Metrics for Data Repositories and Knowledgebases: Working Group Report

Contributors: See Appendix 3

Executive Summary

This report presents the findings of an exploration of the current landscape of biomedical data repository metrics conducted by the NIH Lifecycle and Metrics Working Group and the NIH Metrics for Repositories (MetRe) Working Group. The working groups (WGs) gathered input from the biomedical data repository community using a two-part process. In the first phase, the MetRe WG identified several NIH-funded repositories and developed a list of metrics that are most commonly collected. This list of metrics was used to inform the development of a survey instrument that the Lifecycle and Metrics WG used to collect input from the broader biomedical repository community in Phase 2. This report describes the findings of these two activities, providing insights into the current state of data and repository metrics in the biomedical research community and complementing efforts underway in the broader metrics community.

This report includes input from representatives of 13 NIH repositories from Phase 1 and 92 repository managers in Phase 2. The metrics these respondents reported using are divided into several broad categories, including (from most to least commonly collected) User Behavior Characteristics, Scientific Contribution/Impact, and Repository Operations, and the respondents from the two groups reported similar patterns in the metrics they collect. The majority of respondents in Phase 2 (77%) also indicated a willingness to share their metrics data – an encouraging finding given that such metrics can be helpful to NIH in better understanding how datasets and repositories are used. Many respondents from both groups reported that they were using Google Analytics to collect metrics, primarily in the areas of User Behavior Characteristics, given its ease of use and ability to accurately track such metrics. However, many respondents also indicated that they would like to collect additional metrics but currently do not or cannot because of lack of tools for doing so.

The findings of this report provide a better understanding of the metrics currently used within the biomedical repository community, which can inform future NIH efforts to help develop this space and to understand patterns of use across datasets and repositories. NIH should also maintain awareness of developments in the broader repository metrics community to ensure alignment.

Introduction

Data repositories and knowledgebases are essential to increasing the information value of the scientific research enterprise and have served as important component of the data ecosystems for preserving, archiving, and disseminating of scientific data. As the size and diversity of data collected and stored from biomedical research continues to increase and we transition towards a modernized data ecosystem, the need for making these research data and information FAIR (Findable, Accessible, Interoperable and

Reusable) [1], and the important role of repositories in bringing this to fruition is even more evident. Repositories serve not only as data storage and archival systems for aiding reproducibility of research, but they also help assess the impact of research data. Moreover, repositories are increasingly required to be trustworthy systems to maintain the scientific value of data. The recently formulated TRUST (Transparency, Responsibility, User focus, Sustainability, Technology) principles [2] provide a framework for formalizing the capabilities of a repository to efficiently serve its community. The biomedical research enterprise that the National Institutes of Health (NIH) funds and supports generates large amounts of data that are stored and accessible for public use within these repositories; however, challenges remain not only in evaluating and assessing the value and impact of individual repositories, but also in developing models for long-term sustainability of these resources.

In the first NIH Strategic Plan for Data Science [3], one of the overarching goals is the modernization of the data ecosystem, with the plan also providing a pathway for implementation. The NIH Office of Data Science and Strategy (ODSS) has been providing NIH-wide leadership, in conjunction with NIH working groups, to implement the goals of the strategic plan. The goals of the Data Science Strategic Plan Implementation Tactics Subgroup include:

- Implementation Tactic 2.1.3: Dynamically measure data use, utility and modification
- Implementation Tactic 2.1.6: Employ explicit evaluation, lifecycle, sustainability and sunsetting expectations (where appropriate) for data resources

The Lifecycle and Metrics working group focuses on these two tactics and established a sub-group, Metrics for Repositories (MetRe), based on input from the community provided at the February 2020 Virtual Workshop on Data Metrics [4] to address implementation of Tactic 2.1.3. The MetRe sub-group was formed to learn how dynamically measured metrics can be applied to demonstrate the usage, value and benefits of repositories and knowledgebases. The first step in accomplishing this goal was to understand how a subset of NIH-managed repositories and stakeholders used metrics in operations and procedures, to assess scientific impact of the data resource, and understand the varying actions and considerations that go along with the process. The follow-up step included the solicitation of community input with respect to usage and cost metrics of repositories via a survey instrument, based on lessons learned from the preliminary analyses of previous step. While the MetRe Working Group (WG) recognizes the importance of sustainability in the data lifecycle continuum, it is not in scope for these activities.

Background

Publicly funded data repositories often serve as core resources that can be utilized for archiving and curating data, preserving analysis workflows, and making research datasets accessible to the broader scientific community. The continued operation and success of a repository relies not only on the quality and accessibility of the data stored within it, but also on the broader scientific impact of the use of the data. Repository managers have an interest in demonstrating the impact of their repositories for past, present, and future research endeavors. This impact could be quantified based on different perspectives or characteristics. Metrics provide systematic parameters for evaluating the cost and benefits of a repository to the various repository stakeholders, including research institutions, funding agencies, and the research communities at large.

A number of efforts have been undertaken by the broader scientific community to establish repository metrics and best practices for quantifying for quantifying both repository and data use, value, and impactTo inform this report, a review of studies assessing repository metrics over the last several years was done with a view of understanding the variety, utility and application of metrics being gathered by repositories from various scientific fields. These studies [5-9], published over the last decade, have explored a variety of metrics that can be used in repository performance assessments as well as challenges inherent to the collection of these metrics. The information provided by these metrics give more insights into repository performance by tracking repository access and usage, interoperability, scientific contribution or impact, and the costs associated with repository operations. Additionally, various international certification standard bodies such as CoreTrustSeal [10], DIN31644/NESTOR [11], and ISO16363 [12] certify repositories with a primary focus on operational aspects of a repository, with little attention on usage and scientific impact metrics.

From the perspective of a funding agency such as NIH, metrics can be used to measure scholarly output and impact, meet data access and sharing requirements, and increase visibility and impact of the research area being funded. To better understand the metrics that repositories currently collect and their practices for using and understanding them, the MetRe WG explored the landscape of repositories funded and managed by NIH as well as the biomedical repository community. Using a combination of discussions with repository managers and a survey to repository stakeholders, the WGs gathered information on metrics that are currently used to measure repository and data usage and impact, as well as gained insight into the questions that stakeholders would like to be able to answer using metrics. The metrics described in this report are not necessarily the most desirable metrics, nor a comprehensive solution to answering relevant questions about data and repository use and utility, but this report provides a starting point for understanding the current landscape of metrics within the NIH repository ecosystem.

Approach

The activities described here were conducted in two phases, as shown below in **Figure 1**.

