Ripple Effect COMMUNICATIONS, INC.

INTELLIGENT PROJECT MANAGEMENT™

NIH REQUEST FOR INFORMATION: MANAGEMENT, INTEGRATION, AND ANALYSIS OF LARGE BIOMEDICAL DATASETS

ANALYSIS OF PUBLIC COMMENTS

MAY 10, 2012

This report includes an analysis of the comments received through the NIH Request for Information (RFI): Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics (NOT-OD-12-032).

Executive Summary

In response to the exponential growth of large biomedical datasets, the National Institutes of Health (NIH) Advisory Committee to the Director (ACD) has formed a Working Group on Data and Informatics.[1] The Working Group was charged with the task of providing expert advice on the management, integration, and analysis of large biomedical datasets. As part of the process, the Working Group gathered input from the extramural community through a Request for Information (RFI): "Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics" (NOT-OD-12-032).[2] Ripple Effect Communications, Inc. was contracted to provide third party analysis of the comments received through the RFI; this report provides analysis of the 50 responders to the RFI and summarizes the 244 respondent suggestions. The Working Group will make recommendations to the ACD to assist in developing policies regarding the management, integration, and analysis of biomedical datasets.

The Data and Informatics Working Group (DIWG) identified a total of six issues and seventeen sub-issues as important to consider for enhancing data management and informatics. The six issues were:

- Scope of the challenges/issues
- Standards development
- Secondary/future use of data
- Data accessibility
- Incentives for data sharing
- Support needs

Respondents were asked to consider the identified issues as they responded to the following three questions:

1. For any of the areas identified above and any other specific areas you believe are worthy of consideration by the Working Group, please identify the critical issues(s) and impact(s) on institutions, scientists, or both.

2. Please identify and explain which of the issues you identified are, in your opinion, the most important for the Working Group to address and why.

3. Please comment on any specific ways you feel these issues would or should affect NIH policies or processes.

DATA AND METHODS

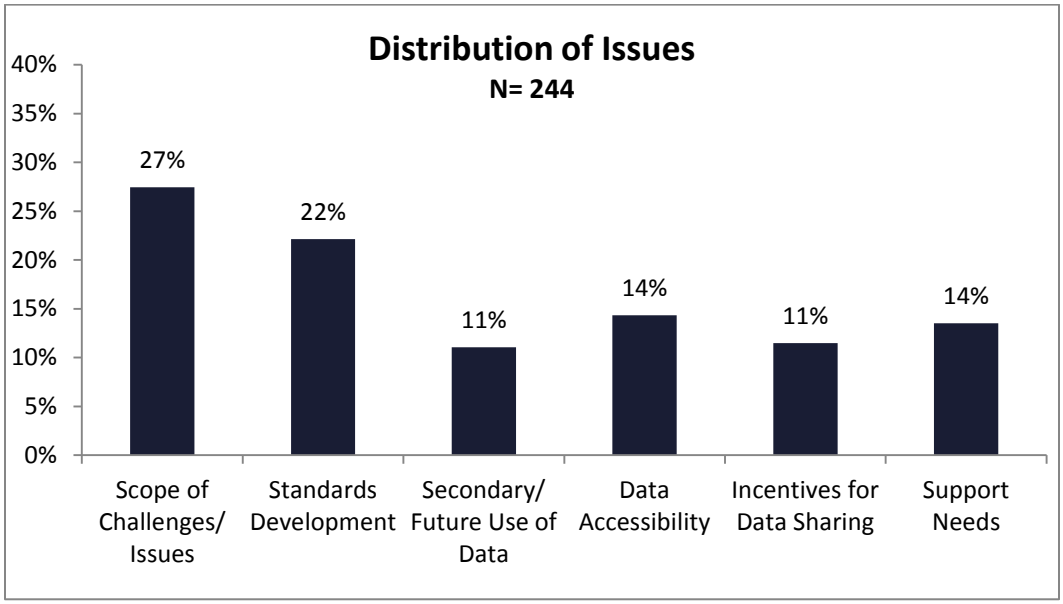
NIH received input from 50 respondents, most of whom provided feedback from a personal perspective (self, 70%; organization, 30%). The 50 respondent submissions were parsed into 244 comments and coded according to the issues identified by the Working Group, as well as by other issues that emerged from the data.

[1] http://acd.od.nih.gov/diwg.htm
[2] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-032.html

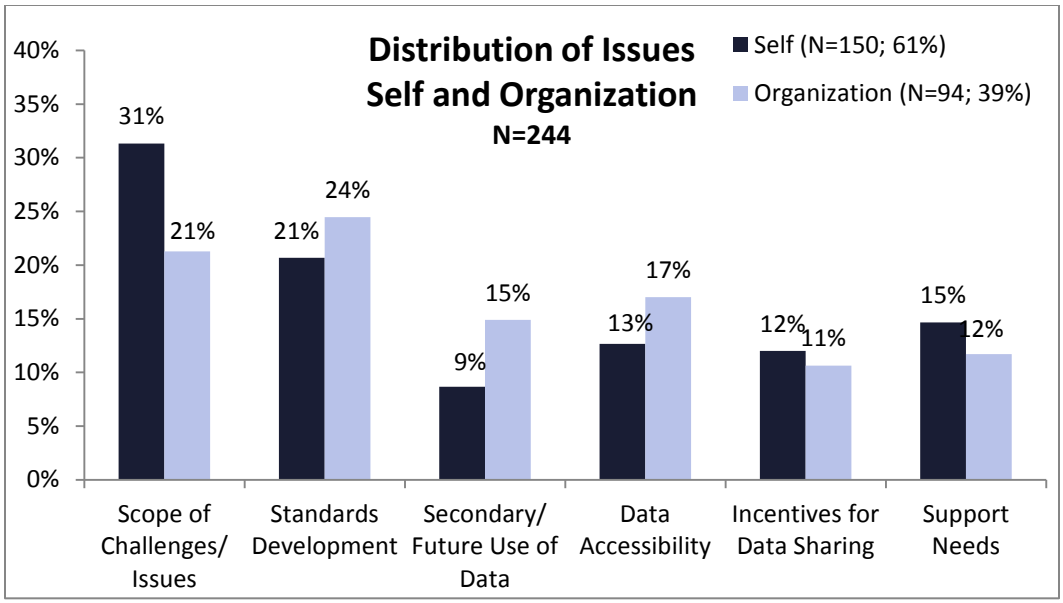
A coding scheme was developed based on six issues and seventeen sub-issues identified by NIH. That structure provided the conceptual foundation, which team members further developed using an iterative, grounded theory approach. The final coding scheme consisted of the six issues and the seventeen sub-issues identified in the RFI, plus three additional sub-issues derived from the data. A total of twenty sub-issues are described in this report. In total, twenty "codes" were applied to the data; these corresponded to the twenty sub-issues.

FREQUENCIES, PRIORITY, AND RECOMMENDATIONS

Of six issues identified by NIH, respondents most frequently commented about the Scope of Challenges/Issues (27%). This issue was followed by Standards Development (22%) and Data Accessibility (14%) to create the top three most frequently-coded issues.



When analyzed by self-reported affiliation, there were slight differences in how the codes were distributed. Those who self-identified as commenting from a personal perspective (self) commented more frequently about Scope of Challenges/Issues, Incentives for Data Sharing, and Support Needs in the review process, compared to those who self-identified as commenting from an organizational perspective (organization).



Priority was assigned to comments when the respondent explicitly stated it was a priority concern. The top three issues when ranked by frequency were the same top three issues when ranked by priority: Scope of Challenges/Issues, Standards Development, and Data Accessibility.

Collectively, respondents recommended that NIH address data and informatics challenges by not only supporting an infrastructure, but also by supporting output and utilization of data needs such as enhanced organization, personal development, and increased funding for tool development.

Contents

Background

NIH REQUEST FOR INFORMATION

In response to the exponential growth of large biomedical datasets, the NIH ACD formed the Working Group on Data and Informatics. The Data and Informatics Working Group (DIWG) was charged with the task of examining issues related to data spanning basic science through clinical and population research; administrative data related to grant applications, reviews, and management; and management of information technology (IT) at NIH. The ACD will make recommendations on the management, integration, and analysis of large biomedical datasets.[3]

To help inform the development of recommendations, the Working Group announced a request for information (RFI), "Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics" (NOT-OD-12-032),[4] to gather input from various sources, including extramural and intramural researchers, academic institutions, industry, and the public. For the RFI, the Working Group identified the following issues and sub-issues as important to consider when developing recommendations:

- Scope of the challenges/issues
    - Research information lifecycle
    - Challenges/issues faced by the extramural community
    - Tractability with current technology
    - Unrealized research benefits
    - Feasibility of concrete recommendations for NIH action
- Standards development
    - Data standards, reference sets, and algorithms to reduce the storage of redundant data
    - Data sharing standards according to data type (e.g., phenotypic, molecular profiling, imaging, raw versus derived, etc.)
- Secondary/future use of data
    - Ways to improve efficiency of data access requests (e.g., guidelines for Institutional Review Boards)
    - Legal and ethical considerations
    - Comprehensive patient consent procedures
- Data accessibility
    - Central repository of research data appendices linked to PubMed publications and RePORTER project record

[3] http://acd.od.nih.gov/diwg.htm
[4] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-032.html

This report includes an analysis of the comments received through the NIH Request for Information (RFI): Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics (NOT-OD-12-032).

  - Models and technical solutions for distributed querying
  - Comprehensive investigator authentication procedures
- Incentives for data sharing
  - Standards and practices for acknowledging the use of data in publications
  - "Academic royalties" for data sharing (e.g., special consideration during grant review)
- Support needs
  - Analytical and computational workforce growth
  - Funding for tool development, maintenance and support, and algorithm development

Respondents were asked to consider the identified issues as they responded to the following three questions:

1. For any of the areas identified above and any other specific areas you believe are worthy of consideration by the Working Group, please identify the critical issues(s) and impact(s) on institutions, scientists, or both.

2. Please identify and explain which of the issues you identified are, in your opinion, the most important for the Working Group to address and why.

3. Please comment on any specific ways you feel these issues would or should affect NIH policies or processes.

The online submission process was open from January 10, 2012 through March 12, 2012. This report is an analysis and summary of the public comments and will serve as a tool for the Working Group to use as part of its process for making concrete recommendations to the NIH Director on ways to improve data management and informatics of large biomedical datasets.

