

Data Management and Sharing

June 10, 2021

Moderator:

Jackie Bendall, Director, Contracts
& Grants Administration, COGR

Panelists:

Sandi Caldron, Data Repository
Outreach Specialist, Purdue
University

Gregory Farber, Director, Office of
Technology, Development and
Coordination, National Institute of
Mental Health, NIH

Lyric Jorgenson, Acting Assoc.
Director for Science Policy and
Acting Director of the Office of
Science Policy, NIH

NEW!

NIH Policy for Data Management & Sharing

Council on Government Relations
Data Management and Sharing Working Group
June 10, 2021

Lyric Jorgenson, PhD
Acting NIH Associate Director for Science Policy
Acting Director of the Office of Science Policy
sciencepolicy@mail.nih.gov



NIH Policy for Data Management and Sharing

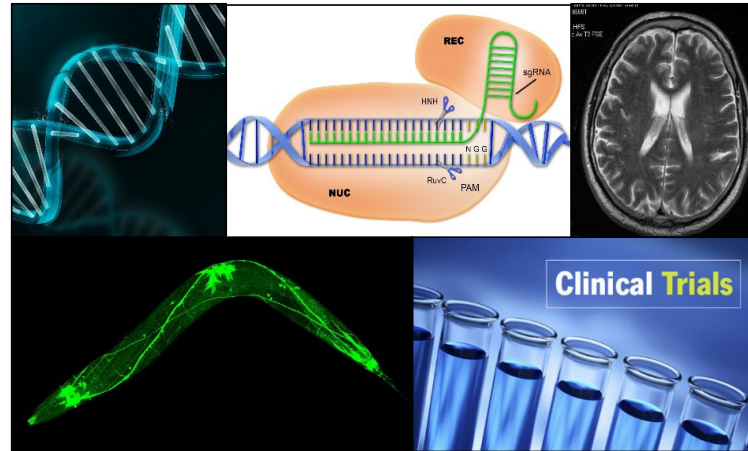
- **Submission of Data Management & Sharing Plan for all NIH-funded research** (*how/where/when*)
- **Compliance with the ICO-approved Plan** (*may affect future funding*)
- **Effective January 25, 2023** (*replaces 2003 Data Sharing Policy*)
- **Supplemental info available to assist**
- **Aims to foster data stewardship**

Why Share Data

*Sparks New Research
Collaborations*

*Enhances Rigorous
Study Design*

*Makes High-Value
Datasets Available*



*Enables Unique Data
Combinations*

*Facilitates
Study Validation*

*Stimulates New
Research Inquiries*



Maximizes Data Collection

(reduces redundancy/maximizes participant contributions)

Fosters Stewardship

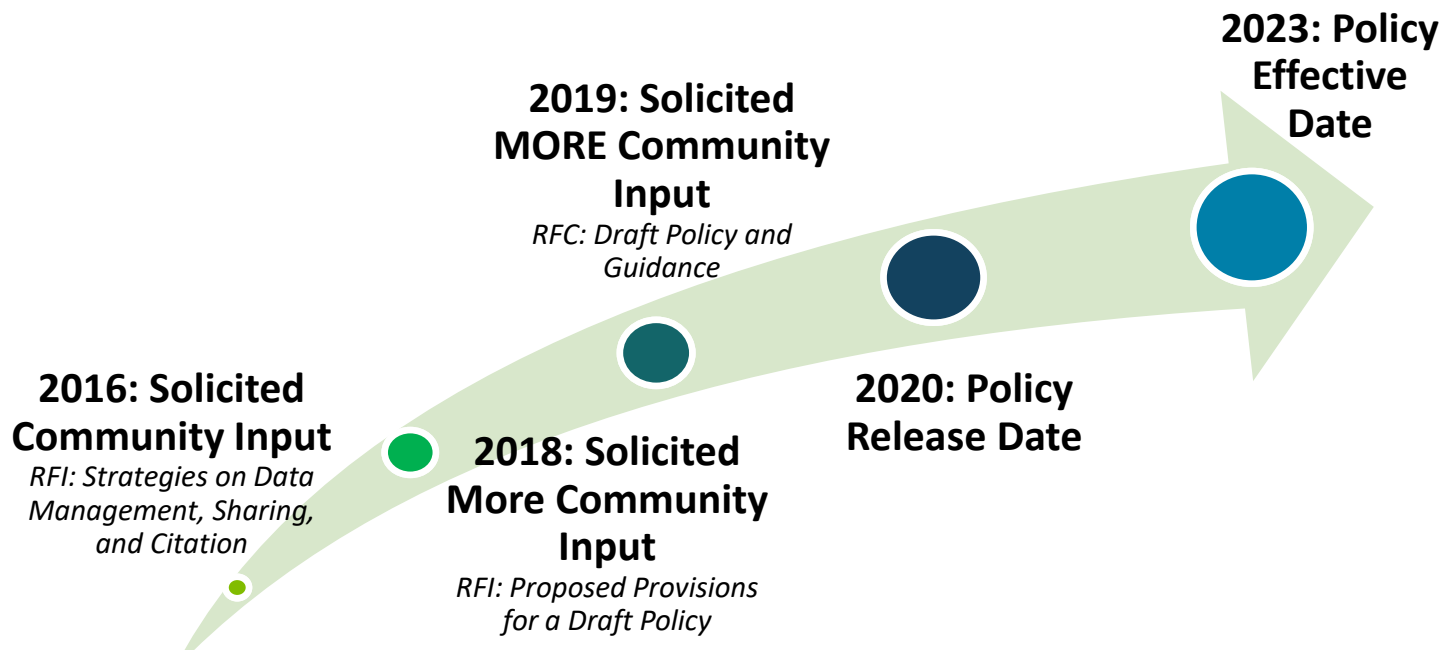
(provides transparency/accountability for taxpayer funds)

Accelerates the Research Enterprise

(for all the reasons stated above!)

