Question 1: Why curate sustain steward data?

This is a domain specific question. However there are commonalities: Re-use, although users drive meta-data creation. Metrics can usefully track use, such as telescope citation, so citation for all data is important in helping to make intelligent curation decisions. Reproducibility: in future keep any data that ties to published results. Keeping data can be a near trivial cost compared to producing the data.

We have use-cases, for example we know archival telescope data gets re-used.

What are the criteria for throwing out data: better data supercedes, there is slow down of new methods development for preprocessing, also what do you save by deleting data? Cost of access can be large, higher than cost of storage. It can be cheaper to keep data than retake or regenerate it in some cases. Otherwise becomes dark data if not accessed in 5 years.

Software and hardware both go obsolete and so, about every 10 years, you must do a migration which is hard and costly. But it is critical to have the meta-data indexed with the raw data, not separately. Need “self-described” data. [as opposed to indexing prioritizing fast retrieval]

So what are standards for trusted repos?

2. Ensuring data remains available after facility is closed down

All trusted repos are required to have an end of life plan. Need to get your house in order before shut down – but monetization not always clear. Need to save not always obvious.

Getting data ready for curation isn’t free (10-20% of project costs) and this has to be build into the cost of the project early on.

Could we partner with interested industry folks like Open Science Data Cloud? Funding from other sources not just NSF.

Broader impacts at NSF could include reuse of data and/or software, metadata, documentation, and if so what is the plan to preserve it?

Annotation and curation cheaper to do during collection rather than after the fact.
Identify data “lifetimes” after which you make a choice about discarding and keeping.

Cost of data preservation can be part of the criteria for keeping – with a baseline of dark storage so data is always kept.

Thinking about access starts with the value of the data: older data from a re-implemented physics experiment doesn't need to be kept, but astrophysics is different – you observe a supernova at a point in time. Need a set of criteria that takes history, needs of community and reproducibility, and costs into account.

If it's cheap to regenerate the data then you can throw it away. Also when another experiment supercedes another.

Tracking how datasets are being used is important. What queries are being submitted to the data?

Multiple copies are important.

If compute is near data that changes things from just data distribution.

Levels of data:

If data can be easily regenerated then can chuck.

Raw data in cold long term storage.

Old data is often a marginal cost compared to the last two or three years of data.

Ecosystem questions:

What are curation requirements? Alex: needs to be useable by the community.