Workshop Report
Sustaining Data Repositories:
A Workshop on Creating and Implementing Sustainability Plans

Funded by the National Science Foundation: Award #1745596

Abstract:

In January 2018, ESA organized and hosted a workshop titled "Sustaining Data Repositories: A Workshop on Creating and Implementing Sustainability Plans." This workshop brought together principal investigators from a wide array of data repositories spanning a number of scientific disciplines. Through case study presentations, facilitated discussion, and group work on a process guide for Data Repository (DR) leaders, the workshop began to address the core question: how do DRs become more sustainable?

The outcomes from this workshop include the creation of a draft process guide and production of three DR sustainability case studies, in addition to a network of more connected DR professionals. Furthermore, the conversations that took place over the course of this workshop revealed that a diverse group of DR leaders from different disciplines faced similar challenges while managing large amounts of data and navigating the culture of data sharing. Group discussions also revealed insights into existing and potential solutions to these challenges, and opportunities for future actions and policies to increase DR sustainability.

Context & Objectives

The volume of data from scientific research is rapidly increasing. Archiving and sharing these data is essential to efficient research, as it allows for the accumulation and progression of knowledge, as well as collaborative discovery. As digital technology has improved, researchers have created effective data repositories (DRs), where these vast amounts of data are stored and accessed by a community of researchers.

This increasing volume of data, along with increasing requirements for public access to data collected using federal funds, means that DRs need correspondingly increased resources for data storage and analysis. In addition, emerging research themes require access to an increasing variety of data types. These trends, along with associated shifts in expectations, are creating new challenges for the data repository community in an increasingly uncertain funding environment. The Ecological Society of America organized a principal investigator workshop to help data repositories meet these new challenges, diversify their funding, and become more financially sustainable. The objectives of this day-and-a-half workshop were to:

- Identify challenges and opportunities specific to data repository sustainability,
- Identify challenges and opportunities of implementing different sustainability models, and
- Produce a process guide to help participants draft their own sustainability plans.

Twenty-two DR leaders, seven advisory committee members, and three ESA staff attended the workshop. Each participant was a PI on a data repository project. Together, they represented disciplines including biology, geology, engineering, computer science, math and physical science, and social science. ESA staff and workshop advisory committee members worked in collaboration with Nancy Maron of BlueSky to BluePrint to design a workshop that would achieve these objectives. The workshop featured

case study presentations, facilitated discussions, and group work on a step-by-step process guide to help Data Repository (DR) leaders create and implement their own sustainability plans. The workshop agenda and list of participants are included with this report in Appendix I and II.

Common Sustainability Challenges

Though the workshop featured attendees from a diverse array of fields, they reported common challenges in managing their data repositories. These challenges were identified during participant introductions, plenary discussions, and small group brainstorming. They are summarized in Table 1 below.

TABLE 1: Summary of Challenges to Data Repository Sustainability

Funding	Finding sustenance funding for research infrastructure is incredibly difficult, and many grants are aimed at funding new research at the expense of saving prior data. Many DRs rely too heavily on a single source of funding, which is not sustainable. Though some DRs have support from host institutions, the repository is not always well-integrated with that host. Other challenges: DRs are growing rapidly while funding flatlines, it's difficult to create public/private partnerships, and there is a lot of funding uncertainty and instability in agencies.
Data management	Lack of data standards, increasing computing costs, the rapidly increasing volume of data, and the challenges associated with making current data available for the long-term were all challenges that participants experienced. Tracking down older data and dark data/the lack of findable data were also mentioned.
DR community/ user culture	Some participants have experienced cultural resistance to data sharing within their user community and have encountered misconceptions that computing is free. Data sharing and management is typically not valued for career advancement. Other challenges included demonstrating value to users, offering a variety of services to users, and communication siloes between different researchers, agencies, and industries.
Unpredictable future	Understanding and predicting future sustainability challenges is daunting, especially when future data sharing policies and funding may be dependent on Congress.

Presentations from several participants offered a more in-depth look at some of these challenges. Eva Huala (The Arabidopsis Information Resource, or TAIR)), Xufeng Wang (NanoHUB), Bob Chen (NASA Socioeconomic Data and Applications Center, or SEDAC), and Myron Gutman (Inter-university Consortium for Political and Research, or ICPSR) gave short presentations on their data repository projects. A brief summary of their presentations is available in Box 1, below. These presentations highlighted the challenges of losing agency funding, matching revenue to costs, communicating value to host institutions (as well as navigating the constraints of university funding), dealing with increasing data-loads, and learning how to collaborate or compete with other data repositories.

