LADDER: Data Management at NIAMS

Scientific Information Technology Branch

Christine Winchester — Intramural Research Project Liaison

CIVIC DIGITAL FELLOWSHIP

NIH

RALPH BUAN
Santa Monica College
Interaction Design

MARS IKEDA
Smith College
Statistical & Data Sciences

LADDER

- **L**aboratories
- **A**dvancing
- **D**ependable
- **D**ata
- **E**nvironments &
- **R**euse

CIVIC DIGITAL FELLOWSHIP

DATA MANAGEMENT MATTERS

- **Modernizing infrastructure** at NIAMS requires knowing how much data you have now so you can estimate what you'll need in the future

- Scientific instruments at NIAMS are replaced every few years. **Newer instruments create larger data files**

- NIAMS/NIH attracts researchers from all over the world. High turnover can lead to more **orphaned data (data created by people who are no longer at NIAMS)**

CIVIC DIGITAL FELLOWSHIP

NIH & DATA POLICY

- January 25, **2023: Final NIH Policy for Data Management & Sharing:** provides institutional support for Open Science
- **Open Science** is a new approach to the scientific research based on:
  - **cooperative** work
  - **diffusing knowledge** using digital technologies and collaborative tools*

**from FosterOpenScience.uk*

CIVIC DIGITAL FELLOWSHIP

OPEN SCIENCE & DATA

- Open Science requires **researchers to create and share high quality data** so it can be found and reused by other researchers
- **Data that is high quality** means:
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable
- You can't make high quality, FAIR data without strong data management at the institute and researcher levels

CIVIC DIGITAL FELLOWSHIP

OPEN SCIENCE BENEFITS

- **Efficiency**
  - Reduce duplication & cost
  - More research from the same data
- **Quality & Integrity**
  - Wider evaluation & scrutiny of data & ideas
  - Early identification of scientific malpractice
- **Innovation**
  - Less delay in reuse of results of research (e.g., articles, data sets) could mean a shorter window between research and new products & services

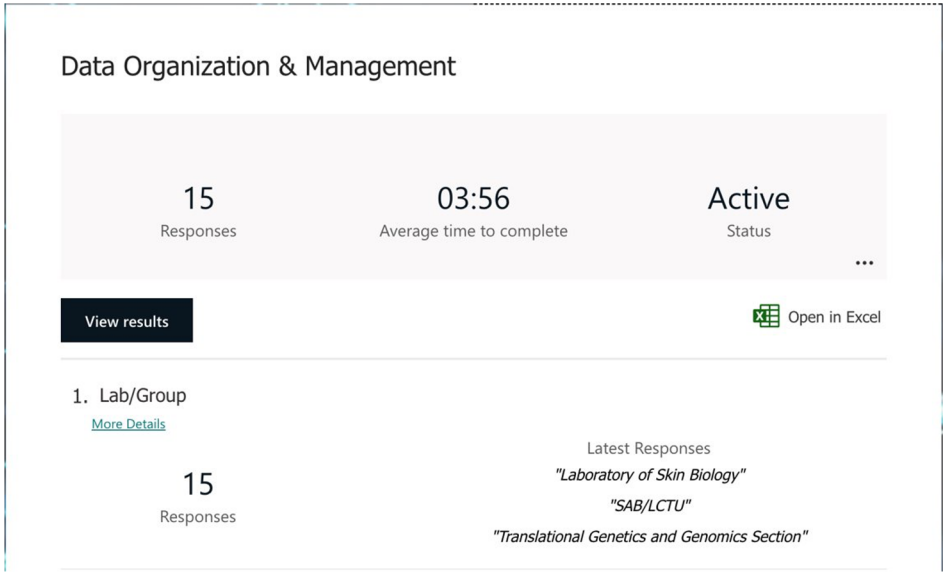
CIVIC DIGITAL FELLOWSHIP

LADDER ACTIONS

- **Review** previous fellows' findings and next steps
- **Research** data management practices available via NIH resources
- **Analyze** existing data storage utilization at NIAMS
- **Understand** partners' (researchers & technologists) needs. Survey took <5 mins., interviews up to 1 hour
- **Develop & share** solutions

CIVIC DIGITAL FELLOWSHIP

Data Management Survey



- 15 total responses
- Behaviors, pain points and roadblocks, and awareness of data management practices
- Preparation for stakeholder interviews