THE ROLE OF RIPPLE EFFECT COMMUNICATIONS, INC.

Ripple Effect Communications, Inc. was engaged by the NIH Office of the Director to perform an analysis of the data received through the RFI. As an independent contractor, Ripple Effect staff is not invested in the ACD committee deliberations and therefore has no bias toward the outcomes of the assessment; however, Ripple Effect is uniquely positioned to bring a continuum of working knowledge and expertise about NIH to the analysis process. Our staff's diverse knowledge about NIH allow an open interpretation of respondents' thoughts and ideas, which not only ensures full expression but also provides context for understanding potentially complicated messages.

Ripple Effect was established in 2006 to provide "Intelligent Project Management"™ to the federal government and is often called upon to provide support in one or more of the following areas: Communications, Program and Policy, Technology, Conference and Events Management, Organization and Process Improvement, Research and Analysis, and Project Management. We assess, plan, manage, and execute projects that aid the government (with the current focus on increasing transparency) in transforming into a "people-centric, results-driven, and forward-thinking" organization.

Methods

We engaged both quantitative and qualitative research methods as part of the analysis process. While focusing on and maintaining the integrity and structure of the issues identified by the Working Group, we remained open to the data. We used grounded theory data analysis methods to capture the ideas that were either pervasive enough to warrant their own codes or went beyond the issues identified by the Working Group.

ABOUT THE DATA

A total of 50 respondents provided feedback to the RFI. Respondents provided a total of 244 comments, which were individually coded. All 50 were received through the online submission process that was open from January 10, 2012 through March 12, 2012. Seventy percent of respondents provided feedback from an individual perspective, while 30% identified an organizational affiliation.

ANALYSIS PROCESS

All submissions were uploaded and organized into a central SharePoint database. The data was parsed into individual comments, coded according to the issues identified by the Working Group, and others that emerged from the data, and then analyzed using both SharePoint and Excel.

Code Development

Code development began using the six issues and seventeen sub-issues identified by NIH as the conceptual foundation of the coding scheme. Team members further developed the coding scheme using an iterative, grounded theory approach, which involved studying the data, suggesting themes for inclusion, reviewing code application by other team members, and resolving disagreements.

Conceptually, the codes that emerged from the data were all at the sub-issue level. In addition to the seventeen sub-issues identified by NIH, three additional "data-driven" codes were developed and applied to the data. The final coding scheme (including code descriptions) included six issues and twenty sub-issues (Appendix A). The table below illustrates the conceptual levels and code names used throughout the report.

Issue	Sub-Issue
Scope of Challenges/Issues	Research Information Lifecycle Challenges/Issues Faced Tractability with Current Technology Unrealized Research Benefits Feasibility of Recommendations to NIH
Standards Development	Reduction of Redundant Data Storage Standards According to Data Type Metadata Quality Control^ Collaborative/Community Based Standards^ General Guidelines^

Issue	Sub-Issue
Secondary/Future Use of Data	Improved Data Access Requests Legal and Ethical Considerations Patient Consent Procedures
Data Accessibility	Central Repository of Research Data Models and Technical Solutions Investigator Authentication Procedures
Incentives for Data Sharing	Acknowledging the Use of Data "Academic Royalties" for Data Sharing
Support Needs	Analytical and Computational Workforce Growth Funding and Development for Growth

^Data-driven sub-issues

Priority

To assess the priority of the issues for each respondent, we included only the comments in which one of the following conditions was met:

1) The comment was included in response to Question 2, “Please identify and explain which of the issues you identified are, in your opinion, the *most important* for the Working Group to address and why.”
2) The commenter expressed priority by using words such as “critical,” “important,” or “essential.”

If no priority was indicated or if the commenter explicitly expressed that the item was NOT a priority, the comment was not included in the priority analysis.

Analysis was a straightforward count of the number of people who identified each issue and sub-issue as a priority. Priority is presented as an order based on the frequency with which each person identified a code, not as a mathematical rank. Analysis of this sub-group is presented in Section Two of the Findings.

NIH Responsibility

To assess how the respondents believed issues would or should affect NIH policies or processes, we captured and quantified comments that either explicitly expressed an action for NIH to take in order to improve data and informatics or that suggested the issue coded fell under the purview of NIH. Specifically, we included comments only when one of the following conditions was met:

1) The comment was located in response to Question 3, “Please comment on any specific ways you believe these or other issues would or should affect NIH policies or processes.”
2) The commenter specifically stated that NIH should be responsible.
3) The comment addressed an existing NIH program.

If the respondent explicitly stated that the item should NOT be the responsibility or purview of NIH or the comment was general and did not explicitly state NIH responsibility, it was not included in the NIH responsibility analysis.

Analysis occurred in two steps. First, we compared the frequency distribution of all sub-issues identified as an NIH responsibility with the overall dataset. Second, we reviewed data for overarching themes that informed explicit recommendations for NIH. Analysis of this sub-group is presented in Section Three.

Findings

Findings are divided into three sections that reflect different conceptual levels of analysis and respond to the questions posed in the RFI. The first section includes analysis in response to Question 1: "For any of the areas identified above and any other specific areas you believe are worthy of consideration by the Working Group, please identify the critical issues(s) and impact(s) on institutions, scientists, or both." This section provides a quantitative overview of the primary categories and issues, as well as a quantitative distribution and qualitative analysis of the twenty sub-issues.

The second section addresses Question 2: "Please identify and explain which of the issues you identified are, in your opinion, the most important for the Working Group to address and why." We coded and quantified the data for respondents that explicitly identified priority issues.

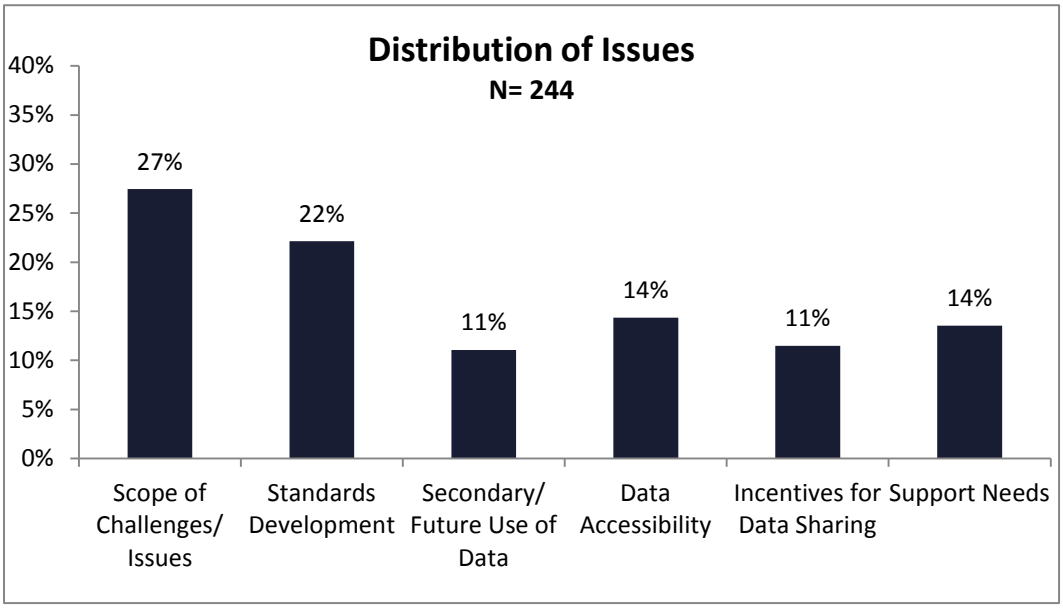
The third section includes a descriptive summary of the ideas commenters presented as relevant to Question 3: "Please comment on any specific ways you believe these or other issues would or should affect NIH policies or processes." We coded and quantified the comments that referred to specific recommendations for NIH.

SECTION ONE: QUANTITATIVE AND QUALITATIVE ANALYSIS OF CRITICAL ISSUES

A total of 50 (100%) responsive submissions were received and parsed into 244 individual comments. Each comment received one code (corresponding to one sub-issue) and was analyzed for frequency and content.

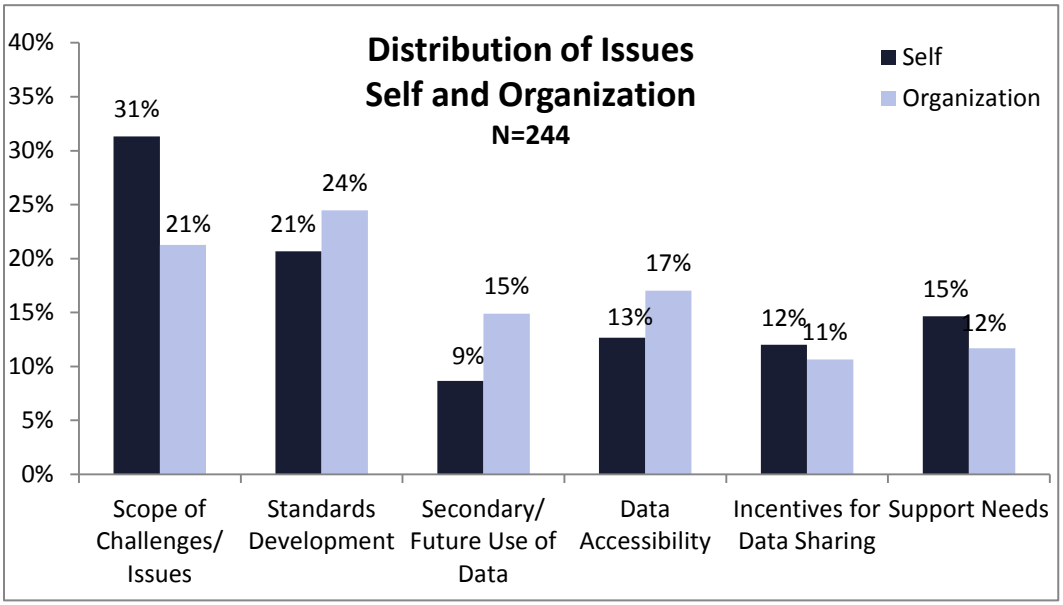
A Quantitative Overview of the Issues

Of the six issues identified by NIH, respondents most frequently commented about the Scope of the Challenges/Issues. The other top issues identified were Standards Development and Data Accessibility. When combined, these top three issues represent approximately two-thirds of all comments.