An Iterative Policy Development Process

- Sought public comment repeatedly
- Tribal Consultation*
**Details provided in [“NIH Tribal Consultation Report: NIH Draft Policy for Data Management and Sharing”](#)*
- Intersection with other government agencies & Secretary’s Advisory Committee for Human Research Protections



The Devil is in the Details...

- **Scope:** All NIH-supported research generating scientific data
 - **Recorded factual material** commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications
 - **Does not include** lab notebooks, preliminary analyses, peer reviews, physical objects
- **Timelines:**
 - For **when to share data**, no later than publication or end of award (for unpublished data)
 - For **how long to share data**, consider relevant requirements and expectations (e.g., repository policies, retention requirements, journal policies) for minimum time frames



'Which brings us to my next point.'

Additional Expectations for Plans

- **SHARING SHOULD BE ...**

- **The default practice**

- All scientific data should be managed; not all scientific data must be shared
 - Maximize appropriate data sharing; Plans may justify exceptions (i.e., ethical, legal, technical factors)

- **Responsibly implemented**

- Plans should outline protection of privacy, rights, and confidentiality
 - Existing laws, regulations, and policies continue to apply

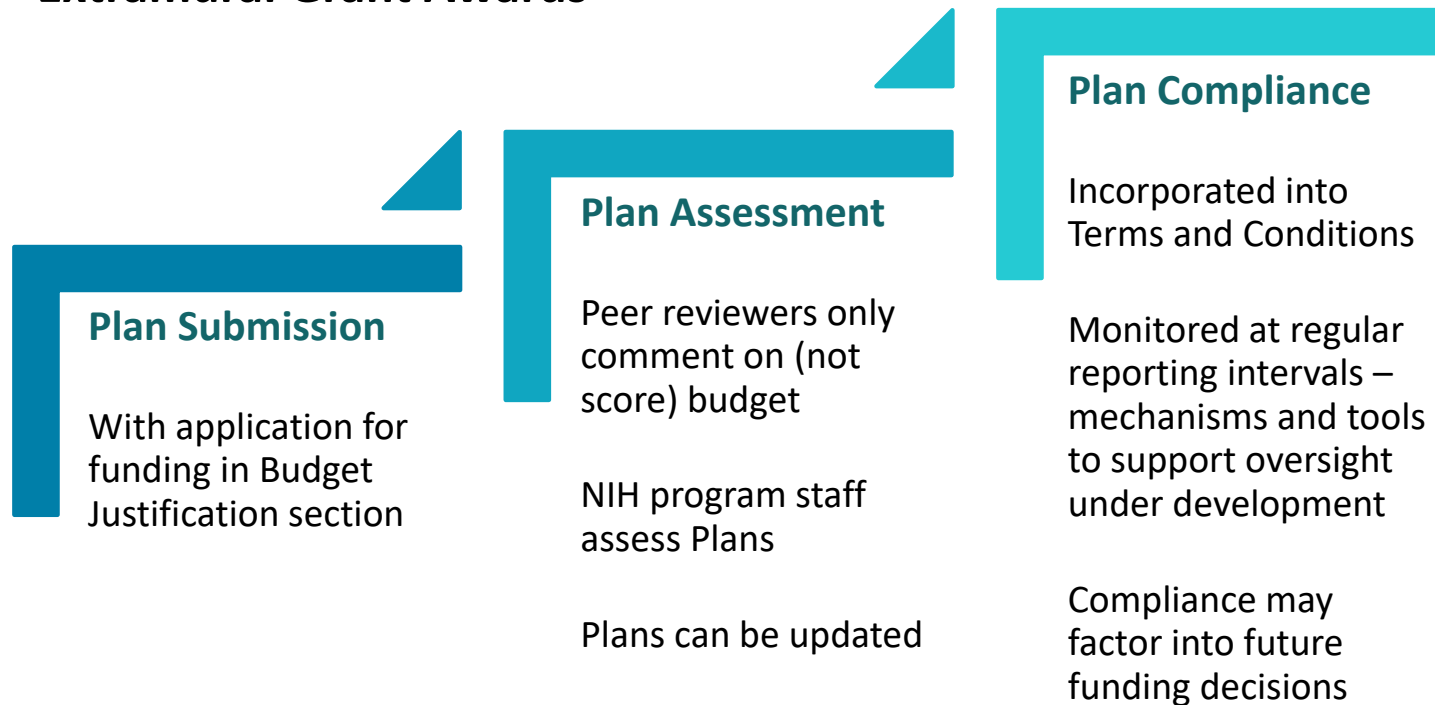
- **Prospectively planned for**

- During informed consent, including communicating how data will be used and shared
 - Data submission, including whether access to data, even if de-identified, should be controlled