Box 1: Summary of DR Sustainability Presentations

Eva Huala is the director of The Arabidopsis Information Resource (TAIR) and Phoenix Bioinformatics. TAIR is a plant genome database that curates available data into one gold standard. Though it was started with NSF funding, this funding was gradually reduced and ultimately lost. In response, TAIR transitioned from being open access to requiring subscriptions. They developed a subscription pricing model that assigned fees based on metered usage of data pages, meaning large universities, companies, and countries pay more than individual researchers. In addition, data are released for free access one year after being added to the repository. This scheme aimed to support scientists and ensure they had access to data, while also allowing TAIR enough income to continue providing high-quality curation and data management. Huala pointed out the fact that outside funding was needed to set this up, particularly in the research and development phase of the transition from free to fee (TAIR relied on external funding from Sloan to cover that transition).

Next, Xufang Wang presented the story of NanoHUB, a repository for nanotechnology simulations. In NanoHUB, researchers upload their simulations, which are then stored in the cloud and accessed by the nanotechnology community via web browsers. In addition to being downloaded by researchers, simulations are often downloaded by educators to use in teaching nanotechnology. Wang described NanoHUB as a bridge between research and education, and described the cultural shift involved in researchers sharing code. Simulation contributors benefit from the tracking tools available through NanoHUB, which allow them to understand how many people are downloading and using their simulations. Some of the challenges facing NanoHUB include the increasing costs of data storage; to address this, NanoHUB developers have gathered ideas from the industry sector and are applying compact models to the large amounts of data being stored. NanoHUB also benefits from a better understanding of its users, and to get this insight has developed tracking tools and worked with the NSF i-Corps Customer Discovery Team. Purdue, the host institution for NanoHUB, supplies substantial support every year, in part based on NanoHUB's strong value proposition for its contributions to the university.

Bob Chen, Director of the Center for International Earth Science Information Network (CIESIN), described the NASA Socioeconomic Data and Applications Center (SEDAC). SEDAC is a line-item on the NASA budget, meaning it has reliable funding. In addition, it's supported by Columbia University, which provides significant financial resources in addition to reputation, intellectual community, and availability of students and facilities through the university. In order to maintain NASA and Columbia support, SEDAC has to continue to communicate its value to host institutions.

Myron Gutman spoke about the Inter-university Consortium for Political and Research (ICPSR), which archives data in partnership with more than 750 international institutions. Gutman emphasized the importance of portfolio funding: for ICPSR, only 20% of revenue comes from subscriptions to the database, while the other 80% is a mix of grants and support from the host institution, University of Michigan.

Opportunities to Overcome Challenges

Between plenary discussions, breakout group discussions, case studies, and presentations from DR managers and NSF representatives, this workshop revealed that academic researchers are already pursuing solutions to the challenges involved in sustaining data repositories. While the points below present challenges to the DR community, they also offer opportunities for DRs to improve their long-term sustainability.

Support shifts in attitudes and practices around data-sharing.

Researcher attitudes towards data sharing can be a challenge for DRs. These attitudes vary across and within disciplines, with researchers holding diverse views on the acceptability and importance of sharing data. When data sharing is seen as a problem, it often stems from a concern about being scooped: researchers fear that data they gather and share will be analyzed and published on by other researchers before they are able to. Many data-sharing guidelines circumvent this concern by only requiring data to be shared upon publication of the research findings, but participants noted that data sharing works best when it happens during the course of the research, not after the fact. Other institutions have restrictions on who can publish on the shared data at certain timeframes, sometimes protecting the privilege of first publishing to the data gatherers (e.g. NIH Data Sharing standards).

Participants also noted a lack of professional recognition for data sharing, and highlighted the opportunity for tenure review boards to recognize contributions to shared data. For example, the one participant noted that the provost at their university recently approved appointments through new criteria (including datasets and data management contributions) in addition to publication records. In addition, participants noted the recent increase in data papers, which are publications of datasets that can contribute to a researcher's professional record.

Funding agencies are also susceptible to cultural attitudes that don't support data-sharing. Participants noted the perception that money spent on data management is money not spent on research, and that there's tension between competing needs for new research and old data management. Participants noted that they had encountered this idea among researchers and funders alike. One participant countered that idea, saying:

> *"I would like to see 100% funding on research, with the understanding that data sharing is the most cost-effective way to do that."*

Others supported a cultural shift towards data storage as an integral part of effective research. Oftentimes, they noted, depositing data in a DR is seen as an "afterthought," resulting in data being dumped into difficult-to-access DRs as time and money run out. These participants emphasized the benefits of viewing data sharing as a key part of the entire research flow, and suggested data management review take place throughout the course of a research grant, not just at the end.