Issues by Respondent Affiliation

Respondents self-identified with one of two types of affiliation: as an independent individual (self) or on behalf of an organization (organization). Of the total 244 comments received, 150 (61%) were from those identifying as "self" and 94 (39%) were from those identifying as "organization." Those who responded from a personal perspective commented more frequently than organizations about Scope of Challenges/Issues, Incentives for Data Sharing, and Support Needs. Those responding on behalf of an organization commented most frequently on Standards Development, Data Accessibility, and the Secondary/Future Use of Data.

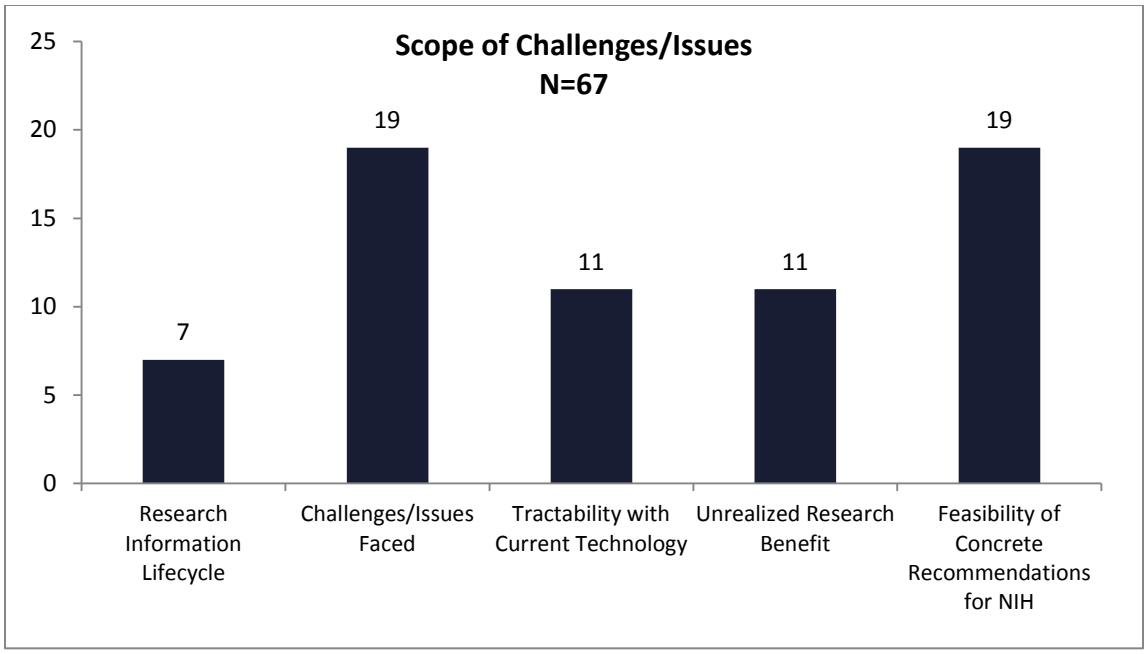


Quantitative and Qualitative Analysis of Issues and Sub-Issues

The six issues and twenty sub-issues, as identified by NIH and derived from the data, are illustrated and discussed here in detail. A graph that summarizes the frequency distribution of comments across all sub-issues is provided in Appendix B. Where relevant, the NIH-identified sub-issues are shown in blue, while data-driven sub-issues are shown in orange.

Issue One: Scope of Challenges/Issues

This issue targeted challenges regarding the management, integration, and analysis of large biomedical datasets. This issue was the most frequently mentioned; approximately one-quarter of all commenters were concerned with the Scope of the Challenges/Issues. Within this category, three leading topics emerged: Feasibility of Concrete Recommendations for NIH, Challenges/Issues Faced, and Tractability with Current Technology. These topics together made up two-thirds of the responses for this issue.



Research Information Lifecycle

For this sub-issue, one respondent outlined a data lifecycle model by describing a scientific community-driven collection of data with a national data infrastructure. In such a community-driven lifecycle, creators of a data set would generate data and input parameters as the first stage. In subsequent stages, other members of the research community would add to the existing data by providing additional context, such as how the data was generated. At the publication and preservation stages, a final detailed description of the data then would be available.

> *An example life cycle is the migration of data from a project collection, to a collection shared with other researchers, to a digital library for formal publication of vetted results, to a reference collection for use by future researchers. (#42)*

When describing the national data infrastructure, one respondent explained that each stage of the community-driven collection would be governed by policies.

Another respondent referred to Charles Humphrey's 2004 overview on research data lifecycles[5] stating that it is applicable to a variety of research disciplines. The respondent noted that, when considering management of analysis of datasets, the roles and responsibilities of the researcher needs to be determined by focusing on documenting the stages of the research lifecycle:

> *Design of a research project*
> *Data collection processes and instruments*
> *Data organization in digital format*
> *Documentation of data analysis process*
> *Publication or sharing of results*
> *Dissemination, sharing, and reuse*

[5]Humphrey, C. & Hamilton, E. (2004). Is it working? Assessing the Value of the Canadian Data Liberation Initiative." *Bottom Line*, *17* (4), 137-146.

Preservation, long-term conservation, and long-term access (#46)

Other comments revolved around hiring technicians involved in designing methodology, requiring electronic notebooks, maintaining better recordkeeping, and preserving and storing data.

Challenges/Issues Faced

This sub-issue referred to the challenges and issues presented by datasets in the biomedical field. Overall, respondents' comments were divided among data infrastructure, the need for well-trained individuals, and data accessibility, although most comments focused on data infrastructure. One respondent specifically stated that there was a lack of data infrastructure:

> *There are two major barriers to sharing of data: 1) Lack of an infrastructure for data sharing. It's not easy to share. Currently, scientists or universities need to set up their own sharing system (we are doing this using DATAVERSE) but there should be a system put in place by NIH/NLM for widespread sharing of data. Once the systems are in place, scientists will use them. (#1)*

One respondent stated that "we have the information, but we do not know how to use it." Others felt that a data system should be created to integrate data types, capture data, and create "space" for raw data.

Regarding the need for well-trained individuals, one respondent spoke passionately about laying off programmers due to lack of funding. Comments were emphatic about how much harder it is to replace a competent programmer than a lab technician.

Regarding data accessibility, most respondents spoke to the difficulty of finding useful data and databases for their particular area of interest, whether it be patient records, health care, or biomedical research. Encountering access issues in our current age of digital technology and electronic records was seen as especially frustrating. One respondent believed that there should be some type of direct access to data records that would facilitate many advances in the biomedical field.

> *What is most puzzling and distressing is that, in spite of our increasingly sophisticated technology and electronic data systems, researchers' direct online access to federal vital records data has become increasingly limited over time, impeding and sometimes precluding potentially valuable etiologic investigations. (#2)*

Tractability with Current Technology

For this sub-issue, there was consensus around a need for tracking current technology for data standards and standardized software. Suggestions to develop standards ranged from performing an analysis of the technology that has been successful or unsuccessful to understanding limitations posed by available computing hardware. Several respondents provided examples of current technology uses and suggestions to accommodate future growth. For example, a suggestion to improve electronic health records (EHRs) was:

> *… to significantly increase the size of the sample (one billion visits per year), the diversity of the population, and the length of follow-up time compared to what is currently feasible. (#4)*

The Nuclear Receptor Signaling Atlas (NURSA) and Beta Cell Biology Consortium (BCBC) were viewed as highly effective efforts that have evolved into successful management of large scale data.

Unrealized Research Benefit

Respondents to this sub-issue consistently agreed that research products involving datasets, data sharing, and administrative data are not being properly utilized. Large amounts of data are not being considered or analyzed. Reasons for such underutilization included poor planning of grant resources, negative results, poor documentation, lack of data sharing compliance, and lack of data retention. Respondents called for open access and offered the Open Government Initiative and the International Household Survey Network as model examples.

> *Great progress has been made in data sharing in many disciplines such as genomics, astronomy, and earth sciences, but not in public health. Developments such as the Open Government Initiative by the US Federal Government and the International Household Survey Network supported by the World Bank provide a promising start but will require a wider support base for a paradigm shift for data sharing in public health. (#31)*

Respondents believed that providing a more open forum to data sources would improve success rates.

Feasibility of Concrete Recommendations for NIH

This sub-issue captured comments that provided feasible recommendations for NIH to improve data sharing, data storage, data management, etc. Many commenters suggested that NIH maintain an up-to-date data directory, create an organizational structure, obtain adequate memory for computer systems, and develop algorithms.

One respondent contributed step-by-step procedures to manage the influx of large datasets.

> *More pointedly, as NIH moves to larger and larger data sets, and federations of data sets, it will discover that the I/O performance of most systems will be inadequate to handle the volume of data in a timely fashion. Solving this problem requires getting many things right, from organizing the data so that it can be accessed efficiently, to picking representations that allow it to be manipulated efficiently in the available memory of the computer systems, to developing algorithms and data management interfaces that work well with peta- to exabytes of data, and, last but not least, to designing the storage and I/O systems to maximize the transfer rate between disks and memory. (#35)*

Another respondent elaborated on the same concern by providing specific examples in software development and hardware configuration.

