Straw-man proposal for a "data collective"

Results of a preliminary requirements-gathering survey

In February 2016, as part of an investigation into how UCSB might provide long-term storage and curation of research data, researchers at UCSB's Earth Research Institute (ERI) and Marine Science Institute (MSI) were surveyed to ascertain their interest in a straw-man proposal for a limited-term data repository dubbed the "data collective," which was described as follows:

> *UCSB needs a long-term storage solution for research data. A few campus units (notably, ERI and MSI) offer disk space and other computational support to their researchers, but the units are not in a position to offer long-term storage guarantees. The UCSB Library's repository (ADRL)* is *committed to long-term curation of its content, but its scope is currently limited to Library collections.*
>
> *We're exploring the possibility of developing a hybrid storage solution that would support both*
>
> - *eventual curation of deposited research data; and*
> - *immediate use of the data, both within UCSB and by external research partners.*
>
> *As currently envisioned, such a storage service—a "data collective"—would be characterized by:*
>
> - *Guaranteed storage for 10 years. After 10 years, a dataset still in the collective, and deemed to be of lasting research value by the owner, would be moved to an appropriate long-term repository with the assistance of Library curators.*
> - *Modest ingest requirements. Copying data into the collective would be controlled, and some minimal data description would be required to help make the data discoverable and to support future curation, but the collective would not have the same rigorous submission requirements as a long-term repository.*
> - *Value-added services, possibly including search, assignation of persistent identifiers, citation generation, and version control.*
>
> *The goal is to create a repository where UCSB researchers can make all of their research outputs (datasets, but also posters, presentations, grey literature, etc.) available in a citable, shareable and discoverable manner analogous to eScholarship for literature. We believe the data collective would fill a vital need for researchers, as a repository for both primary data and research products that do not fit into present publication models,*

and as a necessary component of required data management plans in grant proposals.

45 responses were received. Given that the survey went out to approximately 114 researchers, and taking into account overlapping membership between the two institutes and excluding emeriti, this represents a 40% response rate. The remainder of this report summarizes the responses. Because of the small sample size, we have not attempted to perform any deeper statistical analyses of the data.

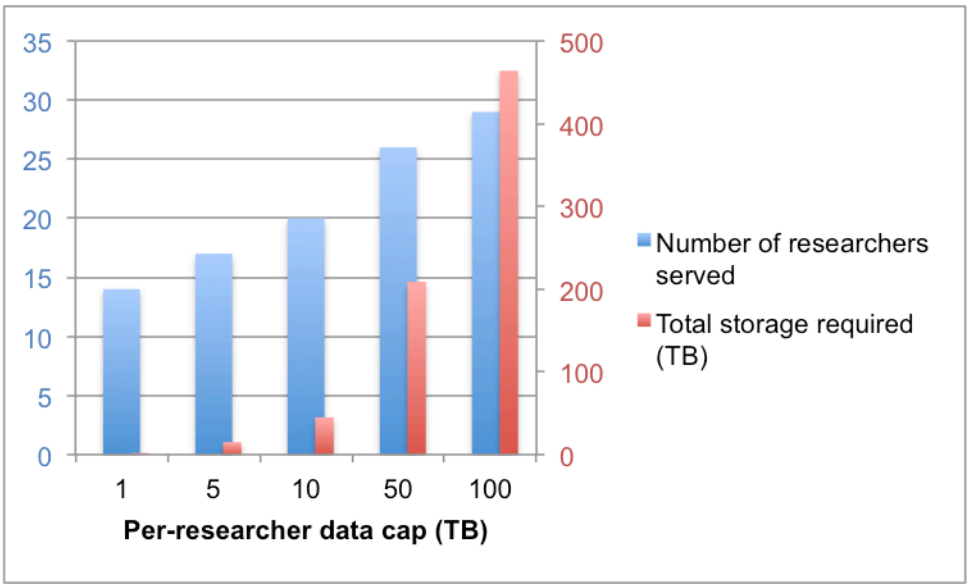
Q1. Would you find such a storage service useful?

