The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Forecasting Costs of Biomedical Data Preservation

A User Guide for Funding Organizations

Summary

Biomedical researchers are generating, collecting, and storing more research data than ever. Preserving those data in discoverable and accessible ways is increasingly important, though doing so generates costs that may be difficult to predict. Allocating responsibility for such costs may further complicate a research endeavor. This guide will help funding organizations identify and consider the major decisions and recommendations for forecasting life cycle costs for preserving, archiving, and promoting access to biomedical data. The guidance presented here reflects the in-depth analysis of the following report from the National Academies of Science, Engineering, and Medicine: *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*,[1] as well as some of the discussions from the following workshop: *Planning for Long-Term Use of Biomedical Data: Proceedings of a Workshop*.[2]

Background

The costs of constructing, maintaining, and accessing biomedical data can vary widely. The National Library of Medicine of the National Institutes of Health tasked the National Academies of Sciences, Engineering, and Medicine with developing a framework for forecasting long-term costs for preserving, archiving, and accessing various types of biomedical data and estimating potential future benefits to research. The resulting National Academies report highlights major cost drivers for biomedical research information resources and puts forth steps for individuals, institutions, and organizations to consider the life cycle costs associated with the data. This user guide summarizes several ways in which biomedical information resources may vary and how each variation is likely to affect costs or utility. It also identifies key areas where funding organizations can contribute to the successful implementation of the framework.

[1] 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.

[2] National Academies of Sciences, Engineering, and Medicine. 2020. *Planning for Long-Term Use of Biomedical Data: Proceedings of a Workshop.* ed. Linda Casola Washington, DC: The National Academies Press. https://doi.org/10.17226/25707.

The life cycle of digital data typically involves the following three major states:

- **State 1: The Primary Research and Data Management Environment**
  In State 1, data are actively captured as they are created, and then analyzed. Those managing or using a State 1 data environment should be focused on standardizing, documenting, sharing, and preserving data and algorithms.

- **State 2: An Active Repository and Platform**
  In State 2, data may be acquired, curated, aggregated, accessed, and analyzed. This is an active information system that usually provides services to a wide range of users. Data are acquired from the primary research environment, from another active repository, or may be revived from archival storage for active use.

- **State 3: A Long-term Preservation Platform**
  In State 3, content is preserved across changes in governance, assessment of data value, and technology. The platform may include an extract of data from a single data set, multiple data sets, or an information system in a system-agnostic format. In this state, data are neither directly analyzable nor easily accessible. Content (e.g., data and code) are preserved in a long-term preservation platform when it is anticipated that the data will not be actively used for the foreseeable future, or if the resources are not available to maintain an active repository.

Data take different forms in each state, and each state includes different activities with different personnel, hardware, and management requirements. It is important to note that the labor and computation needed to transform data from one state to another can require significant resources and data may not transition through the three states sequentially because of the unique needs of the research endeavor or repository.

Funders should note that the biomedical repository landscape spans accessible data repositories hosted by government agencies, national laboratories, research consortia, institutions and hospitals, patient advocacy organizations, researchers, journals, and commercial entities, including consortia of study sponsors.

Value of Data

The perceived value of data influences preservation, access, and archiving decisions as well as decisions made regarding transition of data from state to state. However, assessing the value of biomedical data is challenging and needs to extend beyond monetary costs.

The value of a single data set reflects factors such as its uniqueness, the number of times it is used, the cost per use, and the impact of reuse. The number of different tasks or decisions that the data support may be a good indicator of their value. Data valuation can also depend on the data being findable, accessible, interoperable, and reusable (FAIR), and data standardization and documentation play an important role.

Furthermore, the storage decisions for data can affect their value. For example, while an individual data set on its own may be of limited value, when aggregated with other data, it can potentially increase the value of the entire pool and enable the generation of new knowledge. Thus, data can be "multiplicatively integrative" through adherence to the FAIR principles and exposure through platforms that make them FAIR. If, however, data are shared through a platform where their discoverability is limited and where standards and curation are not enforced, then their value may diminish.

