

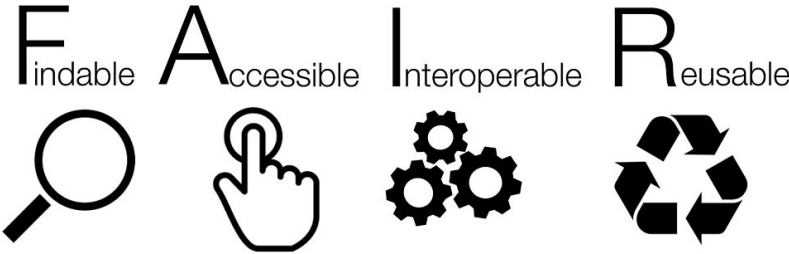


# F.A.I.R Data Policies and Practices for SEES

*Matthew Newville, U Chicago*

The Data Management Plan for SEES states:

a) we will work toward making all data collected with SEES support publicly available, complying with F.A.I.R. data-sharing principles:



b) we will engage with the SEES user community to establish policies and procedures.

Thanks for being here, we need your help.

Materials are at: <https://seescience.org/2024townhall/>



## F.A.I.R Data for Synchrotron Beamline Data

FAIR Data principles say that federally-funded research data should be made publicly available after an *embargo period*. The data should be in usable formats with consistent metadata.

There are many existing repositories at universities or specific fields. Most do not support large datasets (> 5 GB) that are common from synchrotron beamlines.

Several European synchrotrons/user facilities make data publicly available or are working to do so within the next few years.

The DOE-funded synchrotrons/user facilities are not doing this, yet.

SEES is planning to manage an online catalog of data collected at SEES beamlines, including the SEES-funded work at the beamlines it uses part-time. Data collected by SEES will be made public.

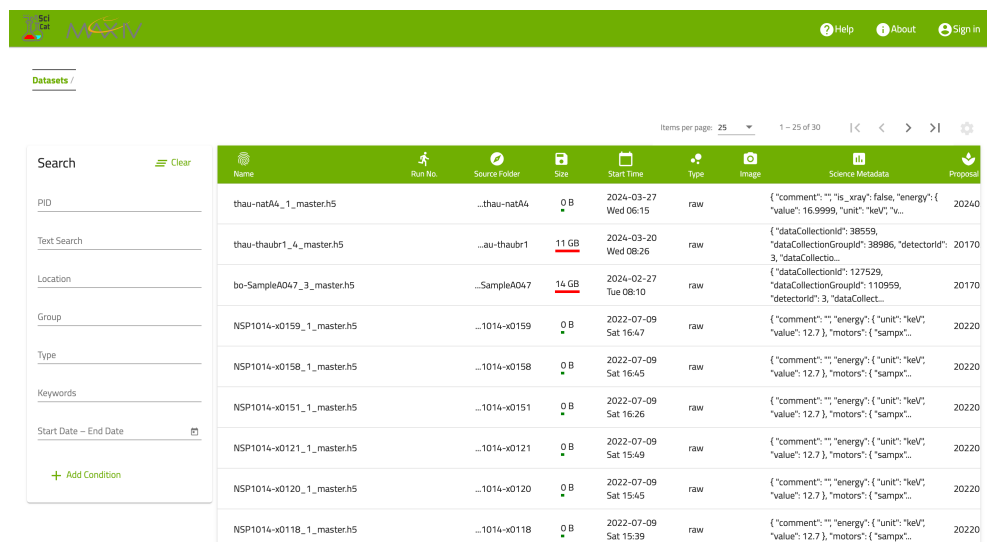
# F.A.I.R Data Implementation Plan (“Concept of a Plan”)

We plan to set up a data catalog using scicat project (from PSI, used at many European facilities – knows about “beam run” at “beamline” by “user group”)

Data will be organized by Experiment ID (Safety Form), with metadata including:

- proposal title and abstract.
- beamline name, info.
- user names and ORCID IDs.
- list of samples.

All data from an experiment run will be given a unique DOI, cataloged, and made available in phases: first only to the experimenters and beamline staff, and then to the public.



Name	Run No.	Source Folder	Size	Start Time	Type	Image	Science Metadata	Proposal
thau-natA4_1_master:h5		...thau-natA4	0 B	2024-03-27 Wed 06:15	raw		{ "comment": "", "is_xray": false, "energy": { "value": 16.9999, "unit": "keV", "v...	20240
thau-thaubr1_4_master:h5		...au-thaubr1	11 GB	2024-03-20 Wed 08:26	raw		{ "dataCollectionId": 38559, "dataCollectionGroupI...	20170
bo-SampleA047_3_master:h5		...SampleA047	14 GB	2024-02-27 Tue 08:10	raw		{ "dataCollectionId": 127529, "dataCollectionGroupI...	20170
NSP1014-x0159_1_master:h5		...1014-x0159	0 B	2022-07-09 Sat 16:47	raw		{ "comment": "", "energy": { "unit": "keV", "value": 12.7 }, "motors": { "samp...	20220
NSP1014-x0158_1_master:h5		...1014-x0158	0 B	2022-07-09 Sat 16:45	raw		{ "comment": "", "energy": { "unit": "keV", "value": 12.7 }, "motors": { "samp...	20220
NSP1014-x0151_1_master:h5		...1014-x0151	0 B	2022-07-09 Sat 16:26	raw		{ "comment": "", "energy": { "unit": "keV", "value": 12.7 }, "motors": { "samp...	20220
NSP1014-x0121_1_master:h5		...1014-x0121	0 B	2022-07-09 Sat 15:49	raw		{ "comment": "", "energy": { "unit": "keV", "value": 12.7 }, "motors": { "samp...	20220
NSP1014-x0120_1_master:h5		...1014-x0120	0 B	2022-07-09 Sat 15:45	raw		{ "comment": "", "energy": { "unit": "keV", "value": 12.7 }, "motors": { "samp...	20220
NSP1014-x0118_1_master:h5		...1014-x0118	0 B	2022-07-09 Sat 15:39	raw		{ "comment": "", "energy": { "unit": "keV", "value": 12.7 }, "motors": { "samp...	20220