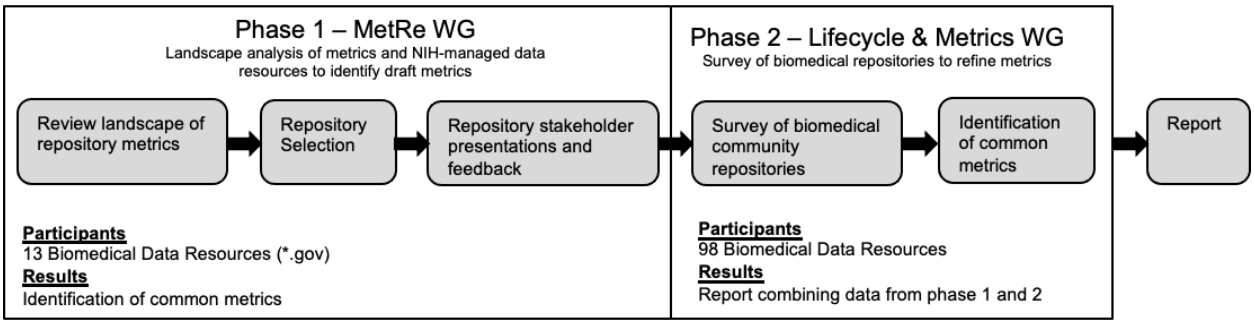

Figure 1. An overview of the approaches included in this report

Phase 1 activities

Phase 1 focused on landscape analysis of metrics and information gathering from NIH-managed data resources (with a '.gov' website) and was conducted from March 2019 to August 2019. Phase 1 activities included 1) review of current metrics; 2) selection of repositories for inclusion in this review ; 3)

repository manager/stakeholder presentations which led to identification of an initial set of metrics used in common among these repositories (**Figure 1**). The repository selection was done first and independent of the other activities, while the information collected from the metrics review and stakeholder presentations were used in an iterative process to arrive at a list of common metrics in Phase 1 and to develop potential questions that were used in the larger repository stakeholder survey (Phase 2).

a. *Repository selection*: A review of the landscape of NIH-funded and -managed repositories across the different NIH Institutes and Centers (ICs) was done to generate a list of potential candidates for further consideration of in-depth assessment. This original list (N = 132) was streamlined to generate a diverse list of 13 repositories (see **Table 1**) that are as representative as possible of research done across NIH ICs. The inclusion criteria in making the final selection included:

1. existence of a .gov website
2. types of study data (including Omics, clinical study, literature, image, nanomaterials, and audio-visual recording/media),
3. data access (e.g., controlled access, registration required, open-access, and mixed access), and
4. function of data resource (e.g., data repository or knowledgebase).

A selection decision was made to focus on data repositories, with the goal for this list to include a variety of repositories managed by NIH across several ICs. Repositories focusing only on biospecimens without associated data were excluded from the selection.

Table 1: Phase 1 Assessed Repositories

Repository Name	NIH Institute or Center	Access Type	Data Type
Chemical Effects in Biological Systems (CEBS)	NIEHS	Open Access	Multiple
ClinicalTrials.gov	NLM	Open Access	Clinical
Data and Specimen Hub (DASH)	NICHD	Controlled	Clinical
Database of Genotypes and Phenotypes (dbGaP)	NLM	Controlled	Genomic
Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics System	Trans NIH and Govt	Controlled	Multiple
GenBank	NLM	Open Access	Genomic
NCI Genomics Data Commons	NCI	Mixed	Genomic
NEI Data Commons	NEI	Mixed	Multiple
Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC)	NHLBI	Controlled	Clinical
NIDDK Central Repository	NIDDK	Controlled	Multiple
NIMH Data Archive (NDA)	NIMH	Controlled	Multiple
The Cancer Imaging Archive (TCIA)	NCI	Open Access	Image
The Immunology Database and Analysis Portal (ImmPort)	NIAID	Registration required	Multiple

Definitions of the various categories of right of data use utilized in the study are as follows:

1. Controlled access: Application and eligibility requirements need to be met to gain access;
2. Registration required: Open to all, but users need to be signed in or registered with the resource to access;
3. Open access: No access restrictions or registration required to access; and
4. Mixed: Has both controlled and open access.

The NIH Office Data Science Strategy (ODSS) [13] defines data repositories as data resources that store, organize, validate, and make accessible the core data related to a particular system or systems. For example, core data might include genome, transcriptome, and protein sequences for one or more organisms. Knowledgebases are defined as resources that accumulate, organize, and link growing bodies of information related to core datasets. They are resources that may contain, for example, information about gene-expression patterns, splicing variants, localization, and protein-protein interactions and pathway networks related to an organism or set of organisms.

b. *Landscape Review of repository metrics*: To jumpstart the review of metrics used by repositories, two major resources were relied upon – 1) a list of metrics identified at the National Library of Medicine and 2) list of metrics gathered from a recent literature review [5-9]. Information from the two sources were combined to generate a master list of metrics. In the exploration of metrics that can best be used to assess repository usage, impact, and sustainability, three main questions were considered by the group:
   1. Which metrics are currently being collected by repositories and why?
   2. Of the metrics being collected, which are common to all participating repositories?
   3. What additional metrics, not already collected, would be valuable to track for the benefit of various repository stakeholders (repository owners/managers, repository users, and funding agencies)?

c. *Repository stakeholder presentations and feedback collection*: Over a three-month period, representatives from the repositories identified for inclusion in Phase 1 were invited to give presentations about their repositories. The content of these presentations included information such as type of data stored, metrics on access and utilization, scientific impact, and other metrics measured or tracked by the repository. Information about the rationale or purpose for metric collection as well as methods and tools utilized in collection or tracking of these metrics was also gathered.

d. *Identification of metrics in Phase 1* : Recorded responses from the repository representatives were mapped to the master list. The repository managers/representatives were asked to review and vet the list for their repositories to identify which metrics on the list are tracked, measured, or collected by their repositories. Expert knowledge from the working group in addition to feedback collected as described above was used in identification of the initial Phase 1 metrics. These Phase 1 metrics are defined as those found to be collected by the majority of the Phase 1 participating repositories, or those deemed to be important for assessing impact and sustainability of a repository by the working group members.

Phase 2 activity - NIH Biomedical Data Resource Community Survey

The Metrics Survey, based on potential questions and building on the metrics identified in Phase 1, was completed in Phase 2, and was deployed to the biomedical community from December 2020 to February 2021. No restrictions were placed regarding respondents to the survey.

Prior to designing the survey, the Phase 1 metrics were ranked based on the number of repositories gathering the metric (see **Table 2**). Among the metrics used by Phase 1 assessed repositories, those metrics most often tracked by the repositories or considered most relevant by the WG were selected to be used in the public survey. The respondents were not provided metric descriptions at the time of the survey. The survey questions are listed in **Appendix 2**.