Plan Submission and Review

Extramural Grant Awards*



**Analogous requirements for contracts, OTAs, IRP*



Supplemental Info to the Policy: Allowable Costs

- **Reasonable costs allowed in budget requests**
 - Curating data/developing supporting documentation
 - Preserving/sharing data through repositories
 - Local data management considerations
- **NOT considered data sharing costs**
 - Infrastructure costs typically included in indirect costs
 - Costs associated with the routine conduct of research (e.g., costs of gaining access to research data)

Supplemental Info to the Policy: Repository Selection

- Encourages use of established repositories
- Helps investigators identify appropriate data repositories
 - E.g., use of persistent unique identifiers, attached metadata, facilitates quality assurance
 - Refers to list of [NIH-supported Data Repositories](#)
- NIH ICs may designate specific data repository(ies)



Resources for Implementation

What's Next?

- **Engage in outreach** to develop additional supplemental information (including tribal-specific considerations)
- **Develop tools** for estimating data management and sharing costs (informed by the [2020 NASEM report on forecasting costs](#) & [April 2021 NASEM workshop on the culture of data management & sharing](#))
- **Develop approaches** for incentivizing good data sharing practices
- **Clarify interactions** with other NIH-wide (e.g., GDS Policy) and ICO-specific data sharing policies
- **Develop FAQs and other resources** to aid policy implementation

~2-year implementation window!

Resources for Implementation

What's Next?

- Engage in outreach to develop additional supplemental information (including tribal-specific considerations)
- Develop tools for estimating data management and sharing costs (informed by the [2020 NASEM report on forecasting costs](#) & [April 2021 NASEM workshop on the culture of data management & sharing](#))
- Develop approaches for incentivizing good data sharing practices
- Clarify interactions with other NIH-wide (e.g., GDS Policy) and ICO-specific data sharing policies
- Develop FAQs and other resources to aid policy implementation

**Your Thoughts:
What else?**

~2-year implementation window!

The NIMH Data Archive Cost Estimation Tool and Data Sharing Expectations

Greg Farber

Director, Office of Technology Development and
Coordination, National Institute of Mental Health

June 2021



National Institute
of Mental Health

Overview

- NIMH Data Archive
- NIMH Data Harmonization Activities
- NIMH and NIAAA Data Sharing Expectations
- Cost Estimation Tool



NIMH Data Archive


- NDA is a cloud-based data commons that (mostly) holds research data from human subjects
 - Federal data repository (<https://nda.nih.gov>)
 - Originally contained data from human subjects related to neuroscience research. We are now holding non-neuroscience data as a service to other NIH ICs. Most subjects have consented to broad data sharing.
 - Data are available to the research community through a not too difficult application process, although some data are freely available.
 - Summary data are available to everyone with a browser.
 - Web services are available for search, data dictionary, validation & submission, GUID generation (<https://nda.nih.gov/tools/apis.html>)
- Started in late 2006, and first data was received in 2008
- The data types include demographic data, clinical assessments, imaging data, and –omics data. There are no formal limits to the types of data that can be stored in NDA.




NDA Home Page

NIMH Data Archive

NDA | ABCD | CCF | OAI | NIAA_{DA}

 NDA

[Contribute Data](#)[Get Data](#)[Tools](#)[Webinars & Tutorials](#)[About Us](#)[MY DASHBOARD](#)




Welcome to the NIMH Data Archive

The National Institute of Mental Health Data Archive (NDA) makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery. De-identified human subjects data, harmonized to a common standard, are available to qualified researchers. Summary data are available to all.

The NDA mission is to accelerate scientific research and discovery through data sharing, data harmonization, and the reporting of research results.

Search NIMH Data Archive




CONTRIBUTE DATA

What would you like to do?

[PREPARE FOR SUBMISSION](#)

[SUBMIT DATA](#)

[ACCESS MY NDA COLLECTIONS](#)





GET DATA

What would you like to do?

[GET DATA beta](#)

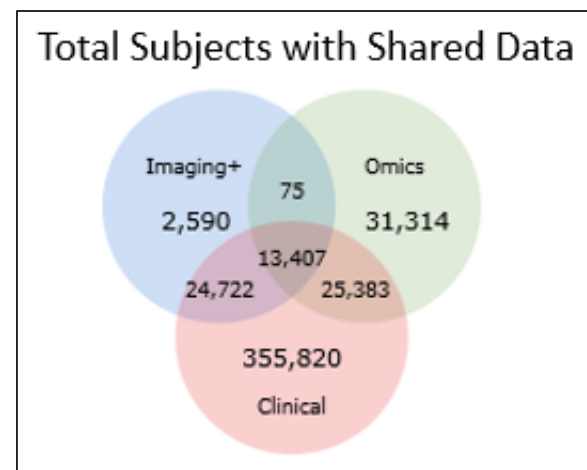
[REQUEST DATA ACCESS](#)

[SEARCH DATA DICTIONARY](#)



NIMH Current Status

- 500,000+ research participants in shared data
- 3,500+ approved users (access lapses after 12 months)
- 3.5 PB of imaging and omics data stored in the Amazon cloud
- 1,500+ collections (a data set associated with a grant award)
- 225+ studies (a data set associated with a publication and assigned a doi)
- 2,500+ different instruments in the NDA data dictionary.
- 300,000+ partially harmonized data elements



NDA Structure – Rows and Columns are the Building Blocks

- It is best to think of NDA as a large (~300,000 data elements by ~530,000 people), sparse, two dimensional matrix.

