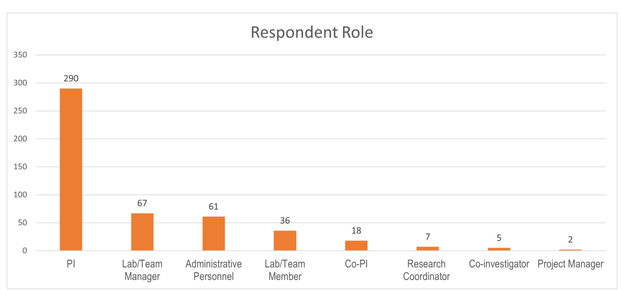
Copy and Pasting vs. Customization: Using Qualitative Analysis of NIH Grantees' Data Management Plans to Shape Future DMP Support

LEVI DOLAN, RUTH LILLY MEDICAL LIBRARY, INDIANA UNIVERSITY SCHOOL OF MEDICINE
ELIZABETH C. WHIPPLE, RUTH LILLY MEDICAL LIBRARY, INDIANA UNIVERSITY SCHOOL OF MEDICINE
HEATHER L. COATES, UNIVERSITY LIBRARY, INDIANA UNIVERSITY-PURDUE UNIVERSITY INDIANAPOLIS, ORCID: 0000-0003-4290-6997

INTRODUCTION

In 2019, our school began requiring all of its NIH grantees to submit data management plans designed to prepare researchers for future requirements, such as the NIH Final Policy on Data Management and Sharing. In 2020-2021, the second year of this mandate, we achieved near 100% compliance with 600+ project-level DMPs submitted. Principal Investigators (PIs) were by far most frequently the team member submitting a DMP:



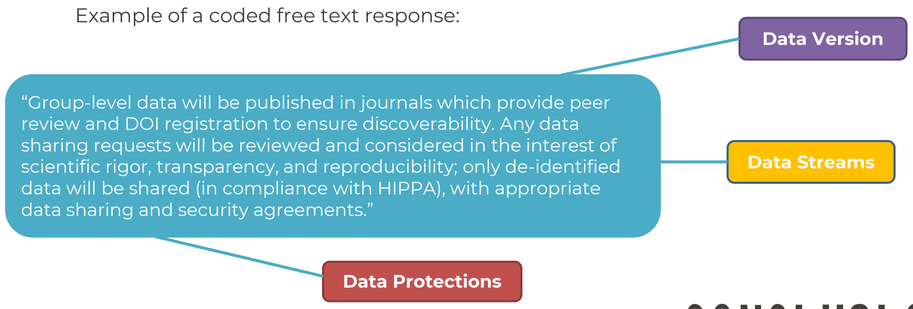
OBJECTIVE

Our objective was to identify themes in the free text responses in the submitted DMPs. The DMP is collected via FireForm, an in-house enterprise-level electronic form and workflow tool. These six free text response fields describe practices related to data collection procedures, QA/QC procedures, data retention periods, data sharing, and permissions to take copies of data when team members leave. Additionally, respondents were asked to identify relevant funder data management or sharing requirements, citations for data reuse, and applicable standards for data and metadata to gauge awareness of these topics.

METHODOLOGY

A codebook was iteratively created based on a randomly selected subset of DMP responses. After developing the codebook with input and discussion from all the authors, two of the authors independently coded the free text responses in the first 300 DMP submissions received in 2021 (out of 702 total submissions), with the third author then reconciling disagreements. Themes that emerged from the codes were then analyzed.

Example of a coded free text response:



RESULTS

The most common themes centered around describing data streams and the conditions and infrastructure used to manage the data during the progress of the research period.