Of the three major categories of metrics identified in Phase 1 (i.e., User Behavior, Scientific Contribution/Impact, and Repository Operations), most metrics assessing the impact of Scientific Contribution/Impact were not selected for the survey, with the exception of metrics tracking the number of projects/studies and number of subjects/cases, due to several considerations. The measurement of scientific contribution impact metrics among repositories are inconsistent, making it hard to compare across repositories. For example, bibliometrics is extremely complex and is currently being examined by other groups (Make Data Count [14,15], COUNTER [16], Scholix [17]). Importantly, the currently used scientific impact metrics are lagging indicators of the usage and utility of the repositories.

In addition to questions in multiple choice format, the survey included several open-ended questions to capture wider and deeper understanding of metrics used by repository stakeholders (**Appendix 2: Repository Survey Questions**). For example, the question "What metrics would you like to collect, but don't currently have the ability or infrastructure to collect?" was included to enable us to identify gaps in our proposed set of common metrics. The survey also collected information about the type of data resource on which the respondents' answers are based, including whether the resource was a knowledgebase, data repository, or hybrid, as well as whether the resource would be considered a generalist or domain-specific repository. The survey was designed using the Qualtrics Platform [18] and publicized through the NIH ODSS website.

Findings

Phase 1 Results

Answering the key questions: The list of repository metrics compiled by the WG served to address the three main questions of interest to the WG, as defined in the approach section. These metrics and their descriptions are presented in **Appendix 1**. The metrics could be further aggregated into the following categories: User Behavior characteristics (dealing with access and utilization e.g., number of users visiting the repository, number of downloads etc.), scientific contribution or impact, interoperability and harmonization (e.g., data quality and metadata), and repository operations and costs.

Metrics tracked by Phase 1 assessed repositories: The aggregate of metrics tracked by the Phase 1 assessed group of repositories is shown in **Appendix 1**. We found that most of the repositories tracked repository usage and user/visitor characteristics (e.g., geographical location, time spent on site, pages most visited while on the site, etc.), using automated and publicly available analytic tools. The

measurement tool most widely used for this purpose within the Phase 1 assessed repositories interviewed was Google Analytics. The repositories are able to capture other useful metrics at various levels of granularity using these analytic tools. The repositories also track some metrics that measure scientific contributions or impact of the repository (e.g., number of publications citing the resource in a year). Less commonly tracked by many of these repositories were metrics related to metadata completeness or data quality metrics and repository operations and cost metrics. The metadata completeness and data quality metrics are especially useful for assessing the findability, accessibility and interoperability of the data and access methods used by a repository, all of which can impact reusability of the data and harmonization of the datasets with other data ecosystems in the future. The metrics identified inPhase 1 (those collected by a majority of repositories or deemed by the working group members to be important for assessing impact) in this list are shown in **Table 2**. The column labeled 'Repository Tracking Count', shows the number of repositories reporting that they track a specific metric.

Table 2: Metrics Based on Phase 1 and Phase 2 analyses

Categories	Common *	Metrics	Description	Repository Tracking Count (Total 13)	Public Survey Tracking Count (Total 119)
User Behavior Characteristics	Y	Number of users	Number of users who can use the (visualization, e.g.) services for the data	13	90
	Y	Page views	Clicks, page scrolling, mouse movement/pointing	11	80
	Y	Downloads	Number of downloads or users downloading data, web or FTP	11	82
	Y	Geography	User IP address based - resolved to country/state	10	63
	Y	New vs. Returning Users	For a defined period, usually three months	10	49
	Y	Dataset submitters	Number of data submitters	9	56
	Y	Visit frequency	Daily, monthly, etc.	8	47
	C	Data Access Requests	How many data requests are made in a specified time period	7	5
Scientific Contribution/ Impact	C	Number of Projects/Studies	Number of Projects or Studies	10	59
		Number of Cases/Subjects	Total number of Cases or Subjects (e.g. individual human participant level data)	10	42
		Total publications	Total number of publications over all the years	8	8
Repository Operations	Y	Storage costs	Total storage cost for repository[#]	4	48
		Cost/dataset (Storage)	Cost per dataset (i.e. Storage)	2	8
		Hardware Costs	Total hardware costs	1	43
		Total download costs	Total download costs	1	8

*Y Indicates that this metric is always a common metric; C: Indicates that this metric is a common metric for some types of repositories (depending on the use restrictions of datasets stored within the resource.)

[#]Additional cost measures including staff costs were not included in this analysis

Phase 2 Results

We received 119 responses to the Metrics public survey of which 92 were from data repository managers or funders, our target audience (**Figure 2A**). Respondents are affiliated with 98 distinct data resources with the majority identified as a data repository (**Figure 2B**).

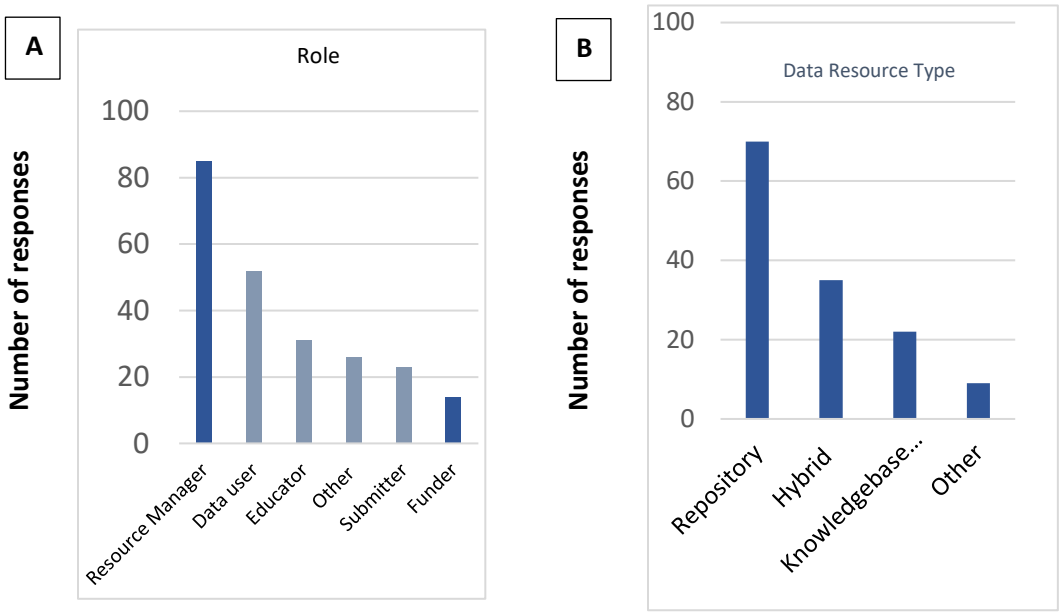

Figure 2. Number of respondents based on **A)** Role and **B)** Type of data resource

The types of metrics tracked were generally consistent across different types of data resources. There were no significant differences in the metrics tracked by knowledgebases versus data repositories as well as generalist versus domain-specific data repositories (**Data not shown**). Respondents also generally reported collecting similar metrics regardless of their role (**Figure 3**).