NDAR_demo.xlsx - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J
1	GUID	clinical assement #1 question #1	clinical assessment #1 question #2	clincial assessment #n question #m	link to raw MRI image	derived volume information from MRI	link to raw EEG	link to genomic data	crosswalk of IDs from other data repositories	
2	NDAR12345	a	1	5					Simons12345678	
3	NDAR12349	b	3	2						
4	NDAR18473	a		4						
5	Ped12345		2				link is here			
6	Cardio12934	a		2	link is here	34				
7	pseudo-GUID 3456	c	3	1						
8										

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from one collection



Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from two collections, no common subjects



Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from two collections, no common subjects, but there are common data elements. Harmonization can be “interesting”.



Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from two collections and there are two common subjects (7 and 8). Potentially there are issues here related to the owners of the two collections accessing other data about a single research participant.

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from one collection with additional derived information from a study added

NIMH Common Data Element Expectations

- NIMH in consultation with the Wellcome Trust and other funders of mental health research, has identified a minimal list of data collection instruments that we will use to harmonize mental health data collection.
- All awards made after May 2021 will be expected to collect data using the common data elements.
- The goal of this effort is to facilitate harmonization of data collected in many different laboratories. Given the heterogeneity in our diagnostic categories, a common measurement framework is essential to make progress.
- A number of other mental health funders are imposing similar expectations on their awardees, and several journals are considering joining this effort.
- Full Details can be found at [NOT-MH-20-067](#).



One Repository for NIAAA and NIMH Data

[NOT-MH-19-033](#) and [NOT-AA-19-020](#)

- Expects all applications to include a resource sharing plan that covers:
 - A summary of the data to be shared
 - The data standards/data dictionaries that will be used to collect the data
 - The schedule to validate the data against the standards/data dictionaries (NDA provides a useful validation tool that can be run without submitting data)
- Awardees will submit data (**not share it with the research community**) every 6 months to allow the awardee to perform validation checks close to the time when data were measured.
 - Data are **shared** after original project ends **or** upon publication
- Reminds the research community that they need to fill out section C.5 “Other Products and Resource Sharing” in the RPPR.
- Enumerates key expectations in NDA Data Sharing Terms and Conditions, which are included in the Notice of Award
- Establishes NDA as the NIMH and NIAAA-designated human genomics data repository



Post Award Expectations

- 1) PIs can't ignore the NDA Getting Started with Data Sharing notifications that they receive shortly after their Notice of Award has been released. All new awardees get an e-mail letting them know that they have to:
 - Submit a data submission agreement
 - Define the data structures they are going to use to submit data
 - Suggesting (but not requiring) that they use the NDA Global Unique Identifier (GUID) to identify subjects
 - Craft informed consent language that is consistent with sharing broad data sharing with qualified researchers (similar to general research use for genomic data)
- 2) If an institutional official receives a copy of a letter after the initial award, it means that there is a problem with data sharing that is being escalated.



Real Time Metrics for Grants Compliance

The screenshot displays the NIMH Data Archive web application. A modal window titled "Alerts for Collection" is open, showing a table of alerts for Collection 1860. The background interface includes a navigation bar with links like Home, Query, Harmonization Tools, and a sidebar with a "Collections (2)" section. The modal window contains a table with three columns: Color for Alert, Alert Title, and Alert Description.

Color for Alert	Alert Title	Alert Description
GREEN	DSA	DSA has been received
GREEN	Data Expected	Data Expected defined
GREEN	Initial Submission	A submission has been received
WHITE	Ongoing Data Submission	No status detected
WHITE	Upload of PII Suspected	No status detected
GREEN	NDA Studies for Publications	No relevant publications without an NDA Study
YELLOW	Publication Statuses	A recent publication status is undefined



Costs for Data Sharing

It is helpful to break down the costs related to data management and sharing into three phases.

- 1) Local data management that is needed as experiments are being conducted and the data are being analyzed.
 - Storage costs in the laboratory
 - Compute costs for the project
 - The infrastructure (hardware and software) may be laboratory based, department based, or institution based.
 - Charges to NIH can either be direct costs associated with the grant or as part of the institutional Facilities and Administrative costs.
 - It would be really interesting for NIH and for academic institutions to look at annual expenses for local data infrastructures as well as the (lack of) security for that infrastructure in comparison to what could be provided through commercial cloud based infrastructure.



Costs for Data Sharing

It is helpful to break down the costs related to data management and sharing into three phases.

2) Data curation and documentation

- These costs often only become apparent when a laboratory sends data to a particular data archive, when data are being analyzed, or when a paper is submitted to a journal.
- There are useful technologies (such as Jupyter notebooks) to keep track of exactly how data were collected (and analyzed).
- It is still true in many laboratories that the metadata containing experimental details aren't recorded appropriately because:
 - Experimental parameters don't change day to day in the laboratory (they think)
 - **Lack of agreed upon standards in a community to describe a particular experiment**
 - Data archives really help the community understand the magnitude of these problems and help the community reach consensus on the metadata that needs to be recorded.