> *What are the non-mainstay innovations that will/could be required? To meet some of the challenges in terms of "population scale" analysis we need a fundamental change in how software is being developed, the methodologies used and the under lying hardware configurations. Such forward thinking seems to be within the remit of the group. Examples of innovations could include: considering how affordable and usable HPC can be made available (e.g. easier to use programmable chips or GPUs, extensions to PIG or other scripting systems for distributed processing/HDFS) or how we can develop scalable/affordable/usable software more easily without introducing constraining requirements on teams (e.g. education, reuse of open-source initiatives (see section 3)). (#14)*

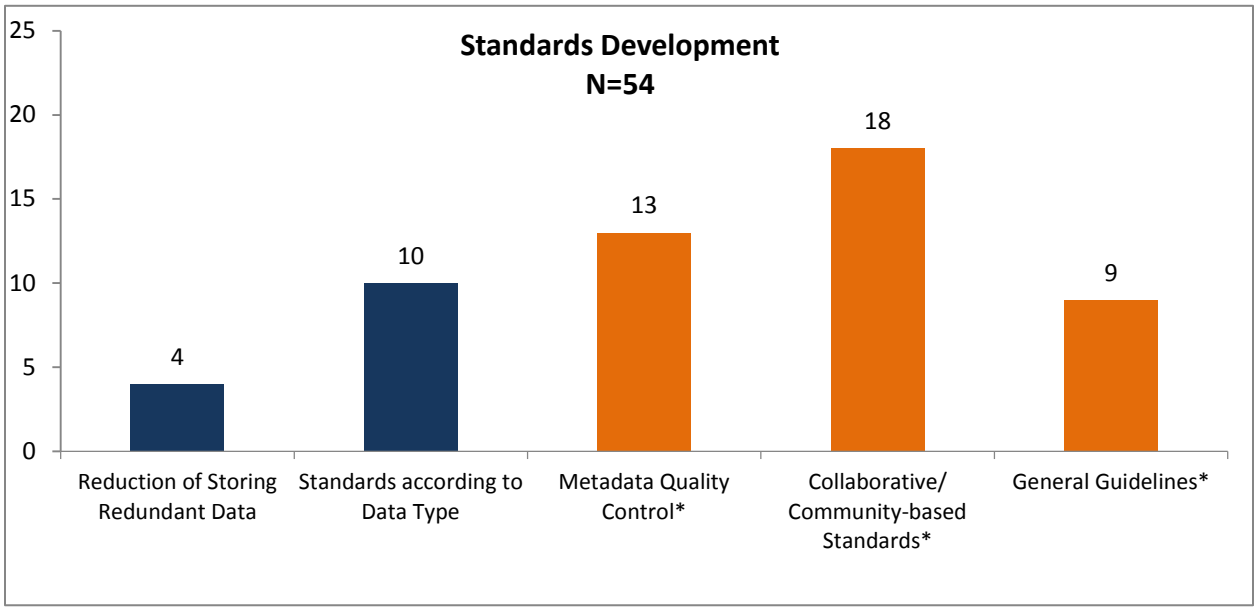
A common suggestion from respondents was the integration of data into a master system. While respondents agreed upon the need for a system, some suggested the goal of this system was data

management while others wanted to create a system for publications, patient records, or enforcement of diversity sampling.

Another respondent identified the need for increased training grants that would provide biostatisticians and bioinformatics specialists with strong scientific backgrounds to provide the appropriate level of technical support to assist with large datasets.

Issue Two: Standards Development

Within this issue, respondents felt that it was important to develop organized standards for current data and to also establish standards for future data. The sub-issues originally identified for this issue were joined by three additional sub-issues that emerged from the data (Metadata Quality Control, Collaborative/Community-based Standards and General Guidelines).



Reduction of Redundant Data Storage

Most comments within this sub-issue expressed the opinion that redundancy is an issue primarily because of the increasing amount of data that is being created without oversight or coordination. Respondents suggested strategies for reducing redundant data:

- Establish standards and policies
- Disseminate and preserve data
- Build a proper support network

One respondent commented that data tends to be dispersed; therefore, cross referencing the data is not simple. Possible solutions to remedy the issue were offered.

> *There is a need for better: i) schema integration, ii) schema mappings to navigate from one data source to another, iii) complex join across databases, iv) support for provenance data, v) flexible resource discovery facilitated by a richer metadata registry. [This] item reflects implicit needs for better metadata that will facilitate the selection and the location of distributed data resources. (#43)*

In general, respondents agreed that the identification of data standards, reference sets, and algorithms were strategies to reduce the storage of redundant data.

Standards According to Data Type

Respondents believed that standards should be developed for distinct data types, such as phenotypes, molecular profiling, imaging, raw versus derived, clinical notes, and biological specimens. One universal theme was the need for a consortium to handle the variety of data types, especially because some respondents believed that creating one general standard would be difficult or impossible.

> *While "universal" standards are theoretically appealing, in practice they have proven difficult, if not impossible, to implement. The WGDI must, therefore, avoid a one-size-fits-all approach and should consider a variety of data sharing models and standards to accommodate the diversity of data types. (#18)*

Respondents emphasized the diversity in data types by highlighting features such as the abundance of non-genomic data associated with patients (EEG reports, imaging, biochemical workups, and reactions to therapeutic interventions). To take this concept one step further, one respondent suggested developing a "biomaterials enterprise interlinked for data access and integration."

> *Coordination of acquisition sites for data uploading is a key factor, as is coordination of databases (or synchronization mechanisms if a federated archive is deployed) by data type, e.g., image data vs. genetic data. Biospecimen banking may be optimally conducted elsewhere or separately from the data coordinating center, with the biomaterials enterprise interlinked for data access and integration as needed by project or user. (#27)*

Additionally, respondents agreed that a system should be developed to create consistency in annotating data standards.

*Metadata Quality Control^*

This sub-issue evolved from the data and captured comments related to organizing data and/or improving data quality control with respect to uniform descriptors, index categories, semantics, ontologies, and uniform formats. One respondent specifically noted that this issue was not addressed in the RFI.

> *The current list of areas does not identify data quality as an area of focus for this agenda. There currently exist no established data quality assessment methods, no established data quality standards, and no established data quality descriptors that could be attached to each data set. In the absence of data quality descriptors, a down-stream user of the data has no ability to determine if the data set is acceptable for the intended use. A data set that is acceptable for one use may or may not be acceptable for a different use. (#7)*

Other respondents agreed that data sets lacked well-formed metadata. They believed that the development of standards for metadata is fundamental in ensuring that data will survive and remain accessible in the future.

*Collaborative/Community-Based Standards^*

Some respondents specifically addressed the process of standards development, stating that community-led collaborative efforts were needed to muster broad support for new standards and reduce competing standards. All should have a voice and a stake in this "information ecosystem": researchers, government agencies, universities, students, publishers, industry, associations, educators, librarians, data scientists, patients and study subjects, the public.

> *Development of such a workforce should be modeled on exemplar efforts such as the NSF DataNets, the Digital Curation Center in the UK, and the Australian National Data Service. This community is needed to help shape and support general policy and infrastructure within and among agencies, and to help spread data expertise into the educational and research communities. At the same time, grass-roots 'communities of practice' must engage disciplinary scientists in order to determine how to implement general agency policies. (#45)*

*General Guidelines^*

Some respondents emphasized the need for guidelines on data management, access and sharing, and some included the necessity for training in guideline usage and compliance. Some respondents specified particular guidelines (e.g., for research funders) for archiving and accessing paper records of public health data for future needs. Some focused on cost issues, others on how to determine who should have access. Some listed concrete suggestions for policy:

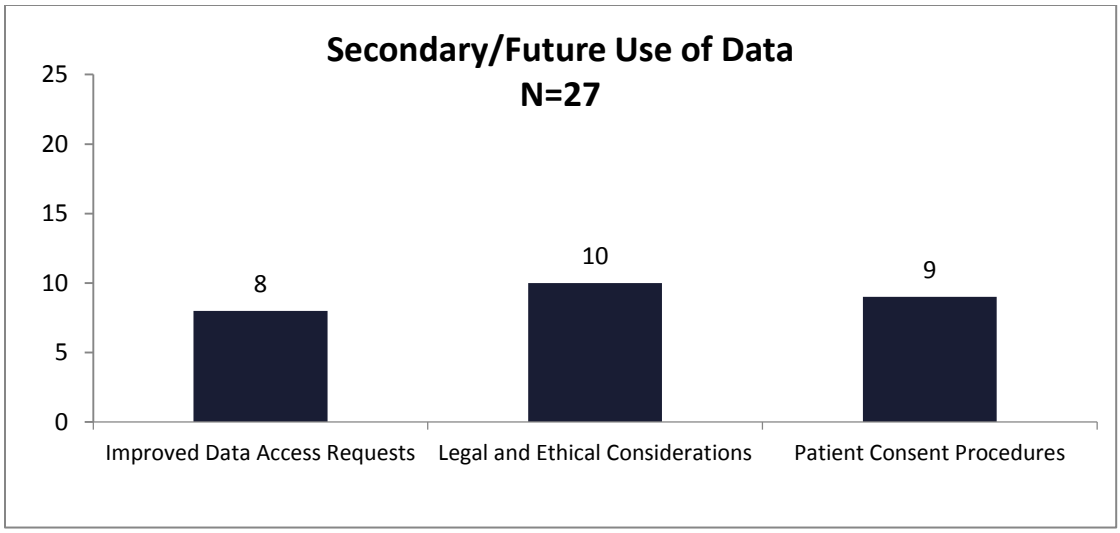
> *Data sharing needs to be built into the research and publication workflow — and not treated as a supplemental activity to be performed after the research project has been largely completed. Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances. Data relevant to public policy should be shared as quickly and widely as possible. (#46)*

All commenters in this category declared that the development of standards and guidelines and policies for data management, access, and sharing, was of critical importance for organizing and utilizing large biomedical datasets.

Issue Three: Secondary/Future Use of Data

Respondents' main suggestion regarding facilitation of the use of data through secondary sources of data was to create commonly-defined data fields with specific structure and standard definitions for methodologies. One respondent spoke to the possible role of the librarian in assisting with building an infrastructure.

> *Again, AAHSL and MLA maintain that librarians have the skills and expertise to assist researchers in understanding the necessity for, and applying the criteria for data definitions so that it can be shared in the future. Librarians can play an important role from the early planning of research proposals to the implementation of data management once a project is funded and should be part of the research team. (#29)*



Others believed that in order to support data for secondary and future use, guidelines and policies would need to be developed to address improvements in data access requests, legal and ethical issues, and patient consent procedures.

Improved Data Access Requests

Several respondents identified the Institutional Review Board (IRB) as a means for improving the efficiency of the request for access to data. In general, respondents felt that IRBs lacked clear guidelines, took a long time to provide approvals back to investigators and project managers, and slowed down the pace of research. The question was posed by a few respondents, "how do we protect privacy without imposing on the pace of many phases in research?" Changes to IRB policies and procedures could improve data access requests.

Legal and Ethical Considerations

Respondents noted that legal and ethical issues complicated data sharing and they relayed concerns that the development of guidelines and regulations for legal and ethical considerations was necessary. In particular, some respondents wanted to ensure that access to secondary data would continue to be free of charge to avoid an unfair barrier for researchers with less funding.

