Survey to SEES User base on FAIR data implementation

Maryjo Brounce, and the SEES UG-EC

26 July 2024

This survey was circulated to the SEES User community via the email. A link to the survey was circulated on the SEES, IsoGeoChem, and MSA-Talk listservs, with two repeat email reminders space 1.5 weeks apart for the duration of the month of June 2024. The survey contained the following preamble:

The SEES organization is required by NSF to incorporate FAIR principles for data management and stewardship at SEES funded beamlines. The FAIR principles seek to make data Findable, Accessible, Interoperable, and Reusable. More information on these principles can be found here: https://www.go-fair.org/fair-principles/. In the context of the SEES User community, the SEES leadership seeks your input as they begin to build FAIR principles into synchrotron-source data management schemes at SEES funded beamlines, including the building of a database, the emplacement of embargo periods for public data accessibility, and deciding which/how various raw data products are uploaded to become publicly discoverable.

To this end, the SEES leadership seeks your input. Please take some time to complete the following survey. The SEES Executive User Group Committee will compile your feedback and provide it to the SEES leadership in advance to planning meetings. Your response will be anonymous by default. If you wish to reveal your identity, please type it in one of your responses.

-The SEES User Group Executive Committee

Maryjo Brounce (chair), Bin Chen (vice chair), Dan Shim, Wenli Bi, Clair Zurkowski, Martin Kunz, Jin Zhang, and Wenlu Zhu

The survey received 36 responses, which are summarized in the following pages.

The survey:

1. What primary data products do you produce at a synchrotron?

2. What are the characteristics of the raw data and its metadata that should be archived (e.g., format, file size…)?

3. What is a reasonable length of an embargo period during which time your data are accessible only to you (two years, five years, upon publication of a manuscript, other)? Please provide any comments you may have,

4. What are the major advantages to a databased containing synchrotron source datasets?

5. What are the major dangers in or hurdles to creating a database containing synchrotron source datasets?

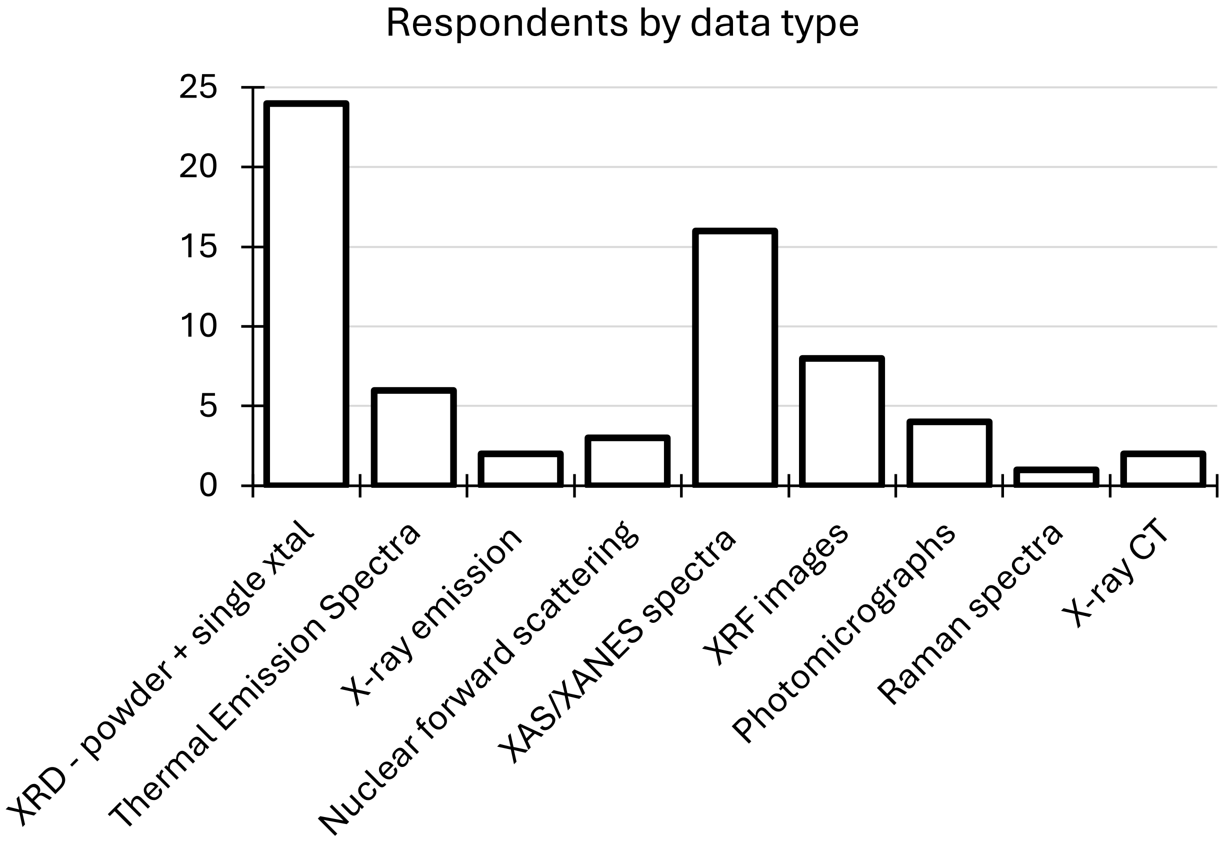
6. Please describe any ideas or concerns you have with respect to FAIR data reporting requirements, and the construction of a database that complies with these requirements.