Costs for Data Sharing

It is helpful to break down the costs related to data management and sharing into three phases.

3) Data repository charges

- Some repositories have a single up-front charge
- Some repositories have baseline funding from an organization (institutional library, funding agency, non-profit) with deep pockets that makes the repository available to a research community
- A few repositories allow free access to the research community (who donate most of the data) but charge industry to access the same data. This model is only open when the data are of clear value to an industry and the repository is mature enough to demonstrate that value.



NIMH Data Archive (NDA) Cost Estimator

- This tool helps researchers estimate how much staff time and effort will be required to work with NDA staff to deposit their data (phase 2 in the previous slides).
- The tool does not currently help researchers think about the other costs associated with sharing data.



NDA

https://nda.nih.gov

Apps

NIMH Web Applica...

News

NIH

NIMH Web Applica...

Jobs

Traffic

Travel

The BRAIN Initiative...

gender

Public - HCP Wiki

Instructions for Aut...

The B4Stage4 Philo...

Data

NIMH Data Archive

NDA

Contribute Data

Get Data

Tools

Webinars & Tutorials

About Us

MY DASHBOARD

Welcome to the NIMH Data Archive

The National Institute of Mental Health Data Archive (NDA) makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery. De-identified human subjects data, harmonized to a common standard, are available to qualified researchers. Summary data are available to all.

The NDA mission is to accelerate scientific research and discovery through data sharing, data harmonization, and the reporting of research results.

CONTRIBUTE DATA

What would you like to do?

PREPARE FOR SUBMISSION

SUBMIT DATA

ACCESS MY NDA COLLECTIONS

GET DATA

What would you like to do?

GET DATA [help](#)


REQUEST DATA ACCESS

SEARCH DATA DICTIONARY

NDA

[Apps](#)
[NIMH Web Applica...](#)
[News](#)
[NIH](#)
[NIMH Web Applica...](#)
[Jobs](#)
[Traffic](#)
[Travel](#)
[The BRAIN Initiative...](#)
[gender](#)
[Public - HCP Wiki](#)
[Instructions for Aut...](#)
[The B4Stage4 Philo...](#)
[Data](#)

NIMH Data Archive



NDA

[Contribute Data](#)
[Get Data](#)
[Tools](#)
[Webinars & Tutorials](#)
[About Us](#)

MY DASHBOARD

Data Contribution

[Cost Estimation](#) | [Informed Consent](#) | [Certificate of Confidentiality](#) | [Submission Training](#)

Welcome to the NIMH Data Archive (NDA)

This website is designed to help researchers receiving new grants for human subjects research from the National Institute of Mental Health (NIMH) become familiar with the expectations surrounding the sharing of their data and the process set up to deposit and share those data through the NIMH Data Archive (NIMH). This section of the website will walk you through the process to prepare your Data Sharing Plan and submission project, make your data as consistent as possible with data from other laboratories, and report publications to NIMH. Please contact us at NDAHelp@mail.nih.gov with any questions or comments.

NDA accepts human subjects research data related to mental health. If you are expected to or planning to contribute data, you must meet the following prerequisites:

1. **Obtain [Informed Consent](#).**
2. **Collect from participants the PII needed to create an [NDA GUID](#).**
3. **Estimate the cost of data sharing (see [Cost Estimation](#)).**

Once the prerequisites are considered, follow these steps:

1. **To understand NDA, read the [NDA Policy](#).**
2. **Complete the Data Submission Agreement**
The principal investigator must complete the [Data Submission Agreement](#) and have it signed by a Signing Official at their NIH-recognized institution. If you don't know your institution's Signing Official, contact us at the [NDA Help Desk](#) for assistance. Once completed, this form should be emailed to the [NDA Help Desk](#).
3. **Request an NDA Account**
Create an account [here](#) to use in creating GUIDs, submitting data, and managing your project's Collection. All personnel working on the project including the PI will need to create their NDA account.
4. **Define Data Expected for Submission**
Create a list of all the data that you plan to collect, as defined in [NDA Data Dictionary](#) structures along with the number of subjects, the initial submission date and the anticipated share date (based on the [Data Sharing Regimen](#)). This list must be created directly in the Data Expected tab of your Collection, once you have created your account and we have received

Contribute Data

[Guide Notices](#)
[Sharing Regimen](#)
[Data Standards](#)
[Data QA Reporting](#)
[Manuscript Preparation](#)
[Submit Data From a Paper](#)

AutoSave Off NDA_Data_Submission_Costs - Excel Farber, Greg (NIH/NIMH) [E] Share

File Home Insert Page Layout Formulas Data Review View Help Acrobat Power Pivot Tell me what you want to do

Paste Font Alignment Number Styles Cells Editing

H8

NDA Data Submission Planning Cost and Effort Worksheet
<http://nda.nih.gov>
 Need help? ndahelp@mail.nih.gov

Replace sample answers with information tailored to your research project.