The vast majority of researchers (93%) indicated affirmatively.

Q2. If yes, what type of data and how much data do you envision storing in it, and for how long?

Responses to this question are listed at the end of this section. Researchers who supplied dataset sizes in their responses identified, in aggregate, close to half a petabyte of storage required. Additionally, some datasets mentioned for which sizes were not given (remote sensing imagery, model output) are known to be potentially large. Given that these responses reflect only a fraction of the survey pool, it is clear that providing research data storage for even just ERI and MSI will put the data collective in the multiple petabyte range.

The problem of data size is further compounded by the wide distribution of dataset sizes. The following graph illustrates the effect of placing a hypothetical cap on the maximum amount of data a researcher may place in the collective, with the effect that researchers with larger datasets would be excluded. Of the 29 responses for which data sizes could be estimated, approximately half would be serviced if the cap were as low as 1TB. The other half of the responses exhibit a range of dataset sizes, with no clear clustering or inflection point.



In terms of retention times, a number of researchers identified retention times longer than 10 years, indicating that transitioning of data from the data collective to more permanent repositories is viewed as a key feature.

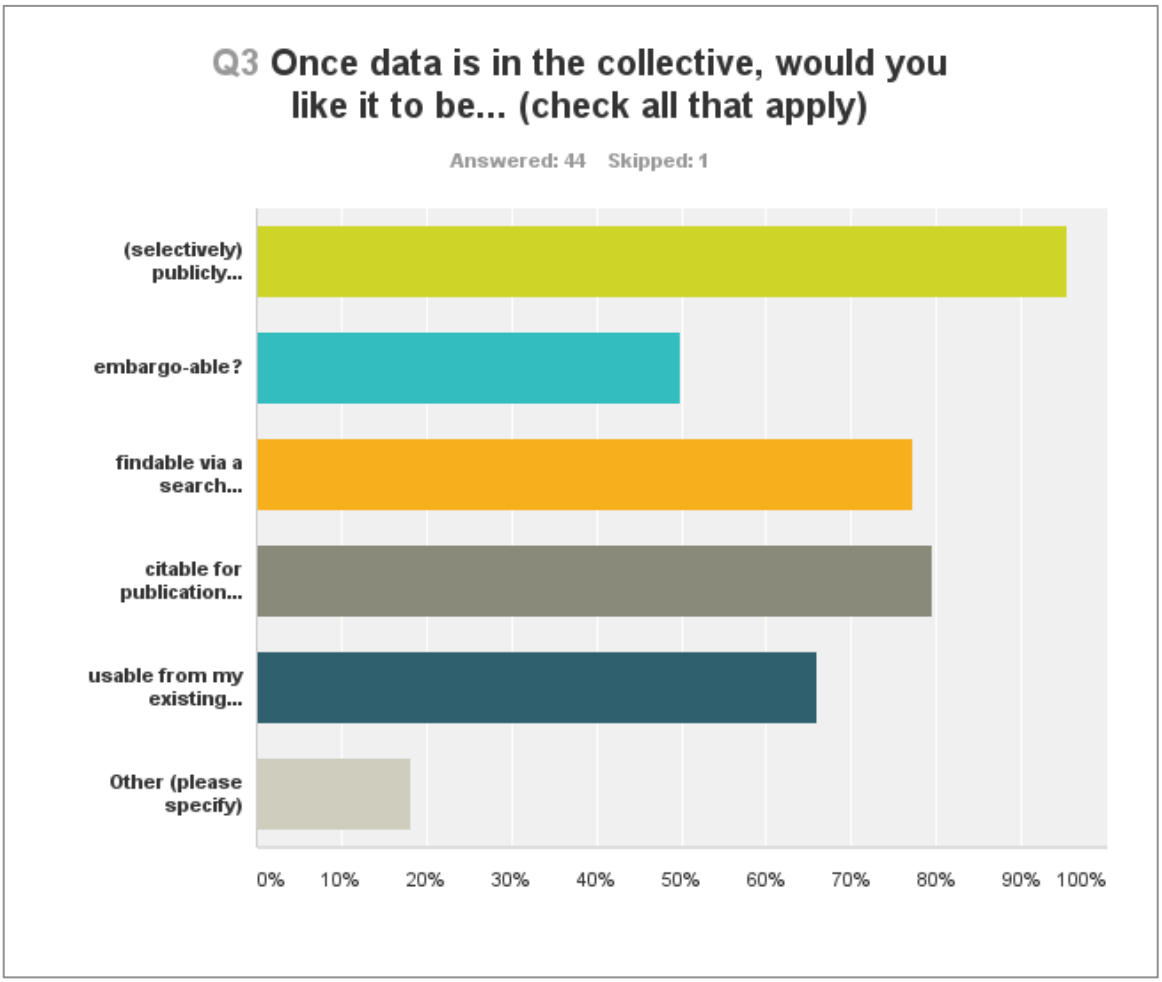
The individual responses to this question as follows:

1. Types: mainly satellite or aerial imagery and results of image analysis (land cover classification, etc.). How much: probably 50-100 GB. How long: 10 years plus possible long-term archiving sounds reasonable.
2. Video-photo data. Storage for 10 years would be good. My data would likely be in the 10-20 Tb range.
3. Experimental data (field and laboratory results), including environmental monitoring data. On the order of several GB total (not huge datasets).
4. Model output; 3 to 5 years.
5. Global and regional satellite data products not available elsewhere. Some TBs, more than 10, less than 1000. 10, 20,... years or until they become obsolete.
6. My data needs are small and would include spreadsheets, figures and data files from custom data reduction software.
7. Digital files of all kinds related to funding (extramural or other) and the data and findings resulting from that.
8. Digital data. About 2-3 Gbytes.
9. Remote sensing products, other data, several years (a decade?).
10. Zooplankton ecology and physiology (both sampling and experimental) biological data takes up relatively little room, likely 1TB >10 years.
11. Bioinformatic data.
12. Genomic and genetic data (e.g. text, FASTA, XML formats); 50Gb to 10Tb per project; 2-5+ years.
13. Model output data, indefinitely.
14. Environmental Sensor Data, Household Survey Responses, Derived Remote Sensing Data Products, UAV Data Collections.
15. Binary/ASCII environmental data collected from oceanographic platforms. Up to several TB.
16. Modeling outputs - over 100 of TB.
17. Ecological field survey and experimental data (tabular, modest volume of < 1 Tb), for at least 10 years; field environmental survey and monitoring data (tabular, moderate volume of 0.5-2.5 Tb), at least 10 years; geospatial mapping and modeling data (mainly rasters, moderate volume of 1-10 Tb), at least 10 years.
18. ~1 MB Excel or text files of geochemical data that are accessible >10 years.
19. Images, PDFs of reports, raw data. Starting 100GB, but will likely increase. Longer term storage.
20. Both ship and satellite data - 10 years.
21. Geophysical, few Gb, indefinite.
22. Field data, such as terrestrial LiDAR data; digital imagery and analyses. Stored as long as required by grantor such as NSF.
23. Data sets, presentations.
24. Databases of research results (experimental and modeling). About 1 GB /yr added and for 10 years.