Finally, participants noted that education initiatives would be key in bringing about a cultural shift around data sharing. They suggested incorporating data sharing and data management into standard science curricula (and highlighted the opportunity for students to gain training by working through the existing volume of uncurated data) and thesis requirements.

Optimize use of changing technology.

One of the noted challenges for DRs is the ever-increasing volume of data to be stored and shared. Data storage and management technology continues to develop, offering improved capabilities to manage this volume of data. Participants noted the importance of setting up DRs in a way that allows them to be responsive to future technological improvements. In addition, participants noted that tools and technology for managing data could be shared across disciplines.

Even with improving technology, participants noted an overwhelming volume of data. Beth Plale, NSF Science Advisor for Public Access, Office of Advanced Cyberinfrastructure, gave a presentation to workshop participants to outline one potential solution. She suggested that scientific communities work to establish tiers of data value. The highest tier includes data of the highest value to science and society (for example: original research, data needed to reproduce figures and statistical tests); the lower tiers are less valuable to science and society (for example: slide decks from a talk). Highest tier data would take top priority for data storage and sharing, allowing stakeholders to focus their resources on the most valuable data and avoid wasting time and money on low-value data. Plale emphasized that this tiering process needs to come from within the scientific community, which can best determine data value.

As a DR, make sure you are indispensable.

This point is at the core of a successful DR value proposition. Nancy Maron presented her findings from case studies of digital repositories and highlighted the fact that a strong value proposition is key to DR success. Maron emphasized value propositions and project sustainability as a cycle in which increasing project value allows for increased funding, which allows for increased project value, and so on.

Remaining valuable to users requires having a good understanding of what users want and need. Participants noted the importance of developing good communication with stakeholders. This is mutually beneficial, allowing users to have their needs known and met and DRs to establish a strong reputation, get a better sense of user needs, and improve their relevance.

One particular area where DRs could be more useful to users is in data curation. Participants noted the high costs (in terms of time, in particular) of data curation, and the difficulty of being responsible for this. Participants seemed to agree that the responsibility for data organization should fall to the researchers. However, there are opportunities for DRs to facilitate data curation. Examples described during the workshop include enforcing common formats for data uploads to a DR (and providing the incentive of automated analysis tools only if the data are correctly formatted), allowing user-driven creation of organizational meta-data fields, ad integration of DRs with ongoing research workflow (i.e. storing lab files throughout the research process, rather than only at the end of a study).

The question of how to quantify the value of DRs came up repeatedly over the course of this workshop. One option is to calculate the dollar value of storing and re-accessing data. One participant described an initiative by the UK's Wellcome Trust, in which database users were asked to estimate the amount of money they would spend if their DR disappeared, and then used these estimates to calculate a dollar value for the DR. Other participants thought that funding agencies should consider taking on the

challenge of calculating how much grant money goes to the collection of new data, versus accessing existing datasets.

Identify new funding models for DRs.

Maintaining and improving DRs will rely on continued revenue to support technology and staff costs. Over the course of the workshop, participants identified many different models by which a DR can create more reliable income streams. Nancy Maron presented her work on case studies of DR management. She found that successful DRs had diverse sources of income, including grants, institutional support, value-added services, and subscription fees.

Participants shared personal experiences and knowledge of the field to brainstorm opportunities for new DR revenue. Value-added services for data depositors was a popular idea. Services that depositors could pay for include: access to analytical tools for researchers to apply to their datasets (e.g. OpenNeuro offers a package of analytical tests for deposited data), as well as dataset tracking to see how their data are being used by others in the field (e.g. NanoHub allows model depositors to see how many others are running simulations using their model). In addition, depositors could pay for an organizational system that allows DRs to improve research workflow, curating and storing data in a way that's useful to the research process. Done effectively, this could also help bring about a cultural change in prompting more researchers to use DRs throughout the research process.

Charging subscription fees for data accessors was also a popular idea, particularly if subscription fees could be tailored to different user groups (e.g. TAIR modulates subscription fees based on usage and ability to pay).

Institutional hosts were also highlighted as providing a valuable source of stable funding to DRs; institutions can benefit from hosting a successful DR, as it contributes to the institution's reputation, and draws in professionals and students.

Continue and increase support agency funding for DRs.

While participants recognized the importance of finding diverse, new revenue sources (including options like charging subscription fees or selling value-added services), they also emphasized the importance of continuing to support data sharing in grant funding. They suggested that funding agencies could require data sharing costs to be a line item in proposal budgets, and that data sharing be an important part of project review. While this is already included in many grants (e.g. NSF Data Management Plans), participants reflected that there is little evidence of it being successful in bringing about effective data curation and sharing. They suggested that funders ensure consequences in project funding if data sharing isn't happening. Furthermore, they suggested funders increase collaboration among DRs by encouraging grant applicants to find existing DRs to which they can send their data, and noted interagency collaborations as an opportunity to focus on supporting DR projects.