Questions	Sample Answers	Comments
1. How many subjects will be included in the research project?	300	NDA expects to receive data for all subject from whom data were collected including affected individuals, controls, parents, and siblings
2. How many sites will collect data from subjects in this project?	2	
3. How many times will data from this project be submitted to NDA?	11	Includes twice-yearly submissions of descriptive (phenotypic) data over the life of the grant and one final submission of experimental data
4. How many data structures will be submitted?	12	This is the number of data structures which are already defined in the NDA Data Dictionary, e.g., ADOS, ADI-R
5. How many unique experiments (e.g. omics, EEG, eye tracking, fMRI) will be conducted in the study?	5	Add the number of experiments only if the project includes genomics data collection.
6. How many NDA aims involving human subjects are defined for this project	3	Researchers are expected to share analyzed data using the NDA Study feature for each aim of the project.
7. What is the hourly rate charged for the Principal Investigator?	\$100	Hourly rate in US Dollars
8. Will the project have a Data Manager?	Yes	Click the cell and use the drop-down arrow to select "Yes" or "No"
9. What is the hourly rate charged for the Data Manager?	\$45	Hourly rate in US Dollars
Total Cost:		\$13,030

Sheet1 Sheet2 Sheet3

Ready 100%

AutoSave On										NDA_Data_Submission_Costs - Excel										Farber, Greg (NIH/NIMH) [E]									
File Home Insert Page Layout Formulas Data Review View Help Acrobat Power Pivot Tell me what you want to do										Share																			
Clipboard Font Alignment Number Styles Cells Editing																													
H15																													
A B C D E F G H I																													
14																													
15																													
16																													
Task										Investigator Hours																			
Data Submission Agreement										Data Manager Hours																			
Review, Complete, Sign and Submit to NDA										Comments																			
Incorporate Informed Consent Language										A Data Submission Agreement is required to receive permission to submit data																			
Define Data Sharing Schedule										http://NDA.nih.gov/contribute_informed_consent.html																			
Review NDA Capabilities										Review and Update Data Expected and Data Submission																			
Request Accounts										Timelines specific to the grant																			
Review Collection and Grant Permissions to Staff										Review of NDA Web site and capabilities; Webinars																			
Insert GUID into Existing Research Workflow										http://NDA.nih.gov/contribute.html																			
Develop Export/Import Procedures (Year 1)										8 hours per site																			
Setup and Test GUID Procedure										8 Hours per Site																			
Develop Data Extract using StaNDAd Assessments										1 hour per Data Structure																			
Ongoing Data Submission (Grant Years 2-5)																													
GUID Generation										Assumes 5 minutes per subject to generate a GUID																			
Hours to Complete Submission (Factored per Submission)																													
Use Validation Tool to Validate/Correct Errors and Submit										Resolve any data quality issues identified by NDA																			
Review Submission in NDA										Validation Tool (4 hours per submission)																			
Review and Resolve Post Submission QA/QC Issues										Factored by Number of Submissions (1 hour per submission for PI, 2 hours per submission for DM)																			
Genomics, EEG, Eye Tracking, fMRI and Imaging										1 hour of training per person																			
Experiment Definition Tool Tutorial										2 hours to define each experiment																			
Define Experiments using Experiment Definition Tool																													
Post Publication Study Setup and Definition																													
Create and Share Study										4 hours per aim times 2 publications (see http://NDA.nih.gov/data_from_papers.html)																			
Total Effort in Hours:										52 174																			
Total Cost:										\$13,030																			
Sheet1										Sheet2 Sheet3																			

What isn't Included in the Estimator

- Time required to define new data dictionaries. This can be significant, but only needs to be done once. Defining a data dictionary involves significant work by NDA staff as well as the group depositing the new data dictionary.
- The cost estimates are for the researcher and do not include NDA staff effort. NDA pays those costs.



NIH Executive Vacancy

- Deputy Director, Office of Extramural Research
- Deputy to Mike Lauer
- The OER Deputy Director's role integrates a broad array of senior-level science administration, extramural research policy, and business functions of the OER, ensuring they are executed and supported through leadership of a sound strategic planning, business management, and communications.
- <https://hr.nih.gov/jobs/executives>
- Applications due June 30, 2021



Purdue University Research Repository

research data management for Purdue

Sandi Caldron
Data Repository Outreach
Specialist
scaldron@purdue.edu

Purdue University

research culture deeply rooted
in the midwest and focused on
the persistent pursuit of the
next giant leap

History and Context

- R1 land grant university in Indiana
- ranked #5 most innovative university in U.S. (US News and World Report 2021)
- ~45,000 students
- ~2,000 tenure track faculty

Research at Purdue


- 135+ research centers and institutes
- Discovery Park - multidisciplinary groups and outside partners
- \$454 million in sponsored research (2018)

Our COVID year

- went virtual in March 2020
- 95% of research labs were back online by July 1, 2020
- 70% of employees working partially or fully remotely
- over 50% of students attended classes in person on a de-densified campus

CENTRALIZED RESOURCES FOR DATA-DRIVEN RESEARCH AT PURDUE


RESEARCH DATA DEPOT



for storing, using and sharing active research data of any kind as well as codes (purchase by the terabyte)

www.rcac.purdue.edu/storage/depot/

GLOBUS



for fast data transfer and sharing with on- and off campus collaborators (no charge)

<https://transfer.rcac.purdue.edu/>

FORTRESS



for extensive, long-term archiving of research data (no charge)

www.rcac.purdue.edu/storage/fortress/


PURDUE UNIVERSITY RESEARCH REPOSITORY



for meeting data management plan requirements and publishing and publicly sharing data

<http://purr.purdue.edu>

COMMUNITY CLUSTER PROGRAM



Purdue's Rice, Snyder and Hammer clusters currently have capacity available for purchase.