25. Regional and global weather and climate model outputs. Difficult question to answer but it could be 1-20 Tb, 5-10 years.
26. 24/7 raw continuous seismic data (typically 200 sample per second, grows at approximately 2TB per year, current data set is ~18TB). Also, smaller project data sets, including an event database segmented out of the continuous database. Would need approximately 40TB of collective space to start which should provide room for approximately 5-10 years of growth.
27. I would not myself store data, but as a modeler could use it.
28. Order .5 - 1 terabyte over 10 years.
29. Video and image files. I am a seismologist and there are already public solutions for long-term data archiving and sharing waveform data. However, no such service exists for video or image data.
30. I presently have need for storage of about 10TB of data, from a number of sources. 1) sea floor image surveys; 2) video surveys; 3) shipboard data from navigations and various sensors; 4) genomic data; 5) other data types including spreadsheets.
31. Type: CSV files, SQL databases, program files. How much: ~ 1TB/year, currently ~10TB. How long: > 10 years.
32. Up to 50 TB of video data files, plus up to 5TB of comma delimited text files.
33. Time series meteorological and temperature data from lakes around the world. < 10 Tbyte. 30 years unless there is another repository for limnological data.
34. Mostly results from analyses of satellite data (the original satellite data are stored reliably by NASA, USGS, and other agencies). Volume would be 10s of TB (not 100s).
35. Largely biological survey data. Some spatial environmental data (custom derivations from publicly available sources).
36. GPS data, descriptions of physical samples, geochronology data (ages of geological materials).
37. Tabular data, plain text, code, posters and supplementary materials; until data are ready to be archived in a disciplinary repository.
38. Geochemical data, both compiled published data and original data. Re quantity, probably not huge -- maybe several Excel spreadsheet's worth. Best would be storage for decades.
39. Image data in TIFF, JPEG, and Photoshop format. 50 to 100 GB total.
40. I have several seismic reflection interpretation projects in the software Kingdom Suite. They range in size from about 4 gigabytes to over 20 gigabytes.
41. Survey data, needs HIPAA compliance.

Q3. Once data is in the collective, would you like it to be...

In terms of services offered by the data collective, the vast majority of researchers (95%) would like their data to be publicly accessible. Additionally, a majority would like features typically found in a full-fledged repository system: the ability to embargo data for a period of time; the ability to search for data; and support for persistent identification and citation of data for publication purposes. Additionally, two-thirds desire that their data be usable from their existing processing environments, which may necessitate that the collective provide filesystem and data service (e.g., OPeNDAP) access.