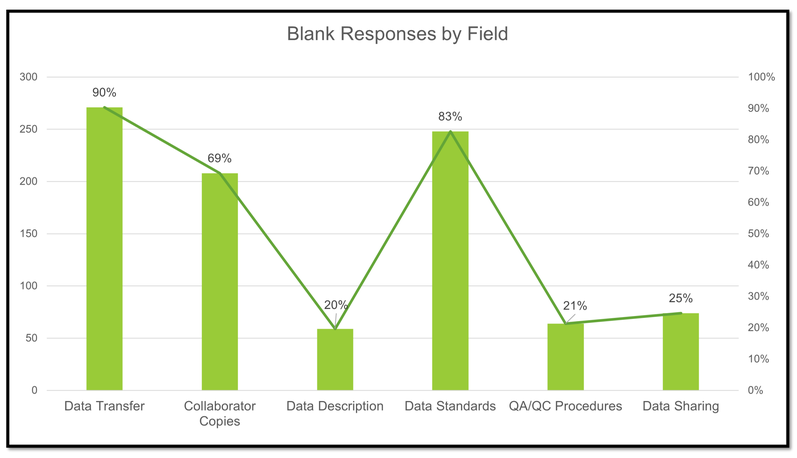
The most common null responses (blank fields) concerned definitions of the scope of continuing access for former project personnel, and identification of data standards relevant to the area of research.

ANALYSIS

Codebook creation proceeded through manual analysis of free text responses. Each author proposed a taxonomy and these were merged to create the following set of qualitative codes:

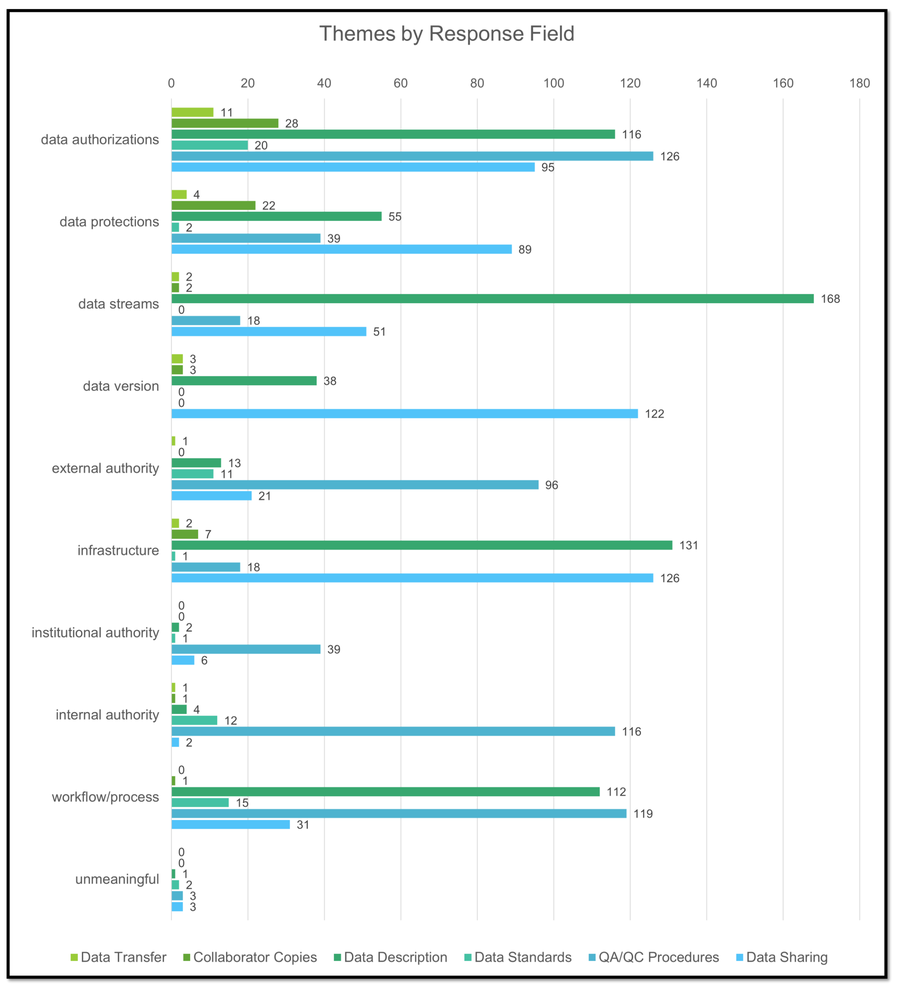
	Code	Definition
	Data Authorizations	Definitions of how data is allowed to be used. Including: • Personnel: Stating which members of the team have access according to their role in the project. • Conditions: Explanations of how conditions are met for data sharing and transfer, and the types of circumstances that authorize data to be disseminated.
	Data Protections	How data may be used within defined limits. Including: • Security: Safeguards which will be put in place within the infrastructure, such as passwords, encryption, or restrictions to physical access to data storage. • Legal protections: References to standards such as HIPPA, IRB, or confidentiality. • Agreements: Mentions of implicit or explicit agreements related to controlled access sharing with a group outside of the research team.
	Data Streams	Definitions of the extent of the data under discussion and indicators of its contents. Including: • Format: Descriptions of data containers through file types, media types, or software tools used. • Content: Parameters which describe the kind of data that has been collected, such as clinical, laboratory, or experimental data. • Subsets: Definition of a particular subset of data, distinct from the data generated by the research project as a whole.
	Data Version	References to the version of the data as defined by: • Timeframe: points within the research process where a version of the data will be shared or restricted. • Preservation: Definition of how a persistent version of the data will be preserved, as in a repository or part of a scholarly publication.
	External Authority	Citation of an external policy, standard, or methodology for quality assurance and/or data standard definition (including hyperlinks).
	Infrastructure	References to any computational infrastructure to be used in data collection, analysis, and preservation such as software tools and platforms.
	Institutional Authority	Citation of an institutional policy, standard, or methodology for quality assurance and/or data standard definition.
	Internal Authority	Citation of an internal policy, standard, or methodology for quality assurance and/or data standard definition.
	Workflow/Process	Descriptions of the processes, workflow, and methodology to be used in data management, or statements that list these things.
	Unmeaningful	Cases where a response does not include enough information to assign any of the other categories meaningfully.