7. Do you have suggestions for ways to streamline data archiving, information management, and data distribution after embargo at SEES beamlines?

Question 1: Self-identification/description:

What primary data products do you produce at a synchrotron?

The survey respondents are dominated by high pressure and XAS/XANES community, though it includes users that identify as producers of X-ray CT, NFS, and Raman spectroscopy as well.



Question 2: What is an archive to you?

What are the characteristics of the raw data and its metadata that should be archived (e.g., format, file size...)?

Respondents describe varied file types including *.tif, *.txt. *.dat, *.cbf, *.ascii and necessary logistical information concerning pixel sizes, beam energy, photon flux, detector distance, storage ring conditions (current, mode of operation). **They also quickly move on nearly unanimously to sample configuration/type/information/ source/treatment/composition, calibration, beamline or online notebooks, images of sample analysis point, meaningful/discoverable sample names etc.** They express immediately a sense that "everything needed to reproduce the analysis/results in the published paper" is necessary for a useful archive, that they would use. Representative quotes included below.

"everything should be archived in a way that anyone else can do the complete data analysis, independently"

"we are still collecting notes in paper notebooks but sometimes also use shared online notebooks"

"XAS is "easier" - the XDI format should be relatively sufficient. XRF imaging is more difficult, as the real need may be to save the full XRF spectral data so that future examination can look at the lowest level of data, not just the "pretty pictures". "

"Metadata are the critical aspect. Unfortunately there won't be one rule fits all, but to start with: sample source (synthesis, geological location, rime-core etc.), treatment (e.g. heated at 400 C under N2 atmosphere), composition, related information, in-situ conditions (p, T, time..)"

"The most important thing would be the sample info, this is the only info that is not recorded by the XRD tiff file"

"html photographs of analyzed points, with meaningful/discoverable/reusable file names."

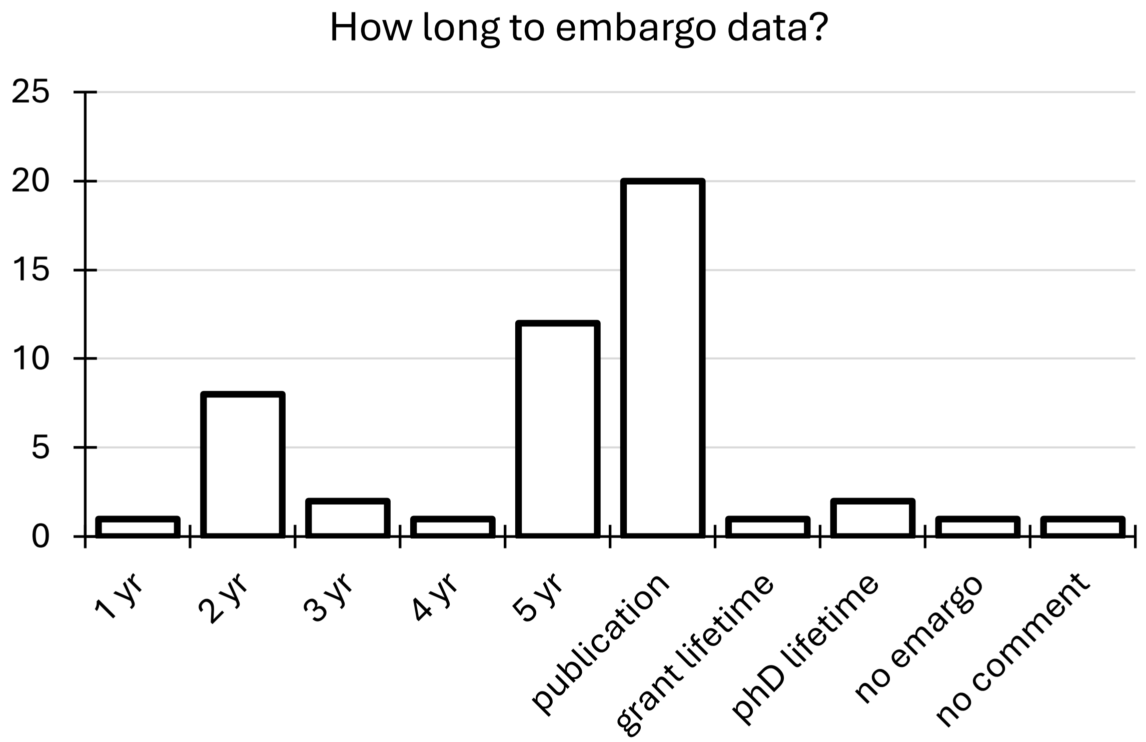
"every spectra linked to a physical sample with an IGSN"

