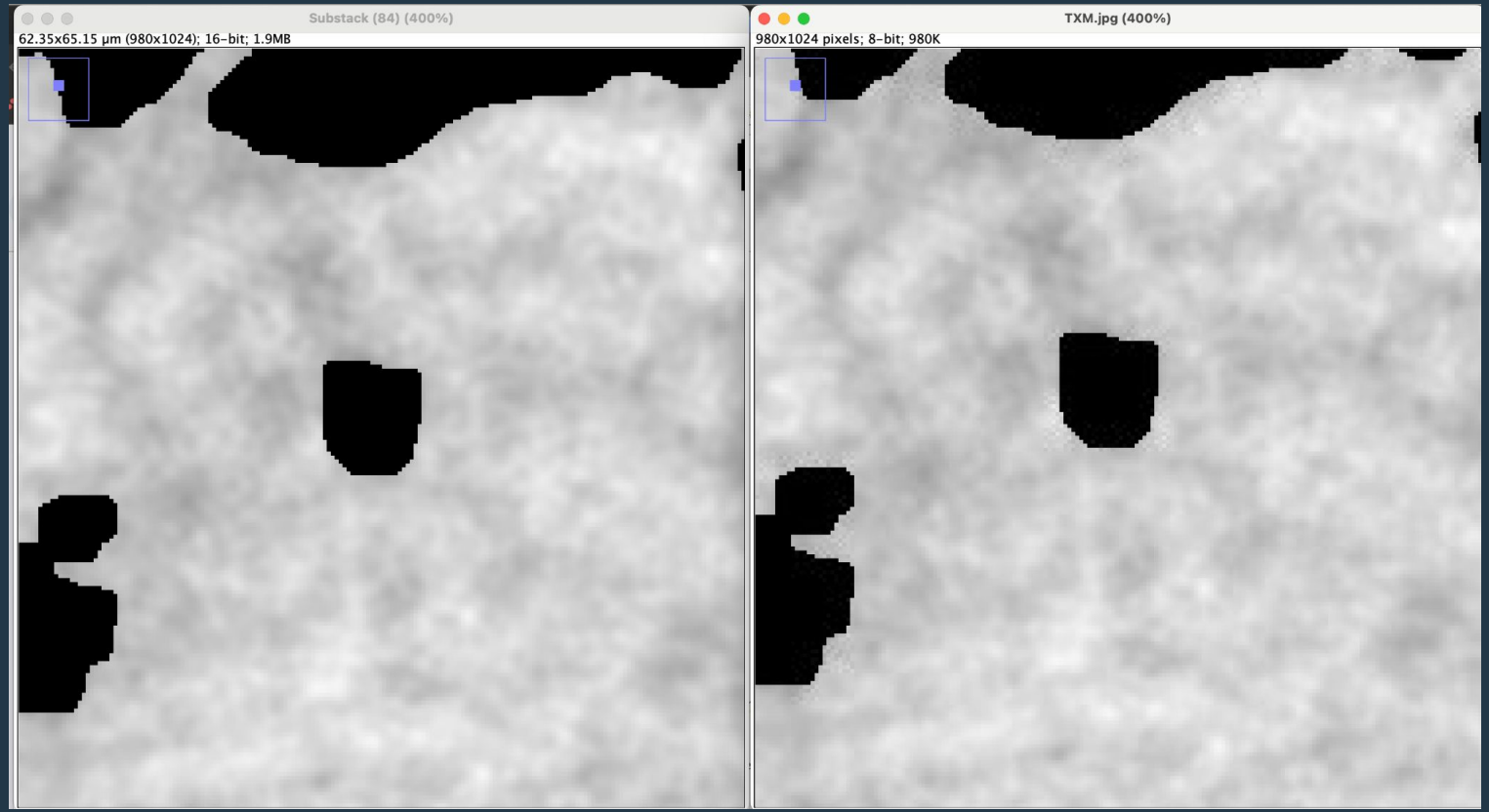
Transformation Exercise

Follow along with the worksheet to look at some examples of transformed images, compare what is lost, what is gained, and what basic tools are available.