Collaborate across disciplines and form new partnerships.

Participants noticed common challenges and opportunities for DRs even from just the sample represented in the room. They were excited about the possibility to expand these conversations to

other groups encountering the same issues and developing solutions. Developing partnerships has a number of benefits, including increasing the amount of relevant data in a DR, providing opportunities for collaborative research projects, and sharing useful tools and technologies. One participant described connecting US seismologists with seismologists around the world through a series of international workshops, allowing the creation of an international seismology DR that enhanced international collaboration.

Participants noted that it was important not to "reinvent the wheel" in developing partnerships. Rather than building new collaborations from scratch, DR leaders should consider joining or enhancing existing networks.

To extend the DR community, participants recommended contact with organizations across a range of disciplines, particularly those that are already working on DR questions, like Earth Science Information Partners, the Research Data Alliance (RDA), and the Committee on Data of the International Council for Science (CODATA). Participants highlighted cross-agency collaborations as an important area for growth. They also suggested reaching out to industry to learn from solutions being implemented in non-academic communities.

Table 2: Summarized Opportunities for Data Repository Sustainability

New partnerships	Could open up opportunities for DRs to work with industry partners, as well as consortia and professional societies. Partnerships could help DRs create new public/private funding agreements, initiate new data projects, and connect with other disciplines.
New policies & grant requirements	Could help overcome cultural barriers to data sharing. Some specific actions could include: • Integrating data sharing education and training in textbooks, online learning, and in university curricula • Automating admin policies to increase institutional knowledge and lower costs • Including data analysis and sharing successes in grant renewals • Including data production, sharing, and impact as a consideration in university promotion processes • Including data experts on review panels • Increasing coordination and education efforts on data IP/copyright/open licensing, and privacy/confidentiality concerns
New funding models & sources	Could open up new revenue streams from private foundations, multiple federal agencies, fees for services/products, and could help DR leaders identify ways to cut costs. There are also opportunities to explore getting financial support from other disciplines and how to turn users into financial contributors.

Making your repository an indispensable tool	Focusing on this can help DRs establish a reputation (in terms of value, quality, & expertise); create better, more useful services, stay relevant (e.g. meet stakeholder needs); stay current (e.g. have updated tech/services); know their users well; and broaden their user base. Knowing what sets a DR apart from others can help identify an operational or regulatory niche.
Collaborating with other repositories	Could promote efficiencies in sharing knowledge, tools, data, best practices, and infrastructure. Data can be linked to multiple repositories to enhance discovery. Collaborations can foster more cross-disciplinary connections.
Changes in technology	Could provide opportunities for efficient data management, improved curation, and new tools (but could also pose new challenges).

Workshop Products, Recommendations, and Next Steps

The products from this workshop include this report, three DR sustainability case studies, and the draft Process Guide. All products are available online at: https://esa.org/sbi/dr_sustainability/

Case studies

Prior to the workshop, Nancy Maron conducted a series of interviews with the leaders of three DRs with different approaches to sustainability: The Cambridge Structural Database (CSD) and the Cambridge Crystallographic Data Centre (CCDC); the Dryad Digital Repository; and TAIR and Phoenix Bioinformatics. Each case study describes the history and background of the DR, the current sustainability model, and the steps and decisions involved in implementing that model. Workshop participants read and reviewed these cases in advance, to spur their thinking and brainstorming around DR sustainability issues.

Process Guide

One of the goals of this workshop was to develop a step-by-step guide for other DR leaders to use when creating and implementing their own sustainability plans.

To do this, participants broke out into small groups, each of which covered one of four topics in the "Data Repository Process Guide." Each group was provided with a partially complete outline to edit, add to, and refine through a two-hour discussion block.

1. Defining your value proposition for stakeholders (including users, contributors, funders, and the general public)
2. Strategic planning and evaluation
3. Developing budgets and reliable, recurring sources of income
4. Implementing and revising your plan

Following the workshop, participants were asked to continue to edit their topic sections and were invited to attend group conference calls to review and approve those edits. After the series of calls, ESA

staff used the notes prepared by the participants to create a complete Process Guide document. This document exists online in a format allowing it to be an easily-accessed, easily-refined living document.

Recommendations

At the workshop's conclusion, participants identified concrete actions and policies that institutions and agencies could implement to increase DR sustainability. These are summarized below in Table 3. The conversations that took place over the course of this workshop revealed that a diverse group of DR leaders from different disciplines faced similar challenges while managing large amounts of data and navigating the culture of data sharing. Group discussions also revealed insights into existing and potential solutions to these challenges, and opportunities for future actions and policies to increase DR sustainability.