> *To facilitate the discovery process through secondary analyses and data repurposing, database access is optimally free of charge to authorized investigators, regardless of location or primary discipline, with costs of data management and curation underwritten by each e-infrastructure funding source(s) (mostly, NIH), at realistically sufficient levels of funding support. Fee-for-access, even by a sliding scale arrangement, encumbers discovery science by limiting it to the financially privileged. Establishing and maintaining a level playing field in access, scientific community-wide, is thus vital to the data informatics or e-structure enterprise. (#27)*

Developing a framework for determining ownership of data from publically-funded projects was cited as necessary to reduce duplicative claims of ownership by investigators and institutions. Policies of global health agencies and the Bill and Melinda Gates Foundation were cited as exemplars that reflect the key principles that should be included in such a framework.

> *The Bill & Melinda Gates Foundation identified eight principles: promotion of the common good, respect, accountability, stewardship, proportionality, and reciprocity. In a joint statement, global health agencies proposed that data sharing should be equitable, ethical and efficient. Most of these principles call for: 1) a recognition or reward structure for data collection efforts, 2) responsibility in data use that safeguards privacy of individuals and dignity of communities and 3) the use of data to advance to public good. (#31)*

Respondents highlighted the need for data security, especially with respect to information released through secondary sources or presumed for future use. Appropriate privacy protection must be guaranteed and considered as part of the original design of the data sharing and management system. One respondent referenced the Genome-Wide Association Studies (GWAS) results "that restricted info can be obtained by asking the right questions about data." (#35)

Patient Consent Procedures

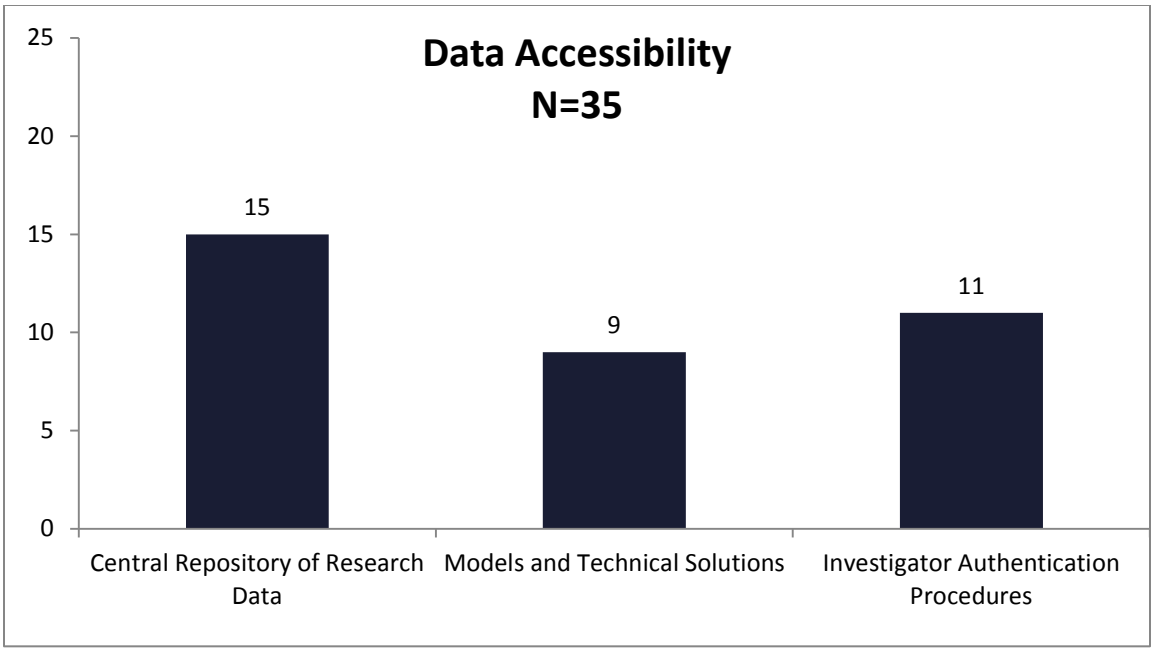
Many respondents believed that the current patient consent procedures are inefficient. One respondent reflected on how the consent process is impeded because there is no clear directive on who owns patient/human subject data.

> *Further, the extent to which data could be shared is constrained by questions of ownership of the data. Funders may feel that taxpayers supported the creation of study-specific data, so that NIH would own the data on behalf of taxpayers. However, in cases where researchers work at health care organizations and build datasets based on the organizations' data, the parent company may reasonably argue that they own the data and that NIH's contribution was a modest value-add. Health care organizations will have a need to shelter their data to protect their business from competition and from reputational risk and a duty to safeguard the confidentiality of their patients. Scientific investigators also have a stake in the ownership of the research data; since they invested their knowledge – including knowledge acquired outside of the study-specific work. (#23)*

Comments from other respondents ranged from promotion of an open-ended policy that would allow patients to designate that their data could be used in an unspecified manner to enactment of stricter access policies with governance and oversight (such as a Data Sharing and Publication Committee to control a HIPAA-compliant data system).

Issue Four: Data Accessibility

Most respondents had suggestions about how NIH could provide guidelines and regulations to assist with making data more accessible. One commenter suggested employing the same methods as the journal *Nature,* including requiring the full disclosure of all materials. Another commenter suggested the use of a distributed-computing paradigm or computing "cloud."



Central Repository of Research Data

Many respondents suggested that a central repository of research data should be developed and handled by NIH. One respondent believed that NIH should work with "university libraries, disciplinary societies, research consortia, and other stakeholders to distribute the many responsibilities associated with establishing and maintaining a trusted repository for digital data" (#15). Others remarked on the financial burden that repositories pose for institutions and emphasized how vital it was for NIH to play a key role to help reduce some of the cost burden.

Respondents acknowledged that there are many existing data repositories and they called for a "directory" of repositories to identify existing datasets. Such a central indexing repository would include links to other repositories, which would help increase access to data. However, respondents recognized that "this is a tremendous undertaking and many datasets that are not federally funded may be excluded from such an approach" (#29). Many suggested that NIH should fund or maintain such repository aggregators.

> *Making public data more visible, navigable, and useful can be accomplished by financing repository aggregators…Financing more projects and tools that promote domain specific databases to push and pull their data to the aggregators and to the Semantic Web will support data sharing. (#49)*

Models and Technical Solutions

One respondent indicated that computational models should be designed to answer specific questions and not for a general purpose. Another respondent called for NIH support so that tools to share data across sites could be streamlined. One comment mentioned the need to develop specialized tools that will provide assistance with "the use and understanding of common data elements and promote open architecture to enable software development for data mining" (#27). These tools will help in data

exploration by alleviating limited usability of a database. A commenter reported that building an infrastructure to query several repositories would add value because new discoveries rely on putting together different pieces of information.

Investigator Authentication Procedures

The comments on this sub-issue identified comprehensive procedures that authenticated the data provided was the investigator's own work. One respondent suggested that NIH create a digital author identifier which would provide a digital signature broadly recognized by datasets.

> *Digital Object Identifiers (DOIs) seem to be the best scheme today. Provenance requires that disambiguated authors be assigned to these datasets and as of today no widely accepted scheme exists to provide this identification. (#17)*

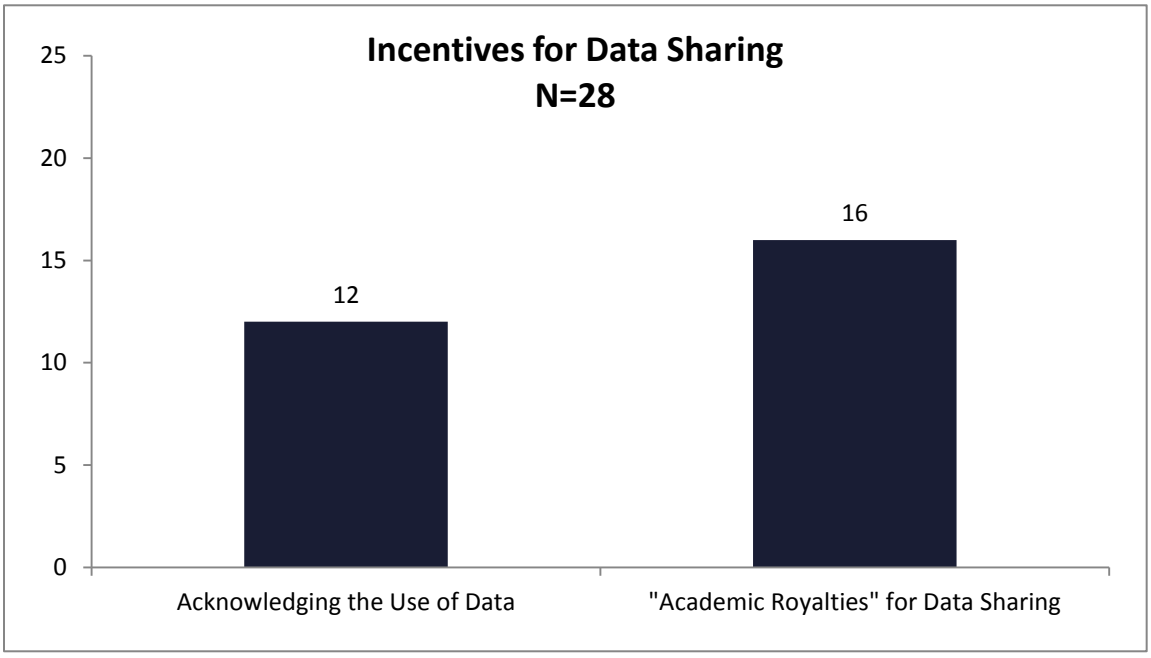
Other suggested procedures included the use of social networking tools for investigators to create a catalog and the protection of rights to the use of intellectual property by investigators.

Issue Five: Incentives for Data Sharing

Respondents either agreed that NIH **promote** policies and incentives to encourage data sharing or that NIH **require** data sharing.

> *The NIH should promote data sharing policies and incentives that will encourage data sharing. Without such incentives, researchers may see data sharing as an overhead activity, requiring time and effort with little reward. (#28)*

> *The NIH must become less passive with regard to enforcing data sharing by its grantees. If grantees are spending federal research dollars, it is incumbent upon them to preserve the research that these dollars purchase. (#38)*



Acknowledging the Use of Data

Developing standards, policies, and practices for acknowledging the use of data was deemed important by respondents, especially since many commented that researchers do the "bare minimum" to satisfy journal and publication requirements. One respondent stated,

> *There should be incentives for researchers to provide consistent and detailed meta-data annotation to the experimental data they are submitting. Special credit should be given during funding decisions to scientists who not only publish good papers, but also whose data are used by many other people.* (#13)

One respondent suggested that cultural differences play a role in the unwillingness to share data because of the fear of being "scooped." Creating clear incentives for data sharing could combat this fear. Specifically, developing a widely-accepted way to identify the creator of a dataset (such as the use of unique identifiers) would enable tracking of the impact and usefulness of data, as well as provide an easy way to reference data as part of an author's publication record.