See also: <https://data.esrf.fr/>



## F.A.I.R Data Implementation Plan (“Concept of a Plan”)

We are in the early stages, with several things to work out:

- get the scicat catalog running on our servers.
- make the F.A.I.R. Data policy clear to all users.
- automatic data transfer from SEES-supported beamlines to storage servers at the APS.
- automatic generation of DOIs and metadata for the catalog from Experiment Approval forms.
- automatic or on-demand creation of download links for using HTTPS, Globus, or NextCloud.

We expect SEES beamlines to support ~1000 individual experiments per year, so these automated steps will need to be robust.

We currently have ~600 TB of storage available for new data.

We may be able to use the ANL tape archiving system to automate archiving and retrieving older data.





## SEES Users F.A.I.R Data Survey

This past summer, the SEES User Group (Chairs: Maryjo Brounce and Bin Chen) sent out a survey on F.A.I.R Data Policies for SEES.

Some Key Findings (36 responses):

- Most responses were positive for FAIR principles, and many of the responses outlined the challenges.
- Working toward more uniform data formats is important but will take real effort.
- Metadata about the sample conditions will be very hard.
- There was some concern about data theft and security.
- The community wants a relatively long embargo period.

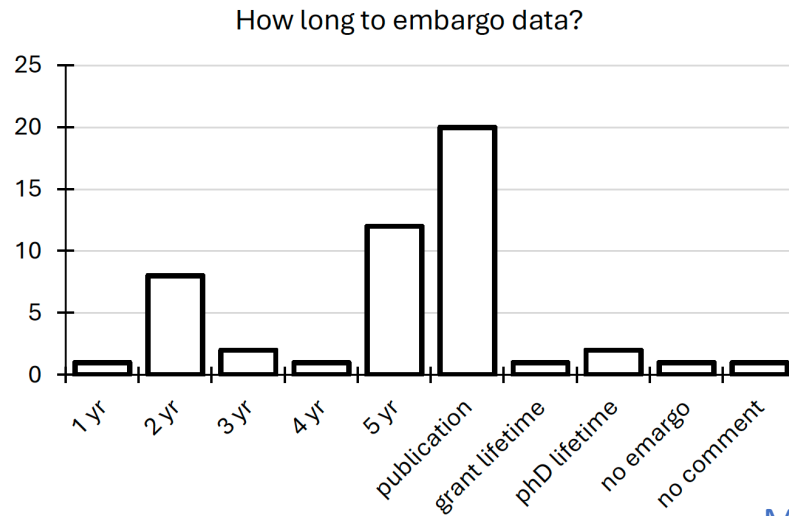
Materials are at: <https://seescience.org/2024townhall/>



## SEES Users F.A.I.R Data Survey: Embargo Period

For the Embargo Period, the Survey results were:

*Most respondents were in favor of a combined “5 years or publication, whichever comes first” approach. They expressed a strong desire to protect the training process for graduate students, and the intellectual property of PIs, especially in the case of developing complicated procedures on long timescales.*



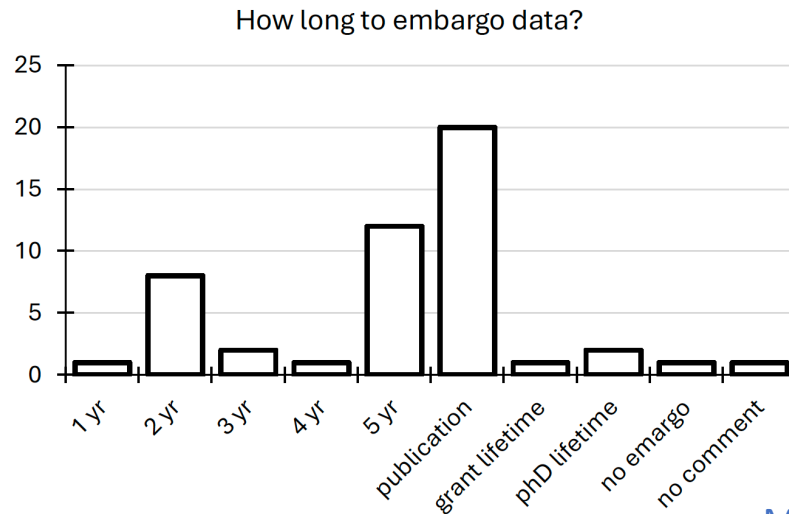
Materials are at: <https://seescience.org/2024townhall/>



## SEES Users F.A.I.R Data Survey: Embargo Period

For the Embargo Period, the Survey results were

*Most respondents were in favor of a combined “5 years or publication, whichever comes first” approach. They expressed a strong desire to protect the training process for graduate students, and the intellectual property of PIs, especially in the case of developing complicated procedures on long timescales.*



“Time of Publication” is challenging:

- We have a hard time tracking publications as it is.
- How will we know which datasets go into a publication?