Table 3: Workshop Recommendations and Next Steps

Education & Academia	• Train students and project managers on the importance of data management and how to do it well. Incorporate data management into curricula and thesis requirements. • Advocate that data management and curation work be considered as important criteria in academic promotions.
Policies & Procedures for Grant-Making Agencies & Institutions	• Make data sharing a requirement on all grant applications. • Provide greater clarification and guidance on what successful data sharing looks like and what metrics should be monitored/documented. • A data repository's data sharing track record should be a more important part of the grant review process. Ensure consequences in funding if data sharing is not happening by the project review stage.
More Collaboration & Networking Across the Data Repository Community	• Organize more workshops for data repository leaders to form connections and network. • Make cross-disciplinary connections to other communities who are tackling similar challenges. Possible venues that could stimulate these connections include: the Research Data Alliance (RDA) and the Committee on Data of the International Council for Science (CODATA)
Topics for Future Workshops and Follow-on Activities	• Establishing best practices in data repository operations (could address cost efficiencies; data management, curation, and storage; quality control; how to adapt to and plan for rapid changes in technology; how to plan for and estimate data management costs). • Exploring the role that commercial partners/programs could play in data repository sustainability (particularly Public Benefit Corporations, SBIR/STTR Programs). • Promoting interagency cooperation on data repository sustainability issues and the future of science, with a goal of developing collective solutions across agencies. • Developing additional case studies on data repository sustainability, potentially in conjunction with a webinar series.

References & Resources

BRTF. 2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://blueribbontaskforce.sdsc.edu/

Downs, R.R. and R.S. Chen. 2012. Towards Sustainable Stewardship of Digital Collections of Scientific Data. Proceedings of Global Geospatial Conference, Québec City, Canada, 14-17 May 2012.
Ember, C. and R. Hanisch. 2013. Sustaining Domain Repositories for Digital Data: A White Paper. http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

Maron, Nancy L., K. Kirby Smith, and M. Loy. 2009. Sustaining Digital Resources: An On-the-Ground View of Projects Today. www.sr.ithaka.org/wp-content/mig/reports/4.17.2.pdf

MPS Open Data. 2016. MPS Open Data Workshop 1: Draft Report. https://mpsopendata.crc.nd.edu/index.php/draft-report.

National Science Board (NSB). 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. National Science Foundation, September 2005. Arlington, VA.

OECD. 2017. Business Models for Sustainable Research Data Repositories. OECD Science, Technology, and Innovation Policy Papers, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/302b12bb-en

Parsons, J.P. and C.S. Duke. 2011. Sustaining biological infrastructure: an ESA workshop report. ESA Bulletin October 2011: 426-432.

Parsons, J.P. and C.S. Duke. 2013. Strategies for Developing and Innovating Living Stocks Collections: An ESA Workshop Report. ESA Bulletin January 2013: 118-130.

Research Data Alliance (RDA). 2016. Income Streams for Data Repositories. RDA-WDS Cost Recovery Interest Group: Ingrid Dillo, Simon Hodson, Anita de Waard. V. 1.00, 10 February 2016 - Final Report, published prior to RDA Plenary 7, Tokyo, Japan.

Data Repositories

Biological and Chemical Oceanography Data Management Office (BCO-DMO)
https://www.bco-dmo.org/
BCO-DMO is a combination of the formerly independent Data Management Offices formed in support of the US Joint Global Ocean Flux Study and US Global Ocean Ecosystems Dynamics programs.

CERN Analysis Preservation Portal (CAP)
https://analysispreservation.cern.ch
CAP is a service for physicists to preserve and document the various materials produced in the process of their analyses (e.g. datasets, code, documentation) so that they are reusable and understandable in the future.

ChemBDDB
https://github.com/hachmannlab/chembddb
ChemBDDB is a big data database toolkit which facilitates the efficient management and sharing of chemical and materials data. It can setup a database, populate it with the data provided by the user, and enable the user to search the database, visualize each molecule in the database, and share the data via the web.

Citrus Genome Database (CGD)
https://www.citrusgenomedb.org/
The Citrus Genome Database, known as CGD, is a USDA and NSF funded resource to enable basic, translational and applied research in citrus. It houses genomics, genetics and breeding data for citrus species and organisms.

CottonGen
https://www.cottongen.org/
CottonGen is a cotton community genomics, genetics and breeding database being developed to enable basic, translational and applied research in cotton.

Databrary
https://nyu.databrary.org/
Databrary is a web-based data library for developmental scientists to securely store, manage, share, discover, and reuse research data, including videos, audio files, procedures and stimuli, and related metadata.