Comparison based on role of respondents:

Among the total respondents to the public survey, ninety-two identified themselves as repository funders and managers, which was our target audience. An analysis of metrics tracked based on the role of the respondent showed that results of all respondents had a trend similar to that for repository managers and funders combined together (please see **Figure 3**). For this report, the responses from the

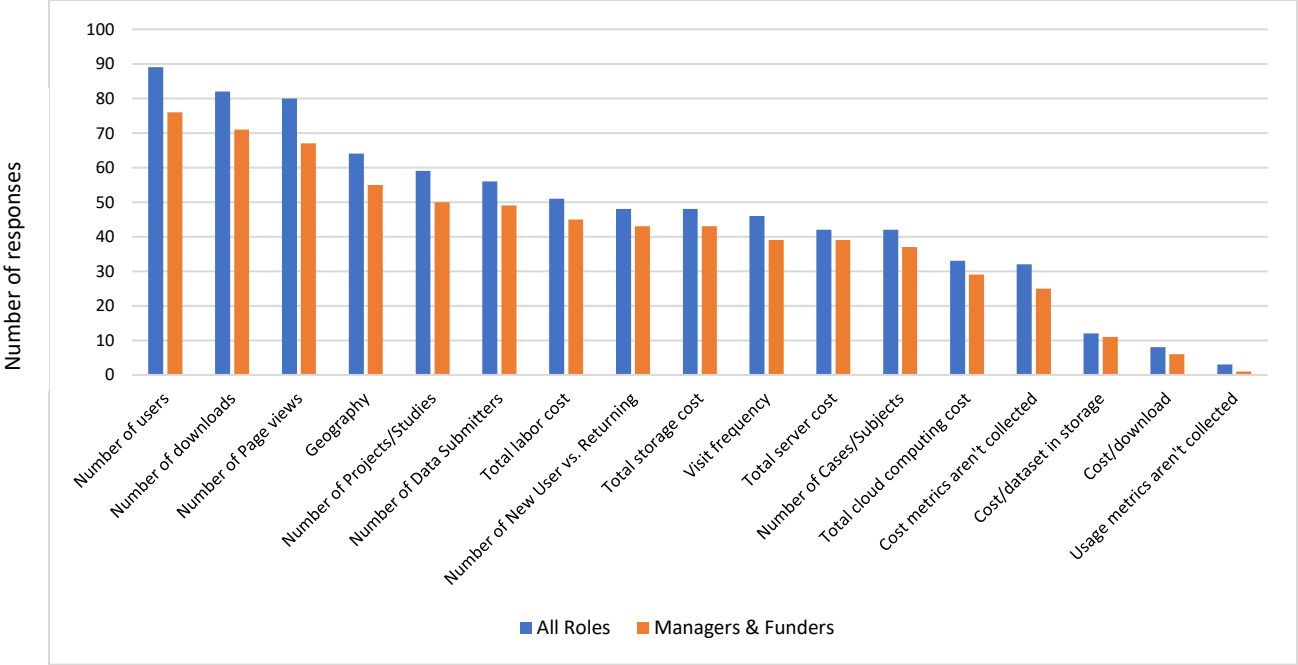

Figure 3. Metrics collected classified on role of the respondents

repository funders and managers are most relevant and therefore, the subsequent analyses will focus on the replies from the ninety-two respondents. Further, since Phase I included responses from repository managers and funders, a focus in the Phase II on the same subset of respondents allowed us to directly compare the results of the two phases.

Metrics tracked by resource managers and funders:

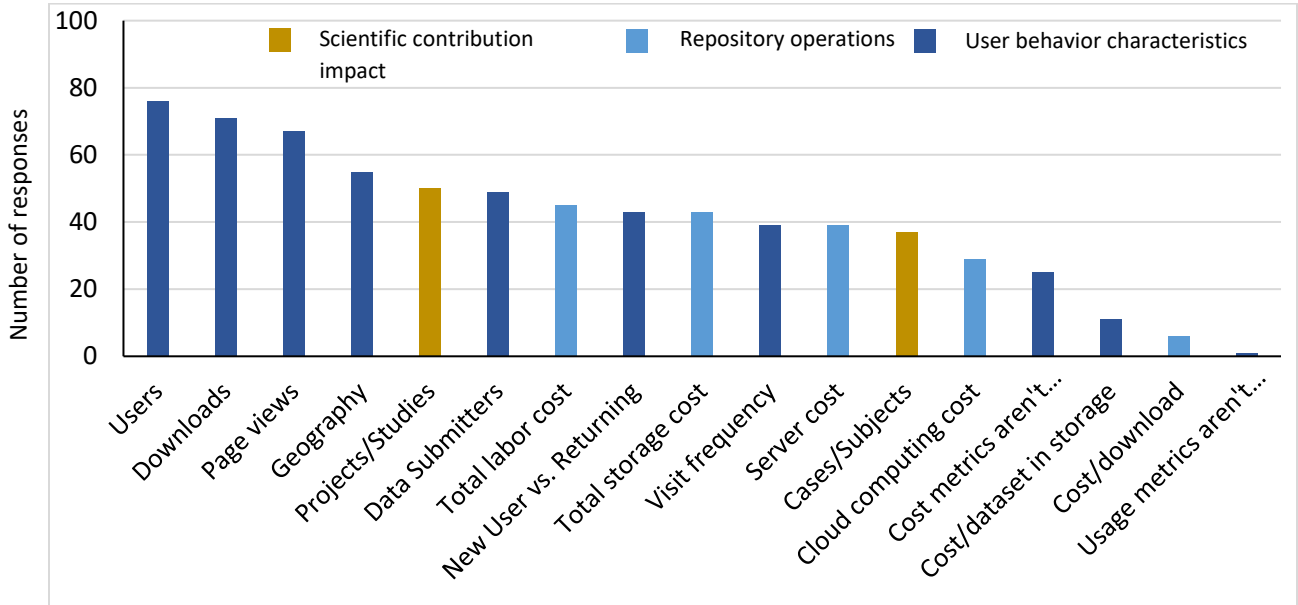

Figure 4. All metrics collected by resource managers and funders

Analysis of all metrics that were reported by the resource manager and funder subset showed a predominance of metrics in the "User behavior characteristics" (in blue). Repository operations and scientific contribution impact metrics were less likely to be tracked (in teal and brown respectively) (**Figure 4**). A significant number of the respondents (66 out of 92 repository managers and funders) indicated that they used Google Analytics, either alone or in conjunction with other tools, as a tool to capture metrics. Google Analytics predominantly allows for tracking of user behavior and flow of traffic on websites.