If we assign a Dataset DOI at the start of each beam time, will they be used in publications?

Materials are at: <https://seescience.org/2024townhall/>



## NSF-EAR Guidelines for Data Management, July 2023

In July 2023, NSF-EAR revised its Guidelines for Data Management Plans for new proposals, requiring F.A.I.R. principles. Selected quotes:

1. *All new data resulting from the project must be made publicly accessible within two (2) years after completion of data collection or generation, via appropriate long-lived FAIR-aligned repositories.... **“Data available upon request” is not acceptable.***
2. *All data in support of peer-reviewed scholarly publications resulting from the project must also be made publicly accessible at or before the time of publication.*
3. *Possible types of “data” to be addressed in the DMP include, but are not limited to: observational, experimental, analytical, and model outputs; derived and compiled datasets; software and code; educational materials; and any other relevant digital products resulting from the project.*
4. *For proposals providing community-serving infrastructure or research services, the DMP should describe the data/sample types to be managed and what guidance or support will be provided to help users meet their data/sample sharing obligations.*

It's not clear if the SEES Data Management Plan needs to be revised to match these new requirements, but this outlines the expectations of NSF-EAR.





## We need your help defining the embargo period

2 years (NSF-EAR), 3 years (ESRF), or 5 years (User survey)?

We understand the cadence of beamtime proposals (one or two beamtime runs per year for a study) means that experiments often take two years to complete.

But we need to organize data by beamtime or experiment, not by a user's definition of “completed project” – we do not know this.

Perhaps we can interpret “*two (2) years after completion of data collection or generation*” as 2 years after the completion of the allocated **beamtime proposal**. If so, then data from the first beam run of a proposal might be embargoed for up to 4 years, while data from the last beamtime would be embargoed for 2 years.

This might vary by facility, but most beamtime proposals have a time scale of a few years.



## Other Questions we have

**2 years (NSF-EAR), 3 years (ESRF), or 5 years (User survey)?**

- Do we need to increase the embargo for work for student's theses? Is that 1 year? How do we manage and enforce that?
- Do your grants have conflicting requirements for data sharing? For example, export or security concerns?
- Can we require all General Users on our beamlines to follow these policies, if the DOE facility does not have a FAIR policy?
- Which experiments at beamlines that are only partly SEES-supported are archived? There may be some subtle cases.
- Can we start with only cataloging the experiments and data, before we move to a centralized data repository? **I think so.**



## We need your help with data formats and metadata.

Making data available does not do much good if the datasets are a mess and hard to decipher.

Working on standards for data in your sub-fields – or at least across SEES beamlines – would be very helpful for everyone.

NSF Guidelines apply to data *and samples*, which will likely be coming your way too. Working on that would help us all out.



## Questions from the registration form:

Q: Is the new proposed data-sharing policy a result of a DOE / NSF directive?

A: Yes, NSF directive.

Q: Are there any plans/examples of open-source analysis scripts being made available at the beamlines for processing beamline-specific data on-the-fly? (to make beamtime more efficient and not relying on user-made scripts that often need to be updated when changes to beamlines, data structures, etc. are made)

A: Many beamlines have such scripts. Improving collection, visualization, and processing workflows would be a great thing to work on.

Q: Is there interest as part of the FAIR data sharing to create a database for common minerals, etc. that are used as standards for XAS data collection?

A: Yes, there is interest. Some online XAS databases exist but probably need more data. We might need good mineral samples too....



## Conclusion: we need your help

Policies for data sharing, including embargo time, are going to be set soon. We expect:

- assign a Dataset DOI at the start of each beam time, to be used in publications.
- catalog metadata about each experiment supported by SEES, and eventually make that data publicly available.
- Propose interpreting “two (2) years after completion of data collection” as 2 years after the completion of the allocated **beamtime proposal**.

We would like to hear from you.

Please let us know if you have other comments on sharing data, data collection, or data management at the SEES beamlines, or would like to have a one-on-one conversation.

*And: Please suggest topics for future Town Hall discussions.*

Materials are at: <https://seescience.org/2024townhall/>

## Bonus Slide: We can provide some data hosting services

SEES can now create DOIs for datasets that we host on our servers.

If you are publishing work using SEES beamlines and would like to create a DOI for a dataset for that publication (data that makes figures, tables, or goes into Supplemental Information), let us know.

We don't have this fully automated yet, but this is our to-do list.