"Academic Royalties" for Data Sharing

Most examples of incentives for "academic royalties" were provisions for special considerations in funding decisions. One respondent suggested a sixth scored review criterion for research awards entitled "data sharing track record" to include:

> *1) the number of publications that re-used the data from your lab and you serve as a coauthor of the papers; 2) the number of publications that re-used the data from your lab and you are not a coauthor of the papers. (#10)*

Another respondent believed that "the incentive to share data for the public good for individual investigators and their institutions will be outweighed by the incentive for personal (and institutional) gain." While this public good versus personal gain theory was seen as a barrier, the respondent thought that an international system may help.

> *An international registration system of collected data in health sciences or publication of datasets after peer review would provide opportunities for considerations of data collection and sharing practices during manuscript or grant reviews and could form an additional basis for promotion and tenure. (#31)*

Respondents shared concerns about unintended consequences of increased data sharing.

> *More significantly perhaps, it is not in the interest of the community if publicly-funded shared data favors researchers with loose ethical standards by granting them exclusive access to a valuable resource. NIH should establish and enforce guidelines to ensure that incentives for data sharing do not compromise existing standards in the scientific community, such as for example standards of academic authorship... (#37)*

> *Policies that support new indicators (e.g., bibliometric measures other than first or senior authored publications) of individual contributions to collective work need to be developed. Further, the federal funding data deposition policy, although requiring data deposition as part of publication, does not yet have a method to track the use of the dataset, nor a dedicated resource for sustaining access to the dataset after deposition. A system for dataset tracking and acknowledgement along with inclusion of metadata and provenance is needed. Such a system would give researchers a rich resource to evaluate*

for extant datasets BEFORE starting experiments of their own, therefore avoiding duplication of efforts and wasted research resources (money and time). (#43)

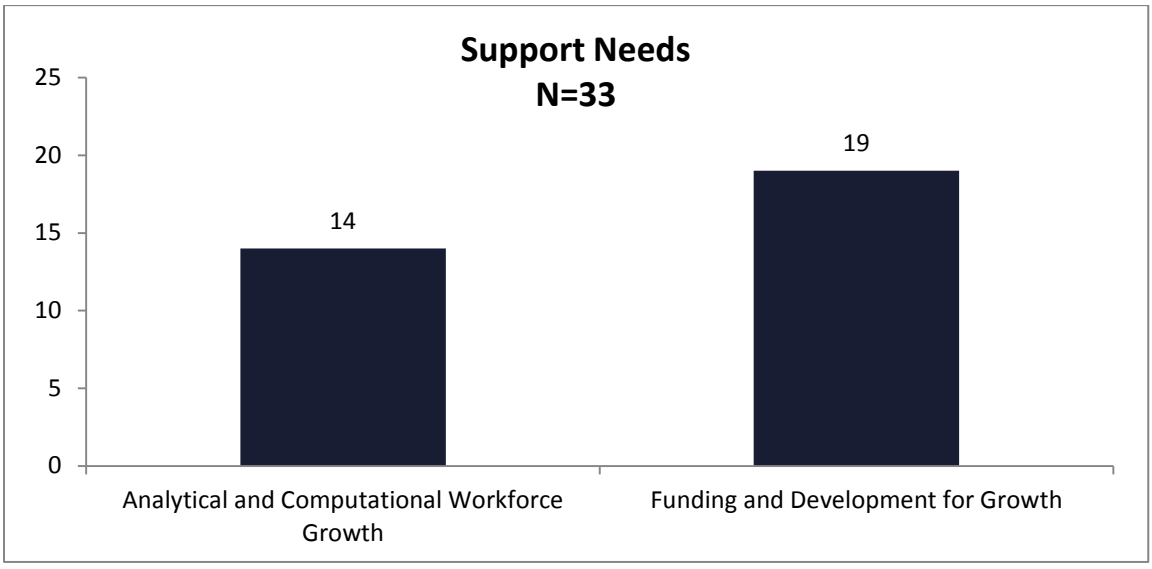
Issue Six: Support Needs

This issue targeted the role of NIH in providing resources to support the needs of the extramural community. Respondents stated that NIH would provide this support through workforce growth or funding and development opportunities.

Analytical and Computational Workforce Growth

Respondents addressed ways in which guidelines, training, and education could meet recent growth in the analytical and computational workforce. Suggestions spanned four topics:

- Need for trained specialists
  Many respondents commented on the lack of biostatisticians and bioinformaticians. Suggestions to increase the workforce included training individuals in data collection, formatting, algorithms, design, programming, and integration, as well as to make the career more attractive.
- Undervaluing of current professionals
  Another point made by respondents was the undervaluing of workers: "professionals supporting data and the infrastructure to make that data available need to be recognized and suitably supported." (#17)
- Development of training programs
  To support an increase of trained individuals in the data information systems workforce, curriculum development will play a major role and should include approaches to data annotation and storage.
- Establishment of data management tools
  Respondents shared their need for help in managing duties and resources; they believed that new management tools would be beneficial in this regard.



Funding and Development for Growth

Comments included the desire both for new programs that support technological developments and additional grants for methodologies and tools to maintain evolving software systems. One respondent wanted tools developed to quickly search and access relevant data. Another felt that tools were available but their values were unknown; therefore, standards to measure the value of tools needed to be developed. In regard to developing new methodology and tools for software efforts, respondents argued for increased funding from NIH. One commenter articulated this response more fully, concentrating on the fact that currently no one has taken on the responsibility of incurring the cost.

> *You have identified issues related to these questions, but the reality is that, at present, no funding agency has the responsibility and resources to do the very real, detailed work needed to create an agreed common physical and software infrastructure for practical long-term management and archiving of the data flows we are now seeing, much less the data flows that are coming soon. (#25)*

Other concerns that arose were the development of lab notebook software, filling of missing repository gaps, and international cooperation.

SECTION TWO: PRIORITY ISSUES

Respondents generally recognized the challenges inherent with managing large datasets. While it was rare for respondents to rank the order of the issues and sub-issues they identified as priorities, some provided a short paragraph or two identifying the issues they felt were most important.

To give a perspective on how many people identified which issues and sub-issues were a priority, we have presented the priority data from the individual perspective (as opposed to code application frequencies, which represent the total number of comments that received a particular code). Of the 50 respondents who provided feedback to this RFI, 36 (72%) identified at least one priority sub-issue.

Priority of Issues

The distribution of the top three issues based on priority criteria matches the distribution of the top three issues found in the overall comment analysis: Scope of Challenges, Standards Development, and Data Accessibility. However, in the priority analysis, the final three issues were Incentives for Data Sharing, Support Needs, and Secondary / Future Use of Data.

Order of Priority by Issue	Number of Respondents (N=36)
Scope of Challenges	24
Standards Development	24
Data Accessibility	16
Incentives for Data Sharing	14
Support Needs	12
Secondary / Future Use of Data	7

Priority of Sub-Issues

A summary of the top ten sub-issues is provided below for overall respondents and self-reported affiliates; a complete list of prioritized sub-issues is provided in Appendix C. Priority order was established based on the total number of respondents that expressed priority for each sub-issue.

Priority of Sub-Issues: Overall

Of the sub-issues, the greatest single priority was placed on Collaborative / Community-Led Standards, followed equally by Central Repository of Research Data, and Academic Royalties for Data Sharing. The sub-issues rounding out the top ten are shown in the table below.

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	10	1
Data Accessibility	Central Repository of Research Data	9	2
Incentives for Data Sharing	Academic Royalties for Data Sharing	9	3
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	8	4
Standards Development	Metadata Quality Control	8	5
Support Needs	Analytical and Computational Workforce Growth	6	6
Support Needs	Funding and Development for Growth	6	7
Scope of Challenges/Issues	Challenges/Issues Faced	5	8
Scope of Challenges/Issues	Unrealized Research Benefit	5	9
Incentives for Data Sharing	Acknowledging the Use of Data	5	10

*N=Number of Respondents

Priority of Sub-Issues: Self

Those who reported from their own individual perspectives expressed greatest priority for Collaborative/Community-based Standards and “Academic Royalties” for Data Sharing. Metadata Quality Control, Central Repositories for Research Data, and Feasibility of Concrete Recommendation for NIH complete the top five priorities for individuals.

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	7	1
Incentives for Data Sharing	"Academic Royalties" for Data Sharing	7	2
Standards Development	Metadata Quality Control	6	3
Data Accessibility	Central Repository of Research Data	6	4
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	5	5
Scope of Challenges/Issues	Challenges/Issues Faced	4	6
Incentives for Data Sharing	Acknowledging the Use of Data	4	7
Support Needs	Funding and Development for Growth	4	8
Support Needs	Analytical and Computational Workforce Growth	3	9

Issue	Sub-Issue	N*	Priority
Data Accessibility	Investigator Authentication Procedures	3	10

*N=Number of Respondents

Individuals who provided feedback from an organizational perspective offered limited comments with regard to prioritizing the issues and, therefore, the analyzed priorities are not presented.

SECTION THREE: RESPONDENT RECOMMENDATIONS FOR NIH

Our analysis for this section involved two approaches. The first approach was to compare code frequency distributions across the entire dataset with the subset of data created to represent specific ideas for NIH. The second approach involved qualitative analysis of the subset of data to identify common themes across respondent suggestions.

Code Frequency Comparison

Comparing the distribution of issues between the total dataset and the subset of NIH Responsibility revealed many differences. The order of frequency distribution for most of the issues differed except for the least identified issue (Secondary/Future Use of Data). The table below illustrates the overall order of frequencies for both subsets.

NIH Responsibility Subset	Total Dataset
Support Needs	Scope of Challenges/Issues
Incentives for Data Sharing	Standards Development
Scope of Challenges/Issues	Data Accessibility
Data Accessibility	Support Needs
Standards Development	Incentives for Data Sharing
Secondary/Future Use of Data	Secondary/Future Use of Data

Qualitative Themes

A number of specific suggestions were presented throughout Section One; in this section, we analyze the subset of NIH Responsibility data to present a more holistic view of respondent recommendations. The recommendations were at the issue and sub-issue level. The table below shows the number of codes marked NIH responsibility according to issues and sub-issues.