Comparison with Phase 1 assessment:

To determine the extent of concordance between the two phases, we compared the results of the Phase 1 analysis to the Phase 2 survey results. Since the Phase 1 approach was open ended and the Phase 2 semi-structured in format, we compared Phase 2 analysis results with a subset of data from Phase 1. The comparison was made in the two categories of "User behavior characteristics" (**Figure 5A**) and "Repository operations" (**Figure 5B**). The metrics that were tracked were consistent across the two phases of the approach (**Table 2** and **Figure 5**).

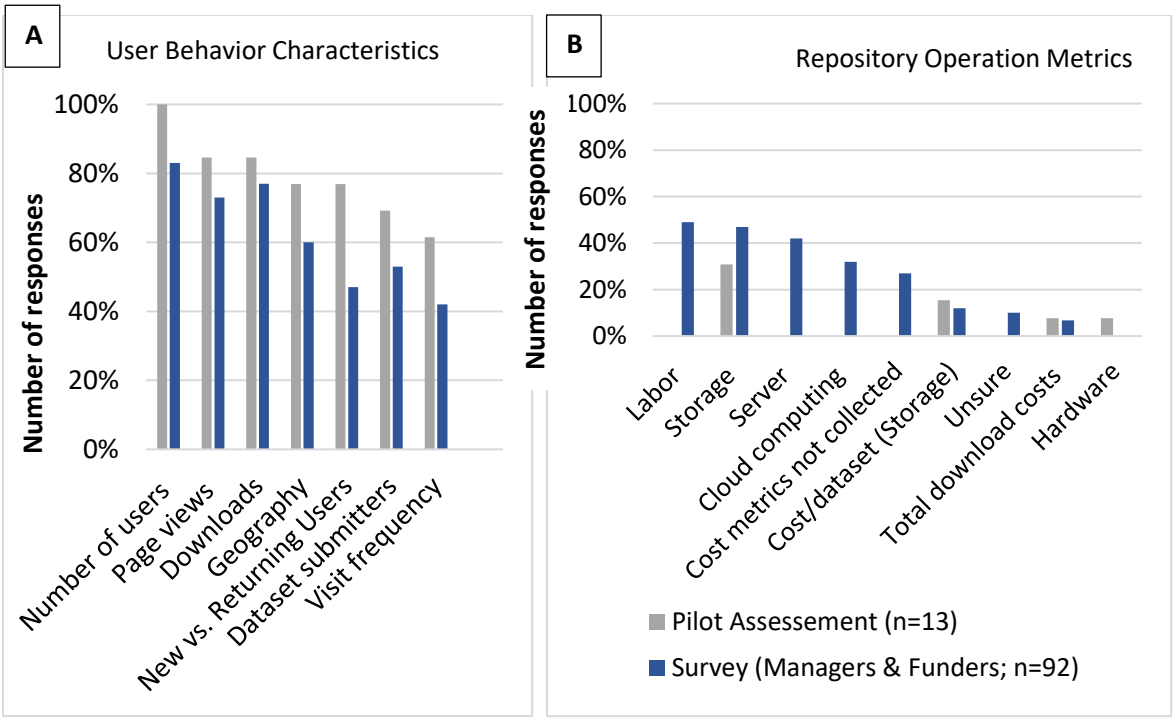

Figure 5. Comparison of Phase 2 community survey with Phase 1 assessment in the categories of **A)** User behavior characteristics and **B)** Repository operation metrics

Willingness to share metrics data:

As a funding agency, NIH has a special interest in improving its support to program management; therefore, the biomedical community's willingness to share metrics data provides input into advancing this overarching goal. Seventy-five respondents, representing 77% of the total responses, indicated a

general willingness to broadly share their repository metrics data (**Figure 6**) with a significant number (n=54, 59%) being extremely willing to share metrics data with the funding source.

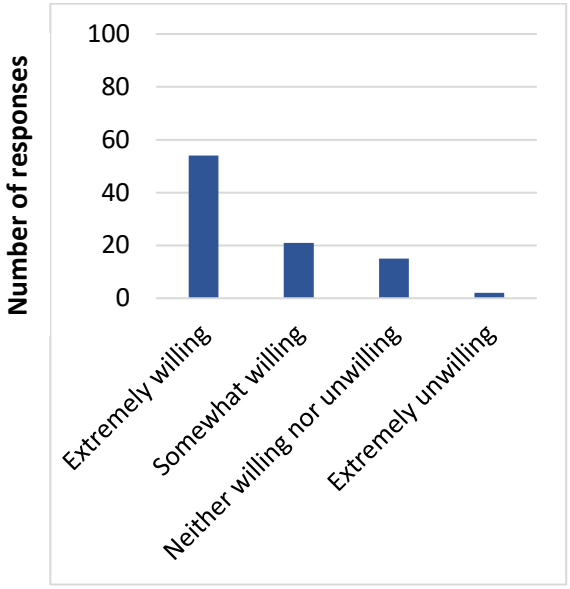

Figure 6. Respondents' willingness to share metrics data with funding source

Metrics to track in the future:

In response to the question "What metrics would you like to collect, but don't currently have the ability or infrastructure to collect?", the free-text responses (88 respondents) primarily identified metrics related to the scientific impact of the repository (57 respondents) (**Figure 7**). This finding may be related to the fact that scientific impact metrics were not included in the multiple-choice options, but also reflects the importance of scientific impact metrics and indicates a need for the community to define such metrics.