"My experiments produce hundreds or thousands of relatively small (few hundred KB) images that record X-ray scattering intensities. In order to be processed for further analysis, these must be paired with metadata from a SPEC (text) file that contains information like X-ray wavelength, diffractometer angles, angles of incidence and exit with respect to a crystal surface, crystallographic unit cell and orientation parameters, etc. For data to be meaningful, additional information would also be required: what is the sample, what was its history, under what environmental conditions was it collected or grown, prepared, processed, handled, and measured. Those are generally recorded in notebooks."

Question 3: Embargo.

What is a reasonable length of an embargo period during which time your data are accessible only to you (two years, five years, upon publication of a manuscript, other)? Please provide any comments you may have.

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. Representative responses included below the histogram.



“At least five years to allow graduate students to process their data and publish the results. We should be protecting our graduate students who need time to learn.”

“two years or upon publication on manuscript; one should be able to extend the two years until the publication of the manuscript, should it happen in close foreseeable future”

“I think publication of a manuscript or maybe 5 years which ever comes first. I think two years can be problematic. We often discover something in our experiments that requires

development of new numerical methods or data analysis techniques and in those situations it can take some time to develop and validate those methods."

"Please consider what "accessibility" means. If someone knows that their spectra will enter the public domain in 1 year, what is to stop them from assigning meaningless or coded file names?"

*"My answer to the embargo question hinges on how FAIR will be *enforced*. If my data are truly FAIR, then I think the embargo needs to go to the end of a PhD thesis so that student thesis data isn't scooped. 5 years."*

"Upon publication of manuscript or five years - whichever comes first (if I haven't published in 5 years it seems reasonable to allow others to access data at that time)"

"The embargo issue can not be addressed by a default answer. Time to publication varies greatly by experiment. Some of the most valuable data is that which goes unpublished."

Question 4: The excitement.

What are the major advantages to a database containing synchrotron source data sets?

Some respondents were enthusiastic, citing value in making data available, searchable, permanent, reproducible, transparent. Some viewed this a protective measure against fabrication and fraud. The ability for future scientists to apply new processing techniques to what will ultimately become "legacy data" was viewed as advantageous. Some respondents stipulated their advantage, that it would only be advantageous if proper descriptions and metadata accompanied the files. Respondents also expressed concerns, saying that there is no advantage to a person who does not understand the sample or the measurement that occurred. **In other words, respondents used this space to underscore the importance of metadata and context for the "raw data products" produced at the beamline.** Representative responses included below:

"Permanence. Reproducibility of science to avoid fabrication, fraud etc."

"making data available and being trustworthy with reporting/publication"

"The ability for future scientists to use and compare to legacy data"

"Standardization of data; processing techniques change with time as do reporting standards. A database will help people keep up with changing techniques and often older data is presented in supplements in ways that make it difficult to access or reprocess."

"someone may reinvestigate the data, if they think that a structural model is flawed, or when better tools are available"

"open access democraticizes science"

"Other people can reuse the data with proper discription of the experiment and metadata"

"I honestly don't see much as it will be almost impossible to understand data without actually participating experiments. Even if we do the best job of making meta data, navigating all the complex record will be extremely challenging. But I would say it is still worth while because the group who produce the data will benefit from it."

"No advantage. If you did not run the experiments, it is pretty hard to know all the contraints for a dataset archived by somebody else."

"No existing databases that I know are particularly good for storing the kind of data we generate"

Question 5: The fear.

What are the major dangers in or hurdles to creating a database containing synchrotron source datasets?

Two primary concern are expressed by many respondents. The first is the fear that without appropriate metadata such as sample and or experiment details and context, the database will be useless, no matter how many resources are dedicated to building and maintaining it. The second is that this will introduce bureaucracy that users will either be overwhelmed by, or subvert e.g., through insufficiently transparent metadata, which would reinforce the first concern, that the database will not be useful. Representative responses below.

"People avoiding FAIR by providing insufficiently transparent meta-data."