DataONE
https://www.dataone.org/
DataONE is a community driven project providing access to data across multiple member repositories, supporting enhanced search and discovery of Earth and environmental data. DataONE promotes best practices in data management through responsive educational resources and materials.

DesignSafe.ci
https://www.designsafe-ci.org/
DesignSafe is the CI component of the Natural Hazards Engineering Research Infrastructure collaboration. DesignSafe embraces a cloud strategy for the big data generated in natural hazards engineering research. It supports research workflows, data analysis and visualization.

Environmental Data Initiative (EDI)
https://environmentaldatainitiative.org/
EDI accelerates the curation and archiving of environmental data. In addition to providing a secure data repository, EDA provides training and resources to promote data management best practices and stewardship.

The Eukaryotic Pathogen Genomics Resource (EuPathDB)
https://eupathdb.org
EuPathDB is an integrated database covering the eukaryotic pathogens in specific genera. EuPathDB offers a unique entry point to similar taxon-specific databases and an opportunity to leverage orthology for searches across genera.

Genome Database for Rosaceae
https://www.rosaceae.org/
This is a curated and integrated web-based relational database providing data mining tools and publicly available genomics, genetics and breeding data for the Rosaceae family (almond, apple, blackberry, cherry, peach, pear, plum, raspberry, rose, and strawberry) to aid basic, translational and applied research.

Genome Database for Vaccinium (GDV)
https://data.nal.usda.gov/dataset/genome-database-vaccinium
The GDV is a curated and integrated web-based relational database. The GDV is being developed to house and integrate genomic, genetic and breeding data for blueberry, cranberry and other Vaccinium species.

Global Natural Products Social Molecular Networking (GNPS)
https://gnps.ucsd.edu
GNPS is a web-based mass spectrometry ecosystem that aims to be an open-access knowledge base for community-wide organization and sharing of raw, processed or identified tandem mass spectrometry data. GNPS aids in identification and discovery throughout the entire life cycle of data, from initial acquisition/analysis to post-publication.

Inter-university Consortium for Political and Research (ICPSR)
https://libraries.psu.edu/databases/psu01060
ICPSR provides access to the world's largest archive of computerized social science data, training facilities for the study of quantitative social analysis techniques, and resources for social scientists using advanced computer technologies.

Interdisciplinary Earth Data Alliance (IEDA)
https://www.iedadata.org/
The IEDA data facility's mission is to support, sustain, and advance the geosciences by providing data services for observational geoscience data from the Ocean, Earth, and Polar Sciences.

IRIS Data Management Center (IRIS DMC)
https://ds.iris.edu/ds/nodes/dmc/
The IRIS DMC archives and distributes data to support the seismological research community. They work closely with members of the seismology community to develop, host and distribute a diverse array of data products, as well as develop web-based tools to allow users to explore their data-holdings interactively.

iRODS Consortium
https://irods.org/
iRods is open source data management software. The consortium maintains and supports a commercial-grade distrobution if iRODS and brings together business, research organizations, universities, and government agencies to guide further software development and facilitate opportunities for education and collaboration.

Mass Spectrometry Interactive Virtual Environment (MassIVE)
https://massive.ucsd.edu
MassIVE is a community resource developed by the NIH-funded Center for Computational Mass Spectrometry to promote the global, free exchange of mass spectrometry data. MassIVE datasets can be assigned ProteomeXchange accessions to satisfy publication requirements.

Materials Data Curation System Software
https://www.nist.gov/programs-projects/materials-data-curation-system
The NIST Materials Data Curation System (MDCS) provides a means for capturing, sharing, and transforming materials data into a structured format that is XML based amenable to transformation to other formats.

NanoHUB
https://nanohub.org/
NanoHUB promotes computational nanotechnology research, education, and collaboration. They host a rapidly growing collection of simulation tools for nanoscale phenomena that run in the cloud and are accessible through a web browser. They provide online presentations, courses, animations, teaching materials, and more.

NASA Socioeconomic Data and Applications Center (SEDAC)
https://sedac.ciesin.columbia.edu/
SEDAC develops and operates applications that support the integration of socioeconomic and earth science data and serves as an information gateway between earth sciences and social sciences.

Netlib
https://www.netlib.org/
The Netlib repository contains freely available software, documents, and databases of interest to the numerical, scientific computing, and other communities. The collection is replicated at several sites around the world, automatically synchronized, to provide reliable and network efficient service to the global community.

Network for Computational Nanotechnology (NCN)
https://nanohub.org/groups/ncn
The NCN is a multi-university center that develops models and simulation tools to predict behavior at the device, circuit, and system level for nanoelectronics, nanoelectromechanics, and nanobio systems. NCN serves as a virtual laboratory to the nanotechnology community through online simulation and education.