RICE: for high-performance parallel computation

SNYDER: for data-intensive computation, requiring large memory systems, especially in the life sciences

HAMMER: for loosely coupled, high throughput, serial computations

www.rcac.purdue.edu/services/communityclusters/


HATHI



a shared Hadoop cluster available to partners in Purdue's Community Cluster Program for working with big data using Hadoop (no charge)

www.rcac.purdue.edu/compute/hathi/

SCHOLAR



for instructors wishing to incorporate supercomputing in a course (no charge)

www.rcac.purdue.edu/compute/scholar/

DIAGRID



a Science-as-a-Service platform for easy Web-based access to software applications used by thousands of researchers around the world with computational resources on the back end (no charge)

www.rcac.purdue.edu/services/diagrid/

XSEDE



the National Science Foundation's Extreme Science and Engineering Discovery Environment, featuring some of the world's largest supercomputing systems (Purdue is an XSEDE partner)

www.xsede.org


SCIENTIFIC SOLUTIONS GROUP



help accessing XSEDE; computational application design and data set management; software development and consulting; and collaboration on grant proposals

www.rcac.purdue.edu/services/software/

ENVISION CENTER



data visualization and analysis; virtual simulation; human-computer interaction; and media creation, including video, animations and publication-quality stills

www.rcac.purdue.edu/envision/

HUBZERO



HUBzero: Web-based platform for deploying computational research codes and visualizing and analyzing results online; built-in social networking to create virtual research communities; and data management and interactive database capabilities

<https://hubzero.org>

More information: email rcac-help@purdue.edu or contact Preston Smith, director of research services and support, psmith@purdue.edu or 765-494-9729 EA/EOU

PURDUE
UNIVERSITY

PURR

Purdue
University
Research
Repository

founded 2011

Research Data Management for Purdue

The Purdue University Research Repository (PURR) provides an online, collaborative working space and data-sharing platform to support Purdue researchers and their collaborators.

Full spectrum support from planning to publication



PLAN

Meet funders' data requirements with our



COLLABORATE

Your own file sharing space and update feed



PUBLISH

Publish your data, get a DOI

PURR Origins

key institutional partners and motivations for founding PURR in 2011

Libraries

long-term stewardship and
access to data as a part of the
scholarly record, library and
information science expertise

Research Office

more competitive proposals
and compliance with funder
requirements

IT

research computing expertise,
e.g., storage engineering, HPC



How PURR works for researchers

The Research Data Lifecycle in PURR

PLAN

expertise

review DMP best practices, and consult with experts

COLLABORATE

private space

use secure, online storage space to share files with collaborators

PUBLISH

public space

meet funder requirements for open access and receive a DOI

ARCHIVE

preservation

PURR partners with the MetaArchive preservation network

IMPACT

tracking

automatic counts of views and downloads, plus citation tracking

[Home](#)[Datasets](#)[Projects](#)[Help](#)

Michael Witt ▾

Michael Witt: Dashboard

[Dashboard](#)[My Groups](#)[My Projects](#)[My Impact](#)[+ Add Modules](#)

WELCOME



Welcome to the Purdue University Research Repository (PURR)!

If you're wondering where to get started, consider viewing our [Data Management Plan \(DMP\)](#) section to

My Projects

[All Projects](#)[+ New Project](#)[FERPAtest](#)

By me | manager

[2012 Latino National...](#)

By James A McCann | manager

2

My Tickets

[All My Tickets](#)[+ New Ticket](#)

Tickets I Submitted

[#1458: Can you write a short KB...](#)

6 months 2 weeks ago 1 comments

Tickets Assigned To Me

[Home](#)[Datasets](#)[Projects](#)[Help](#)[Share your data](#)[Login](#)

220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3

Listed in [Datasets](#)

[About](#)[Supporting Docs](#)[Versions](#)[Citations](#)[Usage](#)

By Marion F. Baumgardner¹, [Larry L. Biehl](#)¹, David A. Landgrebe¹
Purdue University

This publication includes the AVIRIS hyperspectral image data for Indian Pine Test Site 3 along with the reference data for this site including observation notes and photos for the fields within the approximately 2 mile by 2 mile area.