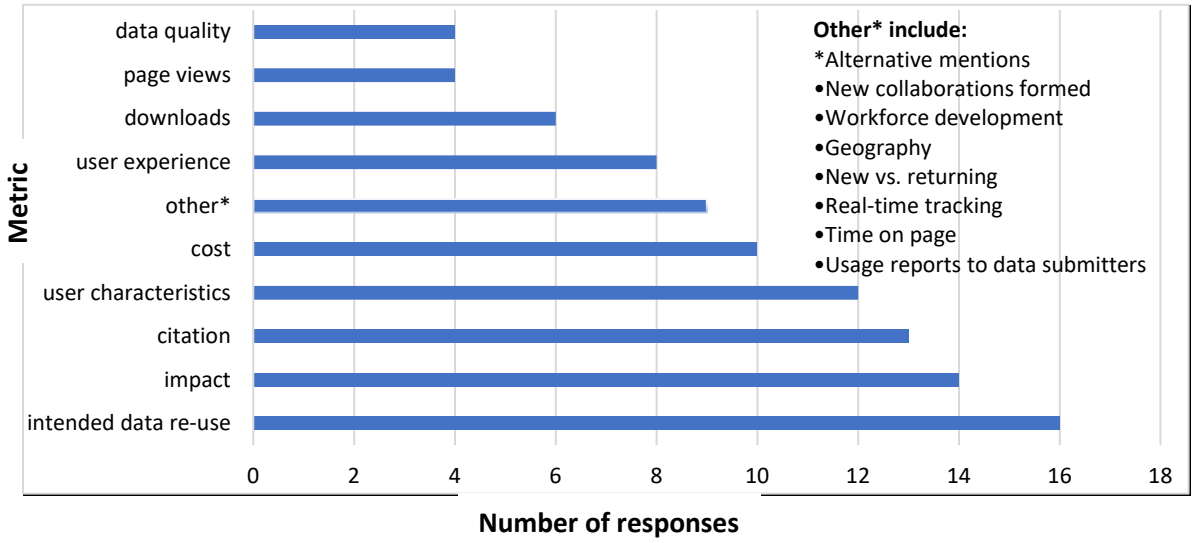

Figure 7. Free text responses of desired metrics currently not collected by respondents

Analysis of free-text responses:

In addition to collecting free-text responses to specific questions, the survey also included an open-ended question about any additional comments the respondent would like to provide, which 29 respondents answered. The two primary themes in these responses were **building on existing efforts related to metrics** and an **interest in further guidance from NIH**. Five comments mentioned existing data metrics efforts, such as Make Data Count, DataCite, and the COUNTER Code of Practice, encouraging NIH (and other funders) to collaborate and build upon these activities to encourage consistency in metrics, enable comparison across funders and repositories, and leverage existing infrastructure. Other commenters encouraged NIH to provide guidance, such as two commenters who pointed to a need for guidance on what NIH would like to see in metrics reported and two who raised questions about how they could get NIH funding for repositories.

Eight commenters used this space to provide additional information about one of their survey responses, such as pointing out that the available responses options were not relevant to them or explaining their thought process in selecting a response. Notably, three of these addressed why they indicated they would not be willing to share metrics with a funder; two shared concerns about privacy and confidentiality and one worried that sharing metrics could threaten their ability to get funding.

Eleven responses provided general commentary, such as the four commenters who indicated they had plans for system upgrades to collect better metrics. The remaining general comments reflected on some of the challenges of using metrics for data, such as the difficulty in translating metrics into actual scientific impact and the overall need for better infrastructure and standardization of metrics.

Alignment with TRUST principles:

Recognizing the importance of the TRUST principles [2] to repository management, we undertook an exercise to correlate the identified metrics to the TRUST principles. **Table 3** provides a mapping between the metrics and the principles. It is important to note that the metrics that were compiled as a part of this project do not comprehensively cover the principles; conversely, all the metrics align with the principles.

Table 3: Alignment of Metrics in survey to TRUST principles (expanded in right column)

Categories	Metrics	TRUST*
User Behavior Characteristics	Number of users	U
	Page views	RU
	Downloads	RU
	Geography	U
	New vs. Returning Users	U
	Dataset submitters	U
	Visit frequency	U
	Data Access Requests	TRU
Scientific Contribution Impact	Number of Projects/Studies	U
	Number of Cases/Subjects	U
	Total publications	U
Repository Operations	Storage costs	S
	Cost/dataset (Storage)	S
	Hardware Costs	S
	Total download costs	S

***T=transparency; R=responsibility; U=user focus; S=sustainability; T=technology**

Discussion

Metrics review and analysis was initially carried out on a subset of NIH-funded and -managed repositories and subsequently expanded to survey across the biomedical community. The findings highlight those metrics that are most used by repositories in this space. This study should be viewed as a somewhat limited assessment, but even with the relatively small sample size, a pattern of metrics collection was observed. Our findings show that User Behavior Characteristics metrics are the most common metrics collected by these repositories. However, metrics that are potentially useful for assessing the FAIRness of datasets, like data quality and metadata completeness, are not tracked widely across the repositories (**Appendix 1**). The repositories directly managed by NIH rarely tracked metrics such as storage and personnel costs, which seemed to be important to the more general sampling set covered by Phase 2 in our study. The metrics most frequently collected by the repositories often have readily accessible existing tools built to support the collection of these metrics, therefore leading to an easier tracking/collection process than what exists for other metrics that are not collected as often, if at all. Therefore, it may be the case that repositories are interested in collecting additional metrics beyond those identified here, but may lack the tools to do so. We also found that the repositories use similar analytics tools and methods in keeping track of these metrics. Factors such as lack of infrastructure to support collection, and lack of incentives to do so may be involved in the low adoption of the collection of some metrics (e.g., metrics for tracking interoperability, operating costs).

The main limitations of this analysis include the relatively small set of repositories used in assessing the current landscape of metrics and that the repository representatives had to provide their feedback on a pre-determined list compiled by the working group. This approach means that some metrics of interest to repositories may not be captured here. In addition, the survey respondents and their affiliated repositories reflect only a subset of the biomedical repository community and therefore these results should not necessarily be considered generalizable to the entire community, or to the broader repository community beyond the biomedical space. It would be critical to continue to solicit feedback from the communities on the topic of metrics that are of import to data resources and their management, as well as to consider other work on metrics currently underway in the broader community, such as the work of Make Data Count [14]. HOW ARE BIOMEDICAL REPOS UNIQUE?

In general, our findings show that these metrics, when properly tracked and utilized, can be used by repository stakeholders to assess usage, performance, and scientific impact of the repositories. They can also be applied to most data repositories and knowledgebases regardless of specific scientific research area. Understanding and filling these gaps in metrics collection may involve more targeted research studies in these specific areas.

Conclusions and Recommendations for Future Work

Based on the two phases of our study, we have identified a set of metrics that are currently used by and may be broadly applicable to biomedical data resources (**Table 2**). Future activities can help further inform NIH's understanding of this space, including reports from the NOSI (https://grants.nih.gov/grants/guide/notice-files/not-od-21-089.html) . As we continue to move towards more Trustworthy repositories [2], standards pertaining to data resource metrics will be integral to the process. Adoption of metrics reporting standards such as the 'Counter Code of Practice for Research Data Usage Metrics' [16, 19], which provide standards for generation and distribution of data usage

metrics for research, will be important in achieving this goal. Use of standardized reporting metrics and definitions will also be useful for development of Application Programming Interfaces (APIs) that can be used in automated generation of core metrics reports and nimble enough to be deployed across different platforms. This report intends to help inform the identification, adoption, and consistency of metric use across repositories.