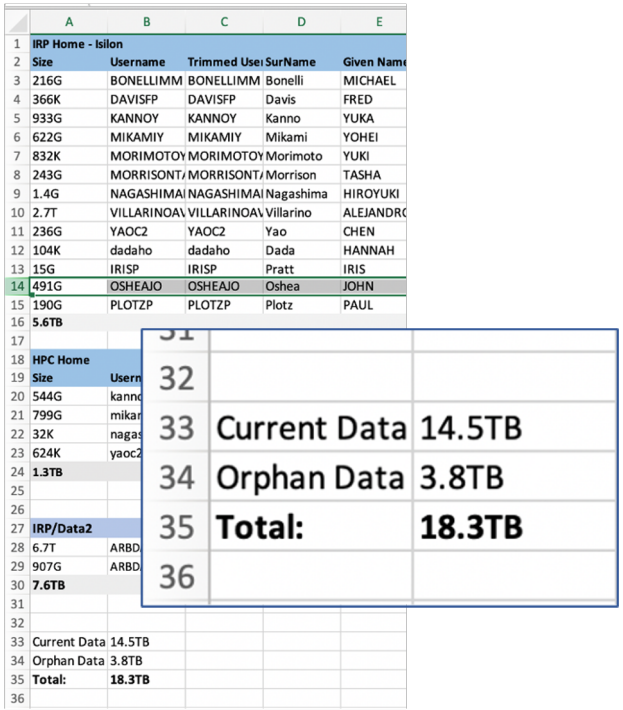
CIVIC DIGITAL FELLOWSHIP

Stakeholder Interviews

- Talked to **Principal Investigators** and **Core Facilities**
- Follow up on survey answers
- Understand current challenges around data management
- Validate potential solutions and practices

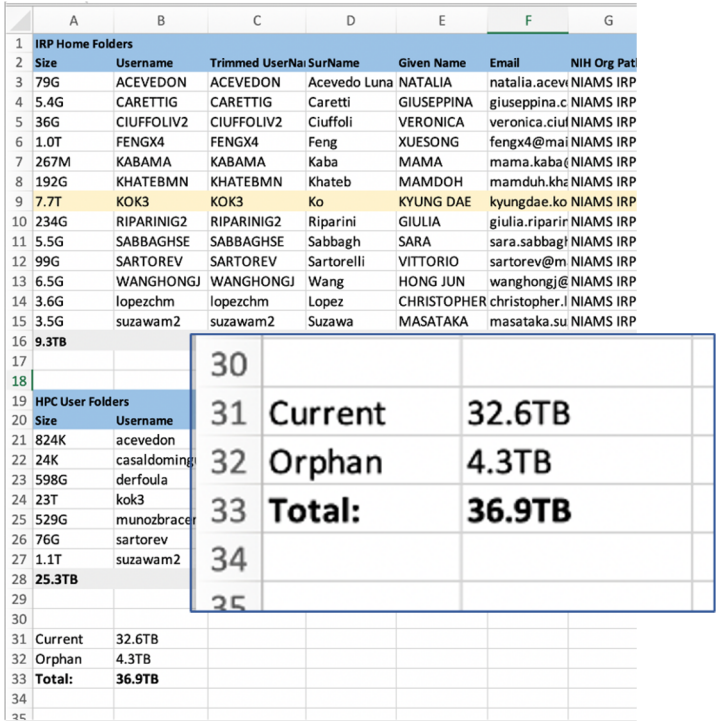
CIVIC DIGITAL FELLOWSHIP

Data Usage Reports



	A	B	C	D	E
1	IRP Home - Isilon				
2	Size	Username	Trimmed User	SurName	Given Name
3	216G	BONELLIMM	BONELLIMM	Bonelli	MICHAEL
4	366K	DAVISFP	DAVISFP	Davis	FRED
5	933G	KANNOY	KANNOY	Kanno	YUKA
6	622G	MIKAMIY	MIKAMIY	Mikami	YOHEI
7	832K	MORIMOTOY	MORIMOTOY	Morimoto	YUKI
8	243G	MORRISONT	MORRISONT	Morrison	TASHA
9	1.4G	NAGASHIMA	NAGASHIMA	Nagashima	HIROYUKI
10	2.7T	VILLARINOA	VILLARINOA	Villarino	ALEJANDRO
11	236G	YAOC2	YAOC2	Yao	CHEN
12	104K	dadaho	dadaho	Dada	HANNAH
13	15G	IRISP	IRISP	Pratt	IRIS
14	491G	OSHEAJO	OSHEAJO	Oshea	JOHN
15	190G	PLOTZP	PLOTZP	Plotz	PAUL
16	**5.6TB**				
17					
18	HPC Home				
19	Size	Usern			
20	544G	kanno			
21	799G	mikar			
22	32K	nagas			
23	624K	yaoc2			
24	**1.3TB**				
25					
26					
27	IRP/Data2				
28	6.7T	ARBD			
29	907G	ARBD			
30	**7.6TB**				
31					
32					
33	Current Data	14.5TB			
34	Orphan Data	3.8TB			
35	**Total:**	**18.3TB**			
36					