"I feel that the effort to maintain this data for the public exceeds the gains of keeping them."

"People who study sensitive materials (e.g., of interest to national security, or proprietary materials) won't be able to use the beamline."

"no context, misinterpretation, more bureaucracy"

"I think one issue that can come up in the high pressure community is that some datasets can be export controlled and there needs to be some mechanism to ensure that availability of these dataset complies with appropriate export control policies. I think there are also some concerns related to intellectual property and making sure that individual receive appropriate credit (see answer below)."

"A growing pain may be that competitive scientists look through the database for "gotcha" data to rebutt publications they don't like."

"data theft, lack of proper credit attribution in publications"

"making sure that important information is included in metadata so that people can work with the data as they wish"

"Misinterpretation of another person's experimental parameters"

"Dangers are others using the data without permission, or using the data without a complete understanding of it."

"that the background information on sample/prep/parameters not well documented; long-term retrieveability needs needs to be supported"

"Determining which metadata are needed and collecting them consistently for diverse types of experiments. Differences in results can be due to many factors about how samples are created

and handled outside of the experiment. For truly innovative or original experiments, it may be difficult to imagine which metadata are needed going forward.
Users need to be trained to manage data.”

Question 6: Thoughts on FAIR.

Please describe any ideas or concerns you have with respect to FAIR data reporting requirements, and the construction of a database that complies with these requirements.

Respondents echoed the same concerns as in previous respondents – making data findable, limiting bureaucracy, and protecting data via embargo – and are neutral to cautiously positive about FAIR data reporting. Representative responses below.

"I have no concerns. I am 100% willing and able to participate in FAIR data provided that data are embargoed for a reasonable amount of time (at least 5 years to protect students)."

"In principle I think it's fine. In practice I think it will be hard to do in a manner that will be meaningful or useful."

"I think "warehousing" data could be straightforward, but the "findable" aspects, and how to make that useful may be a larger challenge."

"I think information on the type of reaction, substrates, products, compounds present in the DACs should be a part of metadata."

"I don't see a downside for sharing data and it's about time. I suspect it will be hard for everyone to agree on the "right" format, and I'm not sure total agreement is necessary. There are probably many different ways and different formats that different groups will use to share their data. Perhaps having a uniform and/or required format raises at least as many problems as it solves. So I recommend first-- just sharing our data, even if formats vary."

"FAIR data reporting sounds great to politicians who think that US taxpayers own our data, but it is not very practical."

"I think this is a good idea"

"If there is a database, it is important that there be complete information about the data, including what the sample material was and how they were collected (such as in a readme file). I am concerned about data being released before the scientist has a chance to publish it themselves."

"please avoid more bureaucracy - this is a serious concern!"

Question 7: Open forum.

Do you have suggestions for ways to streamline data archiving, information management, and data distribution after embargo at SEES beamlines?

Respondents had some suggetions, included below. Some respondents used this space to emphasize: limit bureaucracy, include metadata to make the database useful, and enforce the policy in some way.

"Data formats vary between facilities and beamlines, so interchangeability is key. Since deciding on and imposing a standard format is likely to fail, the best approach is to make formats interchangeable after the fact.
Data curation will be labor-intensive, so it will require real investment by SEES. "

"Probably would be difficult to get users to go through the trouble of submitting their own data to a database, unless adding to the database is contingent upon receiving future beamtime awards."

"In a perfect world, spectra data files link to sample archives (physical samples and their locations in sample databases and sample repositories)....Requires embargo."

"Perhaps this is too strict, and this is just one thought here. In order for SEES to list publications resulting from "SEES-supported beamlines", either on their website or in the NSF report, the publication should have two links: one to the DOI of the publication and a separate link to access the data. Many journals now require a data repository link, so the link can be easily found in the publication itself, so it's just a matter of copying that link or associating it to publication on SEES website.
If there is a SEES database that is created for particular files and types, a link to this database can be given."

"I don't have a suggestion, but I have a request that we try to find ways that don't become another hassle for researchers. Ideally, we would find an archiving infrastructure that is actually useful for the researchers, and it would be to their benefit to use the infrastructure - as opposed to one more bureaucratic thing we need to do for compliance. "

"I am expecting a suggested format for data sharing, possibly with a series of questions and uploading functions"

"the data should be stored right away after the experiment, the process should be straightforward - it could be somehow coupled with an electronic lab book where date, information on the compounds in the DAC, pressure etc. are stored; then wehn the embargo is lifted the data become available to the public and can be searched and accessed via a web browser"

"I have no ideas-- but hopefully there are groups and communities that are already doing this, and we can learn from them."

“Make sure relevant metadata is included in the saved files, so it can be easily extracted during the data upload to the new repository”