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18
2. Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7
3. NIH Strategic Plan for Data Science (2018). Retrieved from https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf
4. https://datascience.nih.gov/data-ecosystem/nih-virtual-workshop-on-data-metrics
5. Cassella, Maria. (2010). Institutional Repositories: An Internal and External Perspective on the Value of IRs for Researchers' Communities. Liber Quarterly: The Journal of European Research Libraries. 20. 10.18352/lq.7989.
6. Bruns, Todd & Inefuku, Harrison. (2015). Purposeful Metrics: Matching Institutional Repository Metrics to Purpose and Audience. 10.2307/j.ctt1wf4drg.21.
7. Parr, C., Gries, C., O'Brien, M., Downs, *et al*. (2019). A Discussion of Value Metrics for Data Repositories in Earth and Environmental Sciences. Data Science Journal, 18(1), 58. DOI: http://doi.org/10.5334/dsj-2019-058
8. Anderson W, Apweiler R, Bateman A, *et al*. (2017). Towards Coordinated International Support of Core Data Resources for the Life Sciences. bioRxiv 110825; doi: https://doi.org/10.1101/110825
9. National Academies of Sciences, Engineering, and Medicine. 2020. Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs. Washington, DC: The National Academies Press. https://doi.org/10.17226/25639
10. CoreTrustSeal, https://www.coretrustseal.org/
11. nestor, https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html
12. ISO16363, http://www.iso16363.org/
13. https://datascience.nih.gov/data-ecosystem/biomedical-data-repositories-and-knowledgebases
14. Kratz, JE and Strasser, C. 2015a. Making data count. Scientific Data, 2: 150039. DOI: https://doi.org/10.1038/sdata.2015.39
15. Kratz, JE and Strasser, C. 2015b. Researcher perspectives on publication and peer review of data. PLOS ONE, 10(4): e0123377. DOI: https://doi.org/10.1371/journal.pone.0123377
16. Fenner, M, Lowenberg, D, Jones, M, Needham, P, Vieglais, D, Abrams, S, Cruse, P and Chodacki, J. 2018. Code of practice for research data usage metrics release 1. PeerJ Preprints, 6: e26505v1. DOI: https://doi.org/10.7287/peerj.preprints.26505v1
17. http://www.scholix.org/
18. Qualtrics XM, https://www.qualtrics.com/
19. The Code of Practice for Research Data Usage (2018). Retrieved from https://www.projectcounter.org/code-of-practice-rd-sections/foreword/

Appendix 1: List of all mapped metrics and their descriptions

Category	Metric	Description	Repository Tracking Count
User Behavior Characteristics	Page views	Clicks, page scrolling, mouse movement/pointing	11
	Downloads	Number of downloads or users downloading data, web or FTP	11
	Time on page	Time on a specific page	8
	Events	Clicks, page scrolling, mouse movement/pointing	7
	Data Access Requests	How many data requests are made in a specified time period	7
	Number of users	Number of users who can use the (visualization, e.g.) services for the data	13
	Geography	User IP address based - resolved to country/state	10
	New vs. Returning	For a defined period, usually three months	10
	Browser, Operating System, Device	Browser, Operating system, used to access the repository	9
	Subsite usage	Did they use more than one part of our site across visits?	9
	Dataset submitters	Number of data submitters	9
	Logged in vs. not	How many users are logged into the system and how many are not	8
	Visit frequency	Daily, monthly, etc.	8
	Data service users	Number of users who can use the (visualization, e.g.) services for the data	4
	Mediation services / staff	Number of users who use mediation services / staff	1
	SUS	Cost of measuring customer satisfaction with repository	0
	Summary statistics	Pages or clicks or events per visit. Generally, more pages or clicks in a visit = higher 'engagement'.	11
	Landing or Entry Page	What was the first page of the visit?	9
	Referrer	Where did they click to arrive at our site?	8
	Exit Page	What was the last page of the visit?	7
	Page Path	Proceeding from a search form to a search result, to a detail page. Did the user follow the intended or 'expected' path?	7
Scientific Contribution	Number of Projects/Studies	Number of Projects or Studies	10
	Number of Cases/Subjects	Number of Cases or Subjects	10
	Total publications	Total number of publications over all years	8
	Publications/year	Number of publications/year	6
	Workforce development	Students or staff trained (workforce development, as a direct result of repository activities)	4
	New research inspired or stimulated	Measure of new research inspired (stimulated); expected to increase if research moves to data-driven	4
	Publication citation	Average and maximal number of citations of the publications	3
	Journals citing dataset	Number of journals citing dataset from data repository	3
	Alternative mentions	Alternative metrics (Wikipedia pages, blogs, and impact on general public. Likely important, but still unspecific	3
	Curricula	Number of curricula (or students reached by curricula) using dataset (likely needs teacher surveys)	2
	Repo/Dataset mentions	Identify mentions of dataset in papers	2

Category	Metric	Description	Repository Tracking Count
	Publication IF	Average and maximal impact factors of journals of the publications	1
	Publication RCR	Average and maximal Relative Citation Ratio (RCR) of the publications	0
	Proposals generated	Number proposals generated (impact on science/innovation)	0
	Policy or management decisions impacted	Number of policy or management decisions impacted	0
Metadata/Data	Detailed metadata	Number of users identifying the applicability of the data (hits on page with detailed metadata)	8
	Data quality policies	Data quality policies (does policy exist; or acceptance rate for data)	6
	Standardized metadata	Percent conforming to standardized metadata (recognized specs)	5
Metadata/Data	Metadata completeness	Percent of holdings with complete metadata (provides an efficient appraisal of usefulness)	4
	Metadata acceptance rate	Repository application of a policy for completeness of metadata (e.g, acceptance rate for metadata)	3
	Standard methods	Standard collection methods described, e.g., potential use for synthesis of data	3
	Time spent applying acceptance criteria	Time spent on applying acceptance criteria (may include volunteer hours)	3
	Temporal coverage	Amount of long-term data (number of years); or general time coverage/period of data disseminated	2
	Metadata creation	Cost of metadata creation (levels: discovery, citation, data quality, machine readable)	2
	Dataset uniqueness	Datasets have been appraised for uniqueness (data that cannot be recollected)	1
	Future Findability, Accessibility	Cost of enabling future discoverability and access	1
	Fraud mitigation or avoidance	Enabling transparency of research; (e.g, FOI-request avoidance?) potential to mitigate or avoid fraud	1
	Quality statistics	Metadata expressing measure of Quality of data disseminated	1
	Importance	number of cases where datasets were important to a study (could break down to Vital, important, referred)	0
	Findability, Accessibility	Cost of enabling efficient discovery and access (development of ontologies, e.g.); could be redundant with other staff costs	0
Performance and Errors	Errors - Server	Application crash aka "500 error	7
	Latency	Network lookup time - how long after a request does it take for our server to respond?	6
	Page load time	How long for users to see the complete page, including graphics or other content	6
	Errors - client	Javascript error, missing image, etc.	4
Repository Operations	Providing user support	Cost of providing user support	5
	Storage costs	Total storage cost for repository	4
	Funding	Total funding amount per year, direct cost per year	3
	Cost/FTE	Cost per Full Time Employee (FTE), number of staff positions	3
	Policy development	Time on Policy development	2
	Distribution/Download costs	Benefit of avoidance of distribution cost to data producer (could be time saved)	2
	Cost/dataset (Storage)	Cost per dataset (i.e. Storage)	2