Option 1: Follow along and do the transformation steps on your own

Option 2: Watch the transformation steps worksheet


Info for TXM.jpg
(Fiji Is Just) ImageJ 2.14.0/1.54f; Java 1.8.0_202 [64-bit];
Title: TXM.jpg
Width: 980 pixels
Height: 1024 pixels
Size: 980K
Pixel size: 1x1 pixel^2
ID: -3
Bits per pixel: 8 (grayscale LUT)
Display range: 0-255
Pixel value range: 0-255
No threshold
Magnification: 0.50
ScaleToFit: false
Uncalibrated
Path: /Users/kenneyml/Downloads/h415p962w (4)/TXM.j
Screen location: 255,198 (1512x982)
SetMenuBarCount: 14 (303ms)
No properties
No overlay
No selection
Info for Substack (84)
(Fiji Is Just) ImageJ 2.14.0/1.54f; Java 1.8.0_202 [64-bi
Title: Substack (84)
Width: 62.3508 µm (980)
Height: 65.1503 µm (1024)
Size: 1.9MB
Resolution: 15.7175 pixels per µm
Pixel size: 0.0636x0.0636 µm^2
ID: -4
Bits per pixel: 16 (unsigned, grayscale LUT)
Display range: 0-35825
Pixel value range: 0-35825
No threshold
Magnification: 0.50
ScaleToFit: false
Uncalibrated
Path: /Applications/Fiji.app/TXM.tif
Screen location: 756,198 (1512x982)
SetMenuBarCount: 13 (283ms)
No properties
No overlay
No selection


Substack (84) (400%)
62.35x65.15 µm (980x1024); 16-bit; 1.9MB
TXM.jpg (400%)
980x1024 pixels; 8-bit; 980K

Group Discussion: Preservation Actions

1. List possible format transformations for your datasets
    a. Consider how any transformation benefits different stakeholders
2. What are the challenges to any particular format transformations or stakeholder perspectives?
3. As a curator, do you understand at what level your institution as promised to “preserve” the data, and what the implications that policy has in practice?

C⇛ U⇛ R⇛ A⇛ T⇛ **E**⇛ (D)

⇓

Evaluate our data

Learning Objectives -E-

Define what we are Evaluating.

Evaluate the practice dataset according to the FAIR and CARE principles.

Findable

*To be **findable (F)** or discoverable, data and metadata should be richly described to enable attribute-based search*

- ❏ (meta)data are assigned a globally unique and eternally persistent identifier
- ❏ data are described with rich metadata
- ❏ (meta)data are registered or indexed in a searchable resource
- ❏ metadata specify the data identifier

Accessible

*To be broadly **accessible (A),** data and metadata should be retrievable in a variety of formats that are sensible to humans and machines using persistent identifiers*

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

Interoperable

*To be **interoperable (I),** the description of metadata elements should follow community guidelines that use an open, well defined vocabulary.*

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

Reusable

*To be **reusable (R),** the description of essential, recommended, and optional metadata elements should be machine processable and verifiable, use should be easy and data should be citable to sustain data sharing and recognize the value of data.*

- ❏ meta(data) have a plurality of accurate and relevant attributes
- ❏ (meta)data are released with a clear and accessible data usage license
- ❏ (meta)data are associated with their provenance
- ❏ (meta)data meet domain-relevant community standards

What are the CARE Principles?

The CARE Principles for Indigenous Data Governance, briefly:

Collective Benefit: Enabling Indigenous Peoples to derive benefit from the data

Authority to Control: Empowering them to control their data

Responsibility: Showing how data “are used to support Indigenous Peoples’ self-determination and collective benefit.”

Ethics: “Indigenous Peoples’ rights and wellbeing should be the primary concern at all stages of the data life cycle”

“Evaluate” Exercise

As a table:
10 minutes:

1. Review the final dataset, as currently visible and compare against the FAIR checklist

Dataset for 'Multi-Step Crystallization of Self-Organized Spiral Eutectics':
https://deepblue.lib.umich.edu/data/concern/data_sets/h415p962w

2. How might the CARE Principles be relevant to scientific image data?

C⇛ U⇛ R⇛ A⇛ T⇛ E⇛ (**D**)

Document curation activities

Document your curation activities

- What did you document in your curation log specific to scientific images?
- Do you have any practices or tools that haven't been mentioned here?
- Does anyone have a Scientific Images Curation 💀👻🎃 *horror story* 🎃👻💀 to share?