32		
33	Current Data	14.5TB
34	Orphan Data	3.8TB
35	**Total:**	**18.3TB**
36		



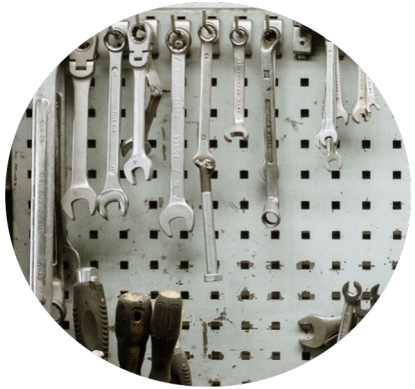
	A	B	C	D	E	F	G
1	IRP Home Folders						
2	Size	Username	Trimmed UserNa	SurName	Given Name	Email	NIH Org Pat
3	79G	ACEVEDON	ACEVEDON	Acevedo Luna	NATALIA	natalia.aceve	NIAMS IRP
4	5.4G	CARETTIG	CARETTIG	Caretti	GIUSEPPINA	giuseppina.c	NIAMS IRP
5	36G	CIUFFOLIV2	CIUFFOLIV2	Ciuffoli	VERONICA	veronica.ciu	NIAMS IRP
6	1.0T	FENGX4	FENGX4	Feng	XUESONG	fengx4@mai	NIAMS IRP
7	267M	KABAMA	KABAMA	Kaba	MAMA	mama.kaba@	NIAMS IRP
8	192G	KHATEBMN	KHATEBMN	Khateb	MAMDOH	mamduh.kha	NIAMS IRP
9	7.7T	KOK3	KOK3	Ko	KYUNG DAE	kyungdae.ko	NIAMS IRP
10	234G	RIPARINIG2	RIPARINIG2	Riparini	GIULIA	giulia.riparir	NIAMS IRP
11	5.5G	SABBAGHSE	SABBAGHSE	Sabbagh	SARA	sara.sabbagh	NIAMS IRP
12	99G	SARTOREV	SARTOREV	Sartorelli	VITTORIO	sartorev@m	NIAMS IRP
13	6.5G	WANGHONGJ	WANGHONGJ	Wang	HONG JUN	wanghongj@	NIAMS IRP
14	3.6G	lopezchm	lopezchm	Lopez	CHRISTOPHER	christopher.l	NIAMS IRP
15	3.5G	suzawam2	suzawam2	Suzawa	MASATAKA	masataka.su	NIAMS IRP
16	**9.3TB**						
17							
18							
19	HPC User Folders						
20	Size	Username					
21	824K	acevedon					
22	24K	casaldoming					
23	598G	derfoula					
24	23T	kok3					
25	529G	munozbracer					
26	76G	sartorev					
27	1.1T	suzawam2					
28	**25.3TB**						
29							
30							
31	Current	32.6TB					
32	Orphan	4.3TB					
33	**Total:**	**36.9TB**					
34							
35							