Issues and Sub-Issues	N*
Scope of Challenges/Issues	20
Research Information Lifecycle	1
Challenges/Issues Faced	2
Tractability with Current Technology	0
Unrealized Research Benefit	2

Issues and Sub-Issues	N*
Feasibility of Concrete Recommendations for NIH	15
Standards Development	18
Reduction of Storing Redundant Data	1
Standards according to Data Type	4
Metadata Quality Control^	4
Collaborative/Community-based Standards^	7
Develop Guidelines^	2
Secondary/Future Use of Data	8
Improved Data Access Requests	3
Legal and Ethical Considerations	2
Patient Consent Procedures	3
Data Accessibility	18
Central Repository of Research Data	11
Models and Technical Solutions	2
Investigator Authentication Procedures	5
Incentives for Data Sharing	23
Acknowledging the Use of Data	7
"Academic Royalties" for Data Sharing	16
Support Needs	36
Analytical and Computational Workforce Growth	14
Funding and Development for Growth	19

*N=Number of codes marked NIH responsibility

Support Needs

To adequately address data and informatics challenges, respondents made several suggestions that NIH support not only an infrastructure, but also output and utilization of data needs, such as enhanced organization, personnel development, and increased funding for tool maintenance.

Increase Funding to Develop and Maintain Data Applications

Comments included suggestions for investing in the development and maintenance of tools. For example, there was interest in new projects that created data repositories. One respondent claimed that NIH supported certain sub-types of data more than others (i.e., genomic/transcription over biological/biochemical). Similarly, others requested less emphasis on translational goals and more on basic science. The creation of an up-to-date directory describing databases and tool development projects was also recommended.

Specific comments were to increase funding for tool development in the areas of technology transfer, data capture, standards compliance, and data integration. Software for lab notebooks that would be freely accessible and available from NIH was suggested (often-cited tasks that the software must accomplish included assisting with documenting lab work, allowing links to figures, storing raw data in several file formats, and providing storage locations). Referring to the issue of exponential data growth,

one commenter requested that NIH not only invest in hardware, but also invest in algorithms and techniques. With the emergence of these new tools, respondents asked for curriculum materials to develop improved understanding of data annotations and storage.

Increase Professionals in the Field/Workforce Growth

Respondents urged NIH to fund projects and programs that placed more bioinformaticians and statisticians in the workforce. Some respondents requested resources for hiring and retaining technically-trained personnel. One comment suggested fellowships that would develop the skills of the workforce that already existed in most institutions, such as librarians.

> *In partnership with computational bio-informaticists and statisticians, librarians undertaking additional training opportunities can address data stewardship principles and practices including: data archival methods; metadata creation and usage; and awareness of storage, statistical analysis, archives and other available resources as part of a data stewardship training curriculum. (#29)*

While respondents called for an increase in funding to ensure growth in the workforce, the comments emphasized the need to "fund people and not projects."

Respondents also suggested support for data curation as a profession, stating that NIH should improve recognition programs for data curation and create alternative career paths. Respondents recommended that NIH stipulate guidelines for data curator positions to be filled by highly-qualified individuals with advanced degrees; these individuals would annotate datasets for high levels of accuracy and ensure data integrity.

Some respondents suggested NIH develop new training programs for data management and sharing. These programs would emphasize coherent strategies for the analysis of large datasets. One respondent suggested the need for new training programs in health agencies to prepare the next generation of investigators and public health staff with the mindset for data sharing.

Data Sharing

The second most cited area in which respondents made recommendations to NIH was in Data Sharing. Many comments suggested the need to make biomedical data more readily available and to address issues regarding the need for incentives to support data infrastructure.

Make Biomedical Data Available

Respondents suggested that NIH develop guidelines and standards. Specifically, they asked for guidelines around comprehensive patient consent procedures that would make data available. Respondents felt that the challenge lies in answering the question of who owns the data: researcher/scientist, institution, or government.

> *Funders may feel that taxpayers supported the creation of study-specific data, so that NIH would own the data on behalf of taxpayers. However, in cases where researchers work at health care organizations and build datasets based on the organizations' data, the parent company may*

> *reasonably argue that they own the data, and that NIH's contribution was a modest value-add. Health care organizations will have a need to shelter their data to protect their business from competition and from reputational risk and a duty to safeguard the confidentiality of their patients. Scientific investigators also have a stake in the ownership of the research data; since they invested their knowledge – including knowledge acquired outside of the study-specific work. (#23)*

One suggestion was to provide a place in the grant application to list shared data; another suggestion was that each researcher's data sharing record be evaluated in peer review. As described in Section One, one respondent suggested a sixth scored review criterion on the data sharing track record.

Respondents indicated the importance of engaging in shared policies and guidelines to determine best practices and systems for data citation. To address this need, respondents recommended accelerating the production of guidelines for researchers to ensure best practices. In line with this suggestion, one respondent was concerned with the ethical compromises inherent when guidelines are not readily available or accessible and suggested that NIH endorse or provide a set uniform data use agreement (DUA).

Incentives to Support Infrastructure

Many respondents called for improved incentives that would help facilitate data sharing by establishing data generators or intramural infrastructure. One respondent thought NIH should promote data sharing; otherwise, investigators may see it as a thankless overhead activity.

> *Without such incentives, researchers may see data sharing as an overhead activity, requiring time and effort with little reward. This perception will not encourage development of high-quality metadata...Better incentives for sharing data, standards for describing data, and clarity of policies for secondary/future use of data are all vitally important to making contribution and reuse of high-quality data a more achievable goal. (#28)*

NIH was encouraged to promote investigator compliance by rewarding and recognizing datasets as research output, thereby ensuring that data sources are included in applications for funding.

Furthermore, some respondents believed that NIH had become less rigid in enforcing data sharing. One respondent recommended that NIH "require" data sharing, not just ask or suggest that it occur.

> *The current policies in NIH RFAs and PAs only ask that applications describe plans to share data and software products created by their publicly funded research. They do not (at least in the cases I have seen) actually require funded projects to share their data. It would not seem unreasonable for NIH to require that projects share data in standard (or at least commonly-accepted formats), especially if those formats were developed thanks to NIH funding in the first place. (#36)*

Standards Development and Data Accessibility

Respondents thought that standardization and data housing would be most efficiently handled by a central source (which was often suggested as NIH). One respondent recommended the development of consortia for each subject, allowing researchers to make decisions specific to the discipline. The Beta

Cell Biology Consortium was used as an example where selection of cell type, treatment, and antibody is well discussed.

Standards regarding regulatory policies and procedures were recommended in an effort to advance the strong need for archiving data by developing technical solutions and standard templates for data sharing. Others suggested the need for appropriate digital signatures, such as digital object identifiers (DOI).

The issue of consistency was frequently mentioned. Some respondents proposed that repository requirements should establish minimal service criteria to be met by repositories as a method of unifying and preserving the data repository. Those who identified this issue as important suggested support to navigate and maintain the repositories since there are many repositories available for different types of data.

> *The researcher must know about all the different repositories in order to search for what they need, and the number of such repositories is only growing…making public data more visible, navigable, and useful can be accomplished by financing repository aggregators. Financing more projects and tools that promote domain specific databases to push and pull their data to the aggregators and to the Semantic Web will support data sharing. (#49)*

While most respondents believed that there were several repositories that met their needs, a few believed that some new repositories should be identified. One respondent suggested that a new database for RNAi data should be widely accepted. Another respondent recommended an alternative repository using the Supplementary Information (SI) section of research articles, thereby allowing publishers to commit to storing the data themselves.

Feasibility of Recommendations to NIH (Scope of Challenges/Issues)

One respondent felt that NIH had fallen behind in data informatics, recommending that NIH move to the cutting edge of the field to catch up with current developments. For example, The Cancer Genome Atlas was not able to release data online until the demise of CaBIG in 2011.

Respondents highlighted the dual problems of large amounts of data produced in many sites and the inadequacy of most systems to handle large volumes. Suggestions for solving these problems were organizing data, picking appropriate representatives, developing algorithms and managing interfaces, and designing systems that maximize transfer between disc and memory.

Others articulated the need for a site to host linked data, stating that their current systems compose a series of “patchworks of exception.”

Collaborative/Community-based Standards (Standards Development)

On the mind of several respondents was the need for NIH to facilitate collaborations for data and informatics topics. Increased collaboration and coordination were consistently identified as important for improving data sharing and data management issues. Respondents called for collaboration on a

variety of levels and emphasized the involvement of everyone, including agencies, institutions, and the U.S. and international scientific communities, in a discussion about the development of data standards.

Collaboration with NIH, Federal Agencies, and Institutions

In addition to NIH, respondents suggested partnerships with sister agencies and grantee institutions to develop approaches for supporting mid-level IT infrastructure as a way to meet agency needs and, in return, avoid inflicting operating inefficiencies on grantee institutions. One respondent highlighted ongoing collaborations to improve the grant making process by the Research Business Models Working Group of the National Science and Technology Council and Federal Demonstration Partnership. This suggestion was for NIH to work in conjunction with working groups in order to facilitate progress towards more developed and maintained IT infrastructure.

Respondents urged NIH to develop data standards to assist investigators who are less familiar in one research area in understanding and using datasets from another research area, thereby leveraging previously funded resources. To facilitate such standards, the National Library of Medicine was suggested to serve as a model for the extramural community in its reliance on the experience and expertise of its librarians.

> *Librarians can be essential team players, not only in helping to develop standards and ontologies, but also in making their research communities aware of the resources available through NIH and other research groups and agencies. (#29)*

The important question, "Who owns the dataset?," emerged from a few commenters. The respondents recommended that NIH, in consultation with researchers, clinicians, and patients, address this issue, giving sufficient weight to the common good.

Community Collaborations

Respondents believed NIH could promote effective coordination of standards by helping to identify problems that standards will solve. Creating initiatives on sharing through use of community reporting standards would encourage good data stewardship. Repeatedly, respondents suggested that NIH support community-initiated efforts for standardized data representation. One respondent used the example of The Gene Ontology to support the notion of collaboration.