NeuroVault
https://neurovault.org/
Neurovault provides a place where researchers can publicly store and share unthresholded statistical maps, parcellations, and atlases produced by MRI and PET studies.

NIST Public Data Repository
https://data.nist.gov/sdp
This repository houses data products developed and distributed by the National Institute of Standards and Technology, which span multiple disciplines of research and are widely used in research and development programs by industry and academia.

OpenNeuro
https://openneuro.org/
OpenNeuro provides a free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data. Users can upload data and collaborate with colleagues or share it with users around the world, browse and download datasets from other contributors, and use OpenNeuro's affiliated website to process applicable data.

Open Video
https://open-video.org/index.php
The Open Video Project collects and makes available a repository of digitized video content for the digital video, multimedia retrieval, digital library, and other research communities. The repository is also intended to be used as a test collection that will enable systems to be compared, similar to the way the TREC conferences are used for text retrieval.

PubChem
https://pubchem.ncbi.nlm.nih.gov/
PubChem is an open chemistry database at the National Institutes of Health (NIH). PubChem collects information on chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, toxicity data, and many others.

The Pulse Crop Database (PCD)
https://www.pulsedb.org/
PCD, formerly the Cool Season Food Legume Database (CSFL), is being developed to identify genes related to traits of interest among other methods to optimize plant breeding efficiency and research by providing relevant genomic, genetic and breeding information and analysis.

Sol Genomics Network (SGN)
https://solgenomics.net/
SGN is a clade-oriented database dedicated to the biology of the Solanaceae family, which includes a large number of closely related and many agronomically important species such as tomato, potato, tobacco, eggplant, pepper, and the ornamental *Petunia hybrida*.

The Arabidopsis Information Resource (TAIR)
https://www.arabidopsis.org/
TAIR maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.

Text REtrieval Conference (TREC)
https://trec.nist.gov/
TREC was started in 1992 to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC has successfully met its dual goals of improving the state-of-the-art in information retrieval and of facilitating technology transfer. The TREC test collections are large enough so that they realistically model operational settings. Most of today's commercial search engines include technology first developed in TREC.

Funded by the National Science Foundation: Award #1745596

Appendix I: Workshop Agenda

Objectives:

- Identify challenges and opportunities specific to data repository sustainability.
- Produce a process guide to help data repository leaders draft their own sustainability plans.
- Brainstorm collaborative, creative solutions to increase the sustainability of data repositories.

Agenda:

Thursday, January 25th:

8:30 am: **Arrival and Registration**

9:00 am: **Introductions and Workshop Objectives**
- Participants briefly introduce themselves, their data resource, and the biggest sustainability challenge they face

9:45 am: **NSF Perspective on Data Repository Sustainability**
- Overview of previous work on sustainability
- Overview of NSF priorities

10:15 am: **Sustainability Overview: Nancy Maron, BlueSky to BluePrint**

10:45 am: **Break**

11:00 am: **Case Study Presentation: Trent Alexander, ICPSR**
Each presenter conducts a 15-minute presentation, followed by a brief (5-minute) period for clarification questions.

11:20 am: **Case Study Presentation: Xufeng Wang, NanoHUB**

11:40 am: **Case Study Presentation: Robert Chen, SEDAC**

12:00 pm: **Plenary discussion: Themes and issues that emerge from case studies**

12:30 pm: **Lunch + Lunch Speaker: Nancy Maron**

1:30 pm: **Plenary discussion: Explanation of breakout group tasks**

1:45 pm: **Breakout Groups:** We will separate into four groups, according to the topics below. Each group will be tasked with developing a section of the process guide based on their topic:

- Group #1: Defining your value proposition for stakeholders (including users, contributors, funders, and the general public)
- Group #2: Strategic planning and evaluation
- Group #3: Developing budgets and reliable, recurring sources of income
- Group #4: Implementing and revising your plan

3:15 pm: **Break**

3:30 pm: **Breakout Reports:**
Each group will have 10 minutes to present their findings, followed by 15 minutes of group discussions to refine their ideas

5:15 pm: **Closing Remarks and Adjourn**

6:00 pm: **Group Dinner**
Delia's Mediterranean Grill and Brick Oven Pizza, Hoffman Town Center
209 Swamp Fox Road, Alexandria, VA 22314

Friday, January 26th:

8:30 am: **Arrival**

9:00 am: **Review of Day One**
Breakout groups quickly (5 minutes) present their final materials, based on feedback received during their report

9:30 am: **Group Discussion on Sustainability Challenges and Solutions**

- Consider the challenges and obstacles to data repository sustainability and how to overcome them.
- Identify frameworks, incentives, and policies to help.