Qualitative analysis was limited to responses which contained free text. To understand areas where the current form is not being well-utilized, the blank responses were summarized:



	Former Collaborator Copies	Continuing Collaborator Copies	Data Description	Data Standards	QA/QC Procedures	Data Sharing
data authorizations	11 (4%)	28 (9%)	116 (39%)	20 (7%)	126 (42%)	95 (32%)
data protections	4 (1%)	22 (7%)	55 (18%)	2 (<1%)	39 (13%)	89 (30%)
data streams	2 (<1%)	2 (<1%)	168 (56%)	0	18 (6%)	51 (17%)
data version	3 (1%)	3 (1%)	38 (13%)	0	0	122 (41%)
external authority	1 (<1%)	0	13 (4%)	11 (4%)	96 (32%)	21 (7%)
infrastructure	2 (<1%)	7 (2%)	131 (44%)	1 (<1%)	18 (6%)	126 (42%)
institutional authority	0	0	2 (<1%)	1 (<1%)	39 (13%)	6 (2%)
internal authority	1 (<1%)	1 (<1%)	4 (1%)	12 (4%)	116 (39%)	2 (<1%)
workflow/process	0	1 (<1%)	112 (37%)	15 (5%)	119 (40%)	31 (10%)
unmeaningful	0	0	1 (<1%)	2 (<1%)	3 (1%)	3 (1%)
blank	271 (90%)	208 (69%)	59 (20%)	248 (83%)	64 (21%)	74 (25%)

The places where each of the coded themes appeared as percentages of the free text fields available in the FireForm are summarized in the table above. Below, the total number of occurrences in each field are depicted as a bar graph.



CONCLUSION

The sample of DMP submissions examined gave evidence that PIs and other DMP authors are more likely to describe the scope of their data and envision general terms of sharing the data than they are likely to provide for contingencies surrounding data access and collaboration during the project period. QA/QC procedures were more likely to be described as elements of an internal lab process rather than looking to domain-specific or organizational data standards. This provides evidence that NIH-funded researchers still have many questions about how data management should best be expressed and monitored in the context of applying general statements of compliance to data sharing and preservation to the circumstances of each research project.

Thematically, streams of scientific data and the infrastructure needed to process and preserve it are areas where the current DMP FireForm is best capturing necessary data elements. There are fewer specifics being written about how data will be maintained, controlled, and shared; more often the language of the DMPs promises that these things are going to happen without describing processes. Next steps will involve extracting generalized sample responses that address customized management and sharing strategies to the specific needs of a given project type. This will involve conversations about how to move from simply stating that compliance will occur to mapping strategies for reproducibility onto the research process.