> *The Gene Ontology was developed by multiple model organism database developers who saw the benefits of collaborating on a common standard. Its wide adoption demonstrates the success of data standards developed collaboratively by researchers trying to solve practical problems. (#26)*

One respondent recommended that NIH require some minimum amount of diversity analysis and reporting on data collected under diversity sampling requirements.

> *It is nonsensical that NIH requires, and goes to great pains to enforce, diversity in sampling; yet has no coincident requirement to conduct and report on differential validities due to race, gender, age, etc. Consequently, very little of this sort of research is ever conducted despite having sufficient data. (#3)*

Global Collaborations

Respondents believed NIH could use its considerable influence to promote and improve collaboration around the world. Respondents suggested that NIH coordinate support between the U.S., Europe, and Asia where uniform standards are often needed. One suggestion was for NIH to work with other funders, such as The Welcome Trust or Biotechnology and Biological Sciences Research Council (BBSRC), to establish consistent data policies where regional laws permit. The ultimate goal would be to make data interoperable, regardless of geographic origin or funding source.

Appendix

A. FULL CODING SCHEME: DESCRIPTION OF ISSUES AND SUB-ISSUES

Issue 1: Scope of the Challenges/Issues

Issue: Understanding the challenges and issues regarding the management, integration, and analysis of large biomedical datasets

Sub-Issue	Description
Research Information Lifecycle	Strategies for managing research information/data from the time it is created until it is terminated
Challenges/Issues Faced	Challenges and issues presented by use of datasets in the biomedical field
Tractability with Current Technology	Ability to manage and control current technology
Unrealized Research Benefit	Acknowledgement that datasets, data sharing, and administrative data have many research benefits that are not being explored
Feasibility of Concrete Recommendations for NIH	Recommendations for NIH action regarding biomedical data

Issue 2: Standards Development

Issue: The development of data standards

Sub-Issue	Description
Reduction of Redundant Data Storage	Identification of data standards, reference sets, and algorithms in order to reduce the storage of redundant data
Standards according to Data Type	Identification and differentiation of data sharing standards according to data type (e.g., phenotype, molecular profiling, imaging, raw versus derived, etc.)
*Metadata Quality Control^*	Development of standardized metadata (uniform descriptions, indexing categories, semantics, ontologies, formats, etc.) to organize data from different sources and improve data quality control
*Collaborative/Community-based Standards^*	Development of processes that involves community-led collaborative efforts
*General Guidelines^*	Development of guidelines for data management, access and sharing, and training for compliance

^Data-driven issues

Issue 3: Secondary/Future Use of Data

Issue: The facilitation of the use of data through secondary sources or data presumed for future use

Sub-Issue	Description
Improved Data Access Requests	Development of procedures and policies that will improve the efficiency of the request for access to data (e.g., guidelines for IRB)
Legal and Ethical Considerations	Development of evolving guidelines and regulations for legal and ethical considerations
Patient Consent Procedures	Development of comprehensive procedures and policies regarding patient consent to share their information

Issue 4: Data Accessibility

Issue: The ability to access data

Sub-Issue	Description
Central Repository of Research Data	Development of a central repository of research data appendices (e.g., developing links to PubMed publications and RePorter project record)
Models and Technical Solutions	Development models and technical solutions from multiple heterogeneous data sources
Investigator Authentication Procedures	Development of comprehensive procedures that authenticate the data provided are the investigator's own work

Issue 5: Incentives for Sharing Data

Issue: The need to have incentives in order to encourage/influence others to participate in data sharing

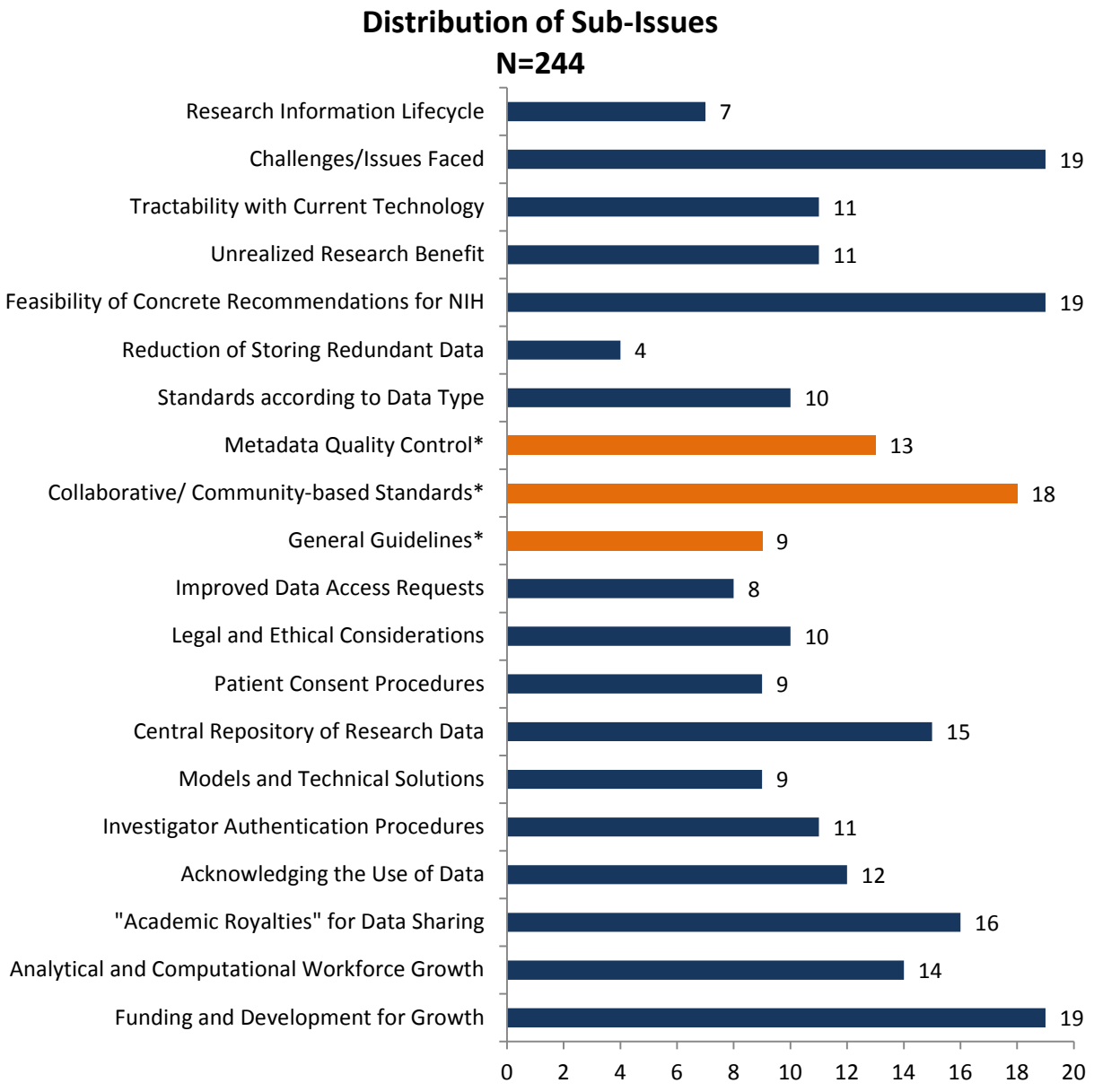
Sub-Issue	Description
Acknowledging the Use of Data	Development of standards/policies for acknowledging the use of data in publications
"Academic Royalties" for Data Sharing	Creation of policies for providing "academic royalties" for data sharing (e.g., special consideration during grand review)

Issue 6: Support Needs

Issue: The role of NIH to provide supportive needs to the extramural community

Sub-Issue	Description
Analytical and Computational Workforce Growth	Provision of guidelines, training, and education to facilitate growth in the analytical and computation workforce
Funding and Development for Growth	Provision of funding and development for tools, maintenance and support, and algorithms

B. SUMMARY OF FREQUENCY DISTRIBUTION ACROSS ALL SUB-ISSUES



C. ORDER OF PRIORITY: ALL SUB-ISSUES

Order of Priority: Overall (N=36)

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	10	1
Data Accessibility	Central Repository of Research Data	9	2
Incentives for Data Sharing	"Academic Royalties" for Data Sharing	9	3
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	8	4
Standards Development	Metadata Quality Control	8	5
Support Needs	Analytical and Computational Workforce Growth	6	6
Support Needs	Funding and Development for Growth	6	7
Scope of Challenges/Issues	Challenges/Issues Faced	5	8
Scope of Challenges/Issues	Unrealized Research Benefit	5	9
Incentives for Data Sharing	Acknowledging the Use of Data	5	10
Standards Development	General Guidelines	4	11
Secondary/Future Use of Data	Improved Data Access Requests	4	12
Data Accessibility	Investigator Authentication Procedures	4	13
Scope of Challenges/Issues	Research Information Lifecycle	3	14
Scope of Challenges/Issues	Tractability with Current Technology	3	15
Secondary/Future Use of Data	Legal and Ethical Considerations	3	16
Data Accessibility	Models and Technical Solutions	3	17
Standards Development	Reduction of Storing Redundant Data	1	18
Standards Development	Standards according to Data Type	1	19
Secondary/Future Use of Data	Patient Consent Procedures	0	20

*N=Number of Respondents

Order of Priority: Self (N=26)

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	7	1
Incentives for Data Sharing	"Academic Royalties" for Data Sharing	7	2
Standards Development	Metadata Quality Control	6	3
Data Accessibility	Central Repository of Research Data	6	4
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	5	5
Scope of Challenges/Issues	Challenges/Issues Faced	4	6
Incentives for Data Sharing	Acknowledging the Use of Data	4	7
Support Needs	Funding and Development for Growth	4	8
Support Needs	Analytical and Computational Workforce Growth	3	9
Data Accessibility	Investigator Authentication Procedures	3	10
Scope of Challenges/Issues	Tractability with Current Technology	2	11
Scope of Challenges/Issues	Unrealized Research Benefit	2	12
Standards Development	General Guidelines	2	13
Secondary/Future Data Uses	Improved Data Access Requests	2	14
Data Accessibility	Models and Technical Solutions	2	15
Scope of Challenges/Issues	Research Information Lifecycle	1	16
Standards Development	Standards according to Data Type	1	17
Secondary/Future Data Uses	Legal and Ethical Considerations	1	18
Standards Development	Reduction of Storing Redundant Data	0	19
Secondary/Future Data Uses	Patient Consent Procedures	0	20

*N=Number of Respondents