10:30 am: **Break**

10:45 am: **Group Discussion on Opportunities and Next Steps**

11:45 am: **Closing Remarks and Next Steps**

12:00 pm: **Adjourn**

Funded by the National Science Foundation: Award #1745596

Appendix II: List of Participants

Karen Adolph
Professor, New York University
Data Repository: Databrary

Sean Ahearn
Professor, Hunter College
Data Repository: Data Science and Analytics (DSA BokMap)

Tim Ahern
Director of Data Services, Incorporated Research Institutions for Seismology (IRIS)
Data Repository: IRIS Data Management Center (IRIS DMC)

Trent Alexander
Associate Director, ICPSR
University of Michigan
Data Repository: Inter-university Consortium for Political and Social Research (ICPSR)

Nuno Bandeira
Associate Professor, University of California, San Diego
Data Repositories: MassIVE and Global Natural Products Social Molecular Networking (GNPS)

Robert Chen
Director, Center for International Earth Science Information Network (CIESIN)
Columbia University
Data Repository: NASA Socioeconomic Data and Applications Center (SEDAC)

Jack Dongarra
Professor, University of Tennessee and Oak Ridge National Lab
Data Repository: Netlib

Chris Gorgolewski
Post-doctoral Researcher, Stanford Center for Reproducible NeuroscienceData Repository: NeuroVault, OpenfMRI, OpenNeuro

Johannes Hachmann
Assistant Professor, University at Buffalo, SUNY
Data Repositories: molecularspace.org, CEPDB, ChemBDDB

Robert Hanisch
Director, Office of Data and Informatics
National Institute of Standards and Technology (NIST)
Data Repository: NIST Public Data Repository

Gerhard Klimeck
Professor, Purdue University
Data Repository: Network for Computational Nanotechnology

Rebecca Koskela
Executive Director, DataONE
University of New Mexico
Data Repository: DataONE

Kerstin Lehnert
Doherty Senior Research Scientist,
Lamont-Doherty Earth Observatory
Columbia University
Data Repository: Interdisciplinary Earth Data Alliance (IEDA)

Dorrie Main
Professor of Bioinformatics,
Washington State University
Data Repositories: PI of Genome Database for Rosaceae, CottonGen, Citrus Genome Database, Cool Season Food Legume Database, Genome Database for Vaccinium

Leah McEwen
Chemistry Librarian, Cornell University
Data Repository: PubChem

Funded by the National Science Foundation: Award #1745596

Lukas Mueller
Associate Professor, Boyce Thompson Institute for Plant Research, Cornell University
Data Repository: Sol Genomics Network

Margaret O'Brien
Information Management, Marine Science Institute, UC Santa Barbara
Data Repository: Environmental Data Initiative (EDI)

Jean-Paul Pinelli
Professor, Florida Tech
Data Repository: DesignSafe.ci

Arcot (Raja) Rajasekar
Professor, University of North Carolina
Data Repository: iRODS Consortium

David Roos
E Otis Kendall Professor of Biology
University of Pennsylvania
Data Repository: The Eukaryotic Pathogen Genomics Resource (EuPathDB)

Zachary Trautt
Materials Research Engineer, National Institute of Standards and Technology (NIST)
Data Repository: Materials Data Curation System Software

Ellen Voorhees
Computer Scientist, National Institute of Standards and Technology (NIST)
Data Repository: Text REtrieval Conference (TREC)

Xufeng Wang
Research Associate, Purdue University
Data Repository: Network for Computational Nanotechnology

Organizing Committee

Cyndy Chandler
Oceanographer Emerita, Woods Hole Oceanographic Institution
Data Repository: Biological and Chemical Oceanography Data Management Office (BCO-DMO)

Abhi Deshmukh
Professor, Purdue University

Myron Gutmann
Director, Institute of Behavioral Science
University of Colorado

Mike Hildreth
Professor of Physics, University of Notre Dame
Data Repository: CERN Analysis Preservation Portal

Eva Huala
Executive Director, Phoenix Bioinformatics
Data Repository: The Arabidopsis Information Resource (TAIR)

Gary Marchionini
Professor and Dean, UNC-Chapel Hill School of Information and Library Science
Data Repository: Open Video

Nancy Maron
President, BlueSky to BluePrint

ESA Staff

Cliff Duke
Director of Science Programs

Ellie Oldach
Science Outreach Intern

Jill Petraglia Parsons
Science Programs Manager

This report was prepared by:
Ellie Oldach, ESA Science Outreach Intern, and
Jill Petraglia Parsons, ESA Associate Director of Science Programs