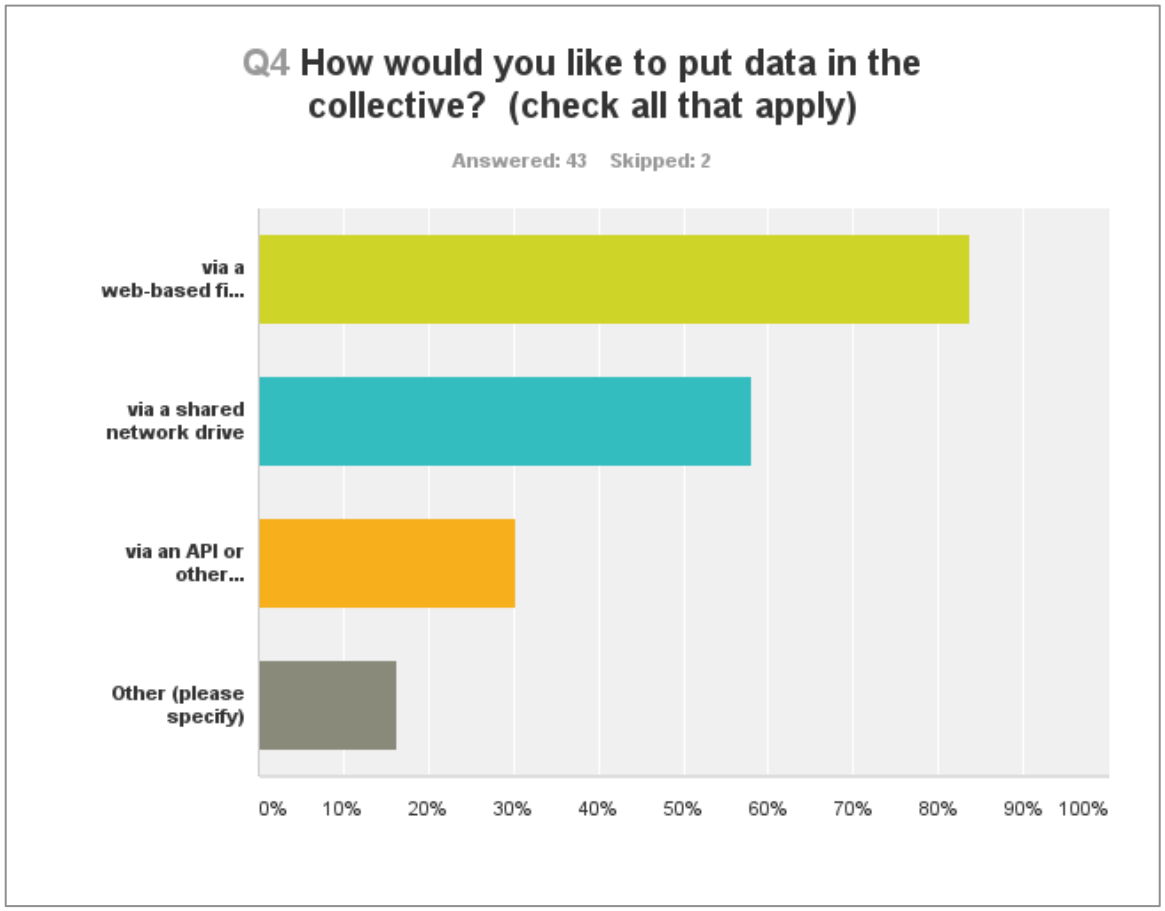


The "other" responses include:

- At a minimum, just plain ol' backup would still be useful.
- At some point it could be public. The data in the project is mostly in other databases, but the projects themselves are valuable to a few people working on these areas with these methods.
- Linked with metadata and compatible with submission to BCO-DMO as appropriate.
- Make fully and publicly accessible after some period, say 2-3 years.
- Some of the individual files would be large. Consider ability to subset.
- The "embargo-able" is important, if I understand correctly, since some data (e.g. from human subjects) has to be maintained as confidential. Similar situation with some industry collaborators.
- With DOI or other persistent identifiers.
- Would be great to have an API interface for data access.

Q4. How would you like to put data in the collective?

A clear majority of researchers would (and, it seems, could) use a web-based file upload interface. A majority also would like to use a filesystem interface such as a shared network drive.



The "other" responses include:

- Anything but a shared network drive.
- Globus Connect.
- We could monthly and/or annually update the data collection. rsync programmatic interface to the storage would require investigator access credentials to the storage collective.
- Whatever the method is fine.

Q5. Any other comments or questions?

The following general comments were received.

- All 3 input options are likely feasible, however I would choose the one that was best suited to off-campus input.
- Although I do think this is a good idea, we have already found a solution. Our lab uses Box and have our own data storage protocols. For formal data records we use ScienceBase, which generates a DOI.

- Another important archive needed is that for physical samples...
- I am highly supportive and would use this extensively.
- I converted two of the projects to the more easily available software OpendTect. But my latest interpretations are not in those. It is likely too much of my time and/or too tedious for me to convert all the projects to this software.
- I like the idea of the Library doing this. However, I question whether UCSB ought to host the storage itself. Please consider Microsoft Azure or Amazon Web Services to hold the actual bytes, and think about a connection that would make it work.
- I will shortly be retired, and would like a safe long-term archival place for both current and historical data as it is digitized.
- I would prefer that UCSB not create its own system, but rather make it easy to use one of the existing systems (e.g., Dataverse).
- Is this system meant to be for files only, or could databases be hosted as well?
- It would be great if there were a UCSB protocol for data management plans that was adaptable and that fulfilled grant requirements.
- It would be helpful to think about ways to facilitate storage of metadata for the data that would be stored. Not sure if there is some standard format that could be used across researchers.
- Most of the data that our research group collects or produces go to public archives with decent metadata standards. I don't anticipate that would change. However, a shorter-term option would be nice for when data are compiled but not yet ready for the archive. For example, this would be a good mechanism for graduate students to clean up their data and ensure a very solid backup, and to start becoming familiar with (and accustomed to) the concept and processes of archiving.
- Sounds like a good idea.
- This would help significantly with the Data Management Plan for NSF proposals.
- To better understand the utility of this service, it would be useful to have examples of proposed uses.