30		
31	Current	32.6TB
32	Orphan	4.3TB
33	**Total:**	**36.9TB**
34		
35		

CIVIC DIGITAL FELLOWSHIP

Insights & Solutions

Clear the Clutter Campaign



Toolbox

CORES Data Portal

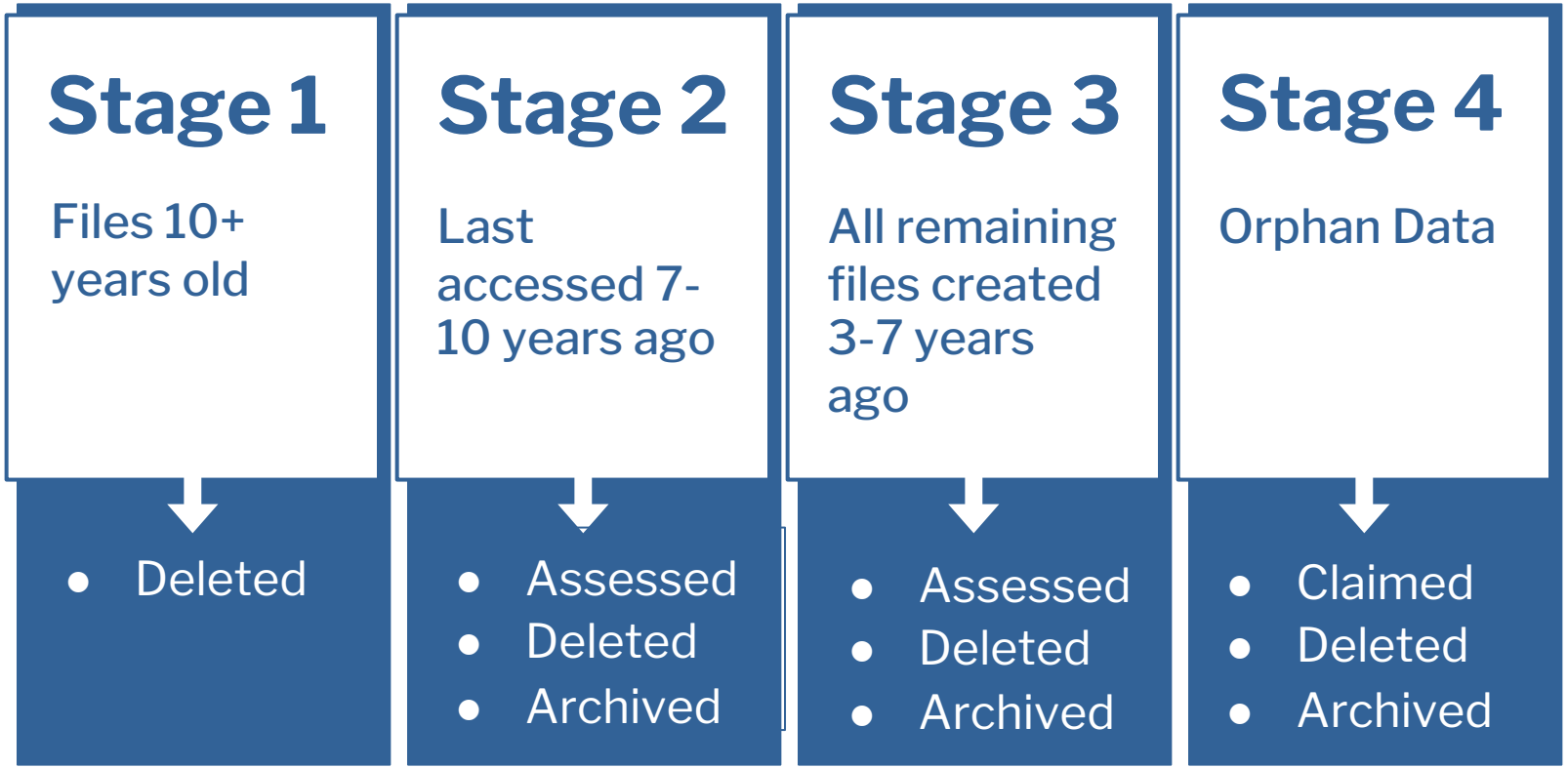
CIVIC DIGITAL FELLOWSHIP

Clear the Clutter Campaign: Insights

Researchers' Blockers	IT's Solutions
Lack of resources: Staff & Time	Divide work into incremental steps so work can be done in stages
Apprehension: Taking the first step	IT deletes data declared of no value at beginning of project
Apprehension: "Making things worse"	Logical structure from easiest to manage data (e.g., oldest) to most difficult (e.g., most recent)

CIVIC DIGITAL FELLOWSHIP

Clear the Clutter Campaign



CIVIC DIGITAL FELLOWSHIP

LADDER Toolbox

Exercises and Workshops

Data Flow Diagrams

Best Practices and Resources

CIVIC DIGITAL FELLOWSHIP

LADDER TOOLBOX

Data Management Plans

Effective January 25, 2023, all research funded or conducted by the NIH that result in the generation of scientific data will require the submission of a data management and sharing plan. It should outline how scientific data will be managed and shared throughout the lifecycle of a research project.