Category	Metric	Description	Repository Tracking Count
	Licensing	Licensing costs	1
	Hardware costs	Total hardware costs	1
	Total download costs	Total download costs	1
	Software development/improvement	Ongoing software development/improvement	0
	Preservation / infrastructure	Preservation/infrastructure	0
	Cost/active user/day	Cost of active user per day	0
	Cost/IP address	Cost per IP address	0
	Cost/publication	Cost per publication	0
	Cost/RCR	Cost per RCR	0
	Cost/Citation	Cost per Citation	0

Appendix 2: Repository Survey Questions

1. What is your role? (Select all that apply.) *
   - Data Resource Manager
   - Funder
   - Submitter
   - Data user
   - Educator
   - Other (Please describe.)

2. What type of data resources do you work with? (Select all that apply.) *
   - Data repository
     Data repositories store, organize, validate, and make accessible the core data related to a particular system or systems. For example, core data might include genome, transcriptome, and protein sequences for one or more organism.
   - Knowledgebase
     Knowledgebases accumulate, organize, and link growing bodies of information related to core datasets. A knowledgebase may contain, for example, information about gene-expression patterns, splicing variants, localization, and protein-protein interactions and pathway networks related to an organism or set of organisms.
   - Hybrid (both repository and knowledgebase)
   - Other (Please describe.)

3. If you manage more than one data resource, please choose one for the rest of the questions. What is the name of this data resource that you fund, manage, or use? *

3a. What is the URL of the data source above? *

4. If your data resource assigns persistent identifiers (PIDs) to data objects, what type of PID are you using?
   - DOI (https://www.doi.org)
   - ARK (https://en.wikipedia.org/wiki/Archival_Resource_Key)
   - Accession number
   - Other (Please describe.)

5. Is this a generalist or domain specific data resource? *
   - Generalist (e.g., multiple data types, no focus area)
   - Domain Specific (e.g., a single data type or focus area)
   - Not sure

6. How is the data made available? (Select all that apply.) *
   - Public access
   - Registration required
   - Controlled access (approval required to access human data)
   - Tiered access (with different permission levels)
   - Other (Please describe.)

7. How often are data submitted or curated? *
   - Daily
   - Weekly
   - Monthly
   - Annual
   - Not active
   - Not sure
   - Other (Please describe.)

8. What type of usage metrics do you track? (Select all that apply.) *
   - Number of users
   - Number of Page views
   - Number of downloads

- Geography
- Number of New User vs. Returning
- Number of Projects/Studies
- Number of Cases/Subjects
- Number of Data Submitters
- Visit frequency
- Usage metrics aren't collected
- Unsure

9. What type of cost metrics do you track? (Select all that apply.) *
- Total storage cost
- Total server cost
- Total cloud computing cost
- Total labor cost
- Cost/dataset in storage
- Cost/download
- Cost metrics aren't collected
- Unsure

10. If you track additional metrics to evaluate the data resource that are not mentioned in the previous question, please describe them:

11. What metrics would you like to collect, but don't currently have the ability or infrastructure to collect? *

12. What tool do you use to collect usage metrics? (Select all that apply.) *
- Google Analytics
- Other commercial tool (Please name other tool below.)
- Open source tool (Please name tool below.)
- Inhouse-developed tool
- Other (Please describe.)
- Metrics aren't collected

13. How do you use the metrics you collect currently? (Select all that apply.) *
- To improve technical capabilities and performance
- To improve user experience
- To facilitate budget and/or resource allocation decisions
- Other (Please describe.)
- Metrics aren't collected

14. How willing are you to share a usage report with a funding source? *
- Extremely willing
- Somewhat willing
- Neither willing nor unwilling
- Somewhat unwilling
- Extremely unwilling

15. Do you have any additional comments?

** Indicates required question.*

Survey information: OMB Control Number: 0925-0648 Expiration Date: 05/31/2021

Appendix 3: NIH Repository Metrics & Lifecycle working group and the Metrics for Repositories (MetRe) working group members

The MetRe group conducted the phase 1 analysis and generated the first version of this report. The Repository Metrics & Lifecycle group undertook the public survey, Phase 2 analysis, and produced the final report. All Working Group activities are under the auspices of the NIH Office of Data Science Strategy.

Tanya Barrett (NLM)[2]
Regina Bures (NICHD) [2]
Elaine Collier (NCATS)[1,2]
Lisa Federer (NLM) [1,2,*]
Kerry Goetz (NEI)[2]
Anupama Gururaj (NIAID) [1,2,*]
Lucy Hsu (NHLBI) [2]
Jennie Larkin (NIA) [1,2]
Sharon Lawlor (NIDDK) [2]
Dawei Lin (NIAID) [1,2]
Fenglou Mao (OD)[1,2]
Matthew McAuliffe (CIT)[1,2,*]
Christine Melchior (CSR)[1,2]
Leonie Misquitta (CIT)[2]
Noffisat Oki (NIAID) [2,*]
Kim Pruitt (NLM)[1,2]
Rebecca Rodriguez (NIDDK) [2]
Eric Sayers (NLM) [2]
Charles Schmitt (NIEHS) [1]
Alyssa Tonsing-Carter (OD)[1]
Vivian Ota Wang (NCI) [2]
Minghong Ward (NLM) [1,2,*]
Susan Wright (NIDA) [1,2]

*The primary authors are denoted by an *. Metrics & Lifecycle group and MetRe group members are denoted by [1] and [2], respectively*.