[Download Bundle \(1 GB\)](#)

Additional materials available

Version **1.0** - published on 30 Sep 2015 doi:10.4231/R7RX991C - [cite this](#)

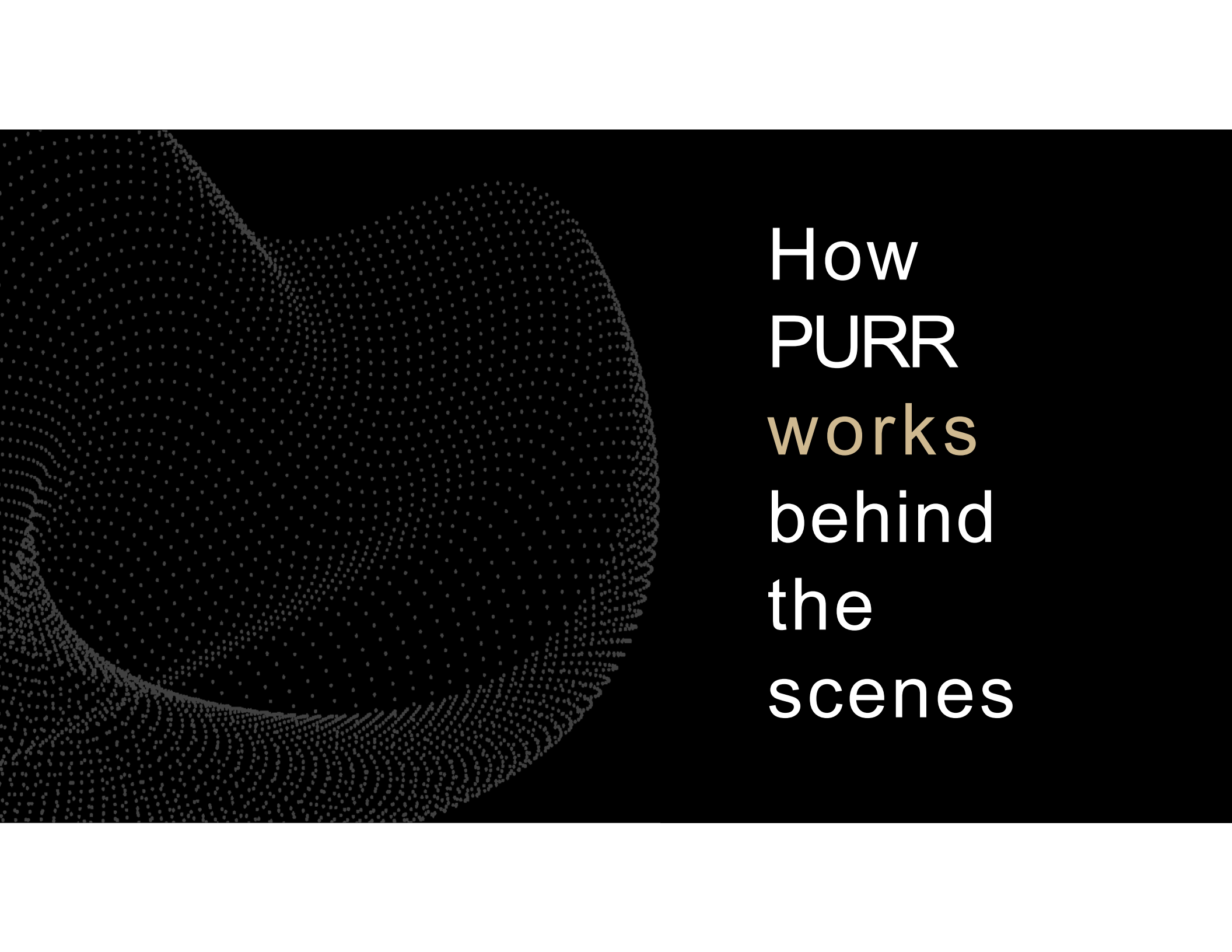
✓ Archived on 25 Oct 2016

Licensed under [Attribution 3.0 Unported](#)

“ 101 citation(s)

42501 total view(s), 169833 download(s)

→ Share: [f](#) [t](#) [R](#) ...



How PURR works behind the scenes

Daily Operations

LIBRARIES RESEARCH DATA

- data curation
- user support
- development and maintenance
custom functionality
- education and outreach

The PURR Team

PURR Staff



Sandi Caldron
Data Repository
Outreach Specialist
Purdue University
Libraries



Jerry Kuang
Digital Library
Software Developer
Purdue University
Libraries



Standa Pejša
Data Curator
Purdue University
Libraries



Michael Witt
PURR Director
Associate Professor
Purdue University
Libraries

Tech Partners

supporting PURR staff as we support users

HUBzero

open source platform for
science gateways providing
core functionality, technical
support

Campus and Libraries IT

project file storage
infrastructure, and some
hosting for custom webpages

MetaArchive

digital preservation
cooperative using the LOCKSS
platform

Collaboration with the research office

Interaction and information sharing between PURR and the research office are built into both automatic processes and human-driven workflows

Proposal Writing

- PURR provides DMP training, boilerplate language, and templates
-

Grant Award

- Research office alerts PURR when a grant that mentioned PURR in the DMP is approved
 - PURR and liaison librarian reach out to PI
-

Project Creation

- PURR project creation process has guardrails for sensitive data
- PURR sends automatic alerts to research office
- Grant funded projects receive more storage
- Research office verifies grant info triggering the storage upgrade

How do our **partnerships** add up?

PURR BY THE NUMBERS

457

DATA MGMT
PLANS

676

GRANT AWARDS

4,610

REGISTERED
RESEARCHERS

1,241

PUBLISHED
DATASETS



Do you have
any questions?

purr.purdue.edu
purr@purdue.edu

Q & A

**Next Session: Discussion Hour: Costing Hot Topics Discussion
@ 5 pm EDT**

Visit us at www.cogr.edu

Thank You