The NIH Final Policy for Data Management and Sharing emphasizes the importance of good data management practices, promotes accountability and transparency, and establishes the expectation for sharing scientific data from NIH-funded or conducted research.

The NIH recommends addressing all of the elements below in your plan within 2 pages or less.

Data type

- ☐ Describe the types and estimated amount/size of scientific data to be generated and/or used in the study.
- ☐ Decide and indicate which scientific data will be preserved and shared based on legal, ethical and technical factors that may affect the extent to which scientific data will be preserved and shared.
- ☐ List metadata, other relevant data, and any associated documentation.

Tools, software, or code

- ☐ Indicate which specialized tools are needed to access or manipulate shared scientific data to support replication or reuse.
- ☐ Specify how needed tools can be accessed and, if known, whether such tools are likely to be available for as long as the scientific data remain available.

Standards

- ☐ Indicate what standards will be applied to the scientific data and associated metadata.

Preservation, access and timelines

- ☐ List the name(s) of the repository(ies) where scientific data and metadata arising from the project will be archived.
- ☐ Specify how the scientific data will be findable and identifiable.
- ☐ Identify any differences in timelines for different subsets of scientific data to be shared.

Access, distribution, and reuse considerations

- ☐ Describe any applicable factors affecting access, distribution, or reuse of scientific data related to: informed consent, privacy and confidentiality protections, whether access to scientific data derived from humans will be controlled, and any legal restrictions.

Oversight of data management or sharing

- ☐ Indicate how compliance with the plan will be monitored and managed, frequency of oversight, and by whom.

File Naming Conventions

Establish a naming convention for your research files that is short, consistent, and descriptive to help you stay organized and identify your files when you need them. Do this before you start generating or collecting files and write it down so others in your lab can follow this standard.

Best Practices

- Avoid special characters or spaces
- Use capitals and underscores instead of periods, spaces, or slashes
- Use versioning
- Write down naming convention for you and your team
- Put the most important information at the beginning of the file name
- For date format, use YYYYMMDD or YYMMDD

Elements to consider using for your naming convention

- Date of creation or data range of experiment
- Location/lab name
- Project name or acronym
- Sample ID or experiment number
- Version number
- Researcher name or initials
- File Type
- Conditions
- Instrumentation

Steps for making your naming convention:

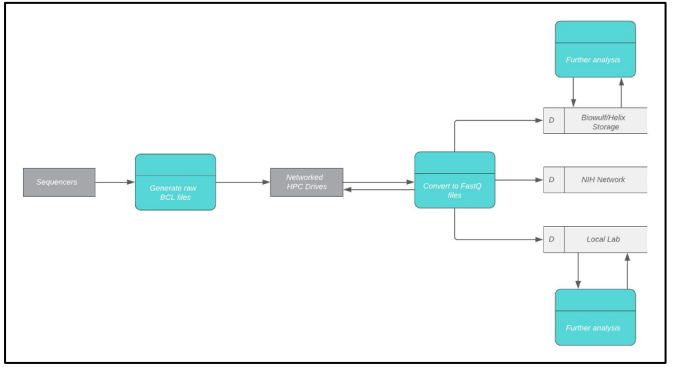
1. Choose a set of files that you'll be making this naming convention for. This could be for an experiment, a particular file type, an instrument(s), etc.
2. Pick 3-5 elements from the above list to include in your convention. Choose elements that will help you differentiate and identify these files.
3. Arrange these elements in a way that makes sense to you. Start with the most important element first.
4. Write it down so that everyone in your team can follow this standard.

Sample naming convention:

Date_ProjectName_FileType_Initials.ext

File name using the above convention:

20210719_DNA_InterviewNotes_RB.docx



FAIR Guiding Principles

The FAIR Data Principles highlight the need to embrace good practice by defining essential characteristics of data objects to ensure that data are reusable by humans and machines.

Findable

Metadata and data should be easy to find by both humans and computers.

Accessible

Once found, users need to know how the data can be accessed.

Interoperable

Metadata use an accessible and standardized language so it can be read and used by a broad range of applications.

Reusable

Data is well-described and provides clear usage of licenses.

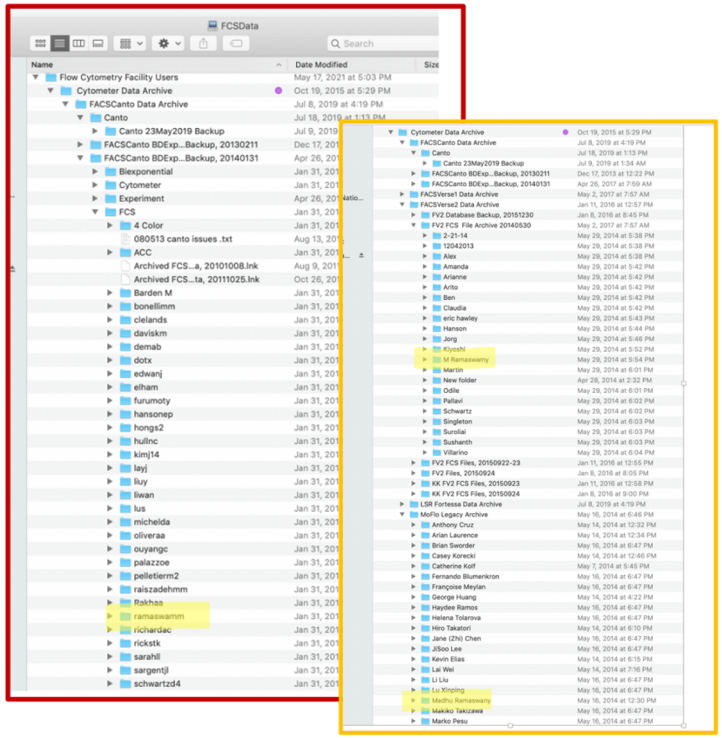
CIVIC DIGITAL FELLOWSHIP

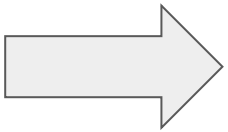
CORES Data Portal: Insights

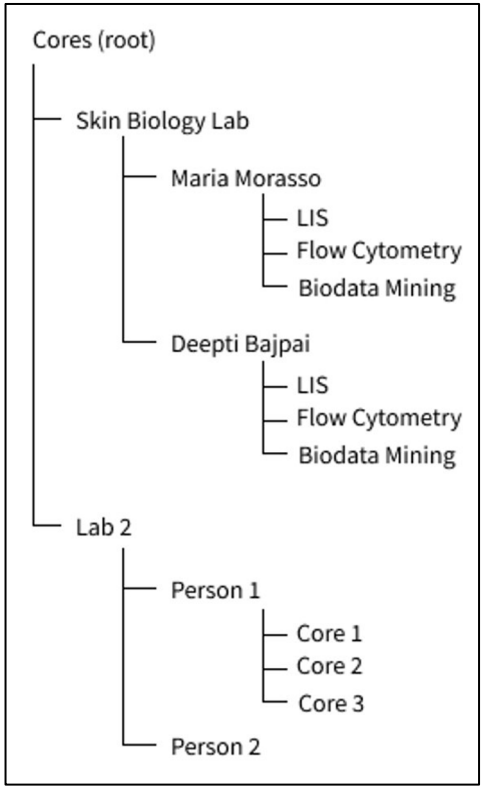
Researchers' & Technologists' Blockers	IT's Solutions
Don't know who owns data	Organize folders by owner
Processes are hectic & patchwork	All CORES handle data the same way
CORES have low staff	Fewer places to look for data will put less stress on current staff of CORES & researchers

CIVIC DIGITAL FELLOWSHIP

CORES Data Portal







CIVIC DIGITAL FELLOWSHIP

Next Steps

- Start the **Clear the Clutter Campaign**
- Make the **Toolbox a living document & resource,** perhaps stored on Github
- **Build out CORES Data Portal** with appropriate security permissions
- As the 2023 Data Management policy nears, **educate researchers on importance of version control & Open Science**
- Consider forming a Data Management advisory board to help keep lines of communication open

CIVIC DIGITAL FELLOWSHIP

Thank You!

- **Christine Winchester, La'Tanya Burton, Diana Mungai**, and SITB
- **Rachel Dodell** and **Ariana Soto** from Coding it Forward
- **Jacqueline Cattell**, **Erin Walker**, and **Allissa Dillman**
- **Martyn Green** and **Christine Cutillo**
- NIAMS **Intramural Research Program** PIs and Core Facilities

CIVIC DIGITAL FELLOWSHIP

Questions