Cost Forecasting

The cost of preserving and providing access to data depends on choices made throughout the data life cycle and on the presence of tools, institutional support, and incentives that affect those choices. These choices often predate the launch of an individual research project in which data are generated. Funder requirements, data management mandates, institutional review board specifications, federal regulations, and journal requirements can all influence costs across the data life cycle. Data management plans that incorporate costs and value across the data life cycle may reduce the cost and time required for later data deposit and sharing.

Cost forecasters will likely need to consult with multiple individuals with varied expertise to minimize uncertainty in the forecast. In most cases, the cost of long-term data preservation will not be accrued by a single individual or institution; the cost burden can shift over the course of the data lifecycle. Understanding where costs will be accrued and who has managerial responsibility for them will inform decision-makers for all data states.

Forecasting Framework

The framework presented should be considered the basis of a cost forecast rather than a one-size-fits-all analytical tool for all applications. How it is applied in any situation depends on the circumstances, needs, and resources available to those involved. The activities, decisions, and cost drivers will be situationally dependent, and the framework will need to be modified to suit the specific purpose. In whatever application, however, the forecaster is encouraged to think beyond the costs associated with the specific data state being developed or managed. In the long term, it is more efficient to think early about how decisions may affect the costs of data

management and access in future data states, the transitions to those states, and to the future value of data to the scientific enterprise.

The report provides an overview of steps to direct a cost forecaster's efforts throughout the process of forecasting data costs for a biomedical information resource. This framework is meant to be a starting point for cost forecasters to begin analyzing the costs of their biomedical data resource. Even so, it is often helpful to see an example first. The committee put together a video think-aloud that walks a user through the thought process and mechanics of the framework.

To identify data characteristics, data contributors, and data users, the cost forecaster will need to work with a range of experts within the research community to identify or develop appropriate metrics to better understand and manage costs. Consulting with information technology professionals, metadata librarians, software engineers, and many others may be necessary to compile the information necessary to identify the major cost drivers. The primary cost drivers often relate to the following:

- **Content** – the amount, kinds, and qualities of data that a biomedical information resource is expected to host. Generally, the larger and more complex a data set, the more costly it will be. Costs can be lowered by greater compressibility and replaceability.
- **Capabilities** – what information resource users are able to do with the data therein. More functionality and capabilities for a data resource typically means greater costs.
- **Control** – aspects of a biomedical information resource that deal with control and oversight of the resource (e.g., quality control measures). Increased controls on the data or the repository result in higher costs.
- **External Context** – relationships between the biomedical information resource and other, external resources. Although cost relationships can vary, costs typically increase if the resource is replicated and if its content is relatively distinct.
- **Data Life Cycle** – aspects of a biomedical information resource's expected evolution over time. Longer-term costs will be incurred if the resource is anticipated to be updated or grow in size. However, outlining a useful life span and moving the resource to offline or deep storage can reduce costs. It is important to keep in mind the trade-off in balancing the allocation of resources to maintain the by-products of past research, and allocating resources toward new research. Expending resources to keep existing data sets available means fewer resources for funding new research activities.
- **Contributors and Users** – a biomedical information resource's users and their characteristics. The wider the audience for a biomedical data resource, the more costly it will be.
- **Availability** – expectations about the availability of the data in a biomedical information resource. This can encompass the reliability of the resource hosting the data, how quickly new data appear, how fast requests for data are serviced, and from

where the data can be accessed. A resource which offers greater accessibility (both to the data and to user assistance) will have greater associated costs.

- **Confidentiality, Ownership, and Security** – data protection and the rights of those associated with the data. Taking measures to ensure higher confidentiality and security will increase costs. Costs may also be increased if multiple parties have ownership or rights to the data.
- **Maintenance and Operations** – obligations for maintenance and operation of the biomedical information resource. Frequent maintenance and more extensive risk mitigation efforts will drive up costs. Costs may or may not be offset by the possibility of charging for use of the resource.
- **Standards and Regulatory Compliance, and Other Governance Concerns** – community conventions, rules, policies, laws, and stakeholder concerns with which the operators of a biomedical information resource may have or want to comply. Greater oversight will incur greater costs, as will using more modern applicable standards.

Once the investigator has considered where the data are coming from, and how they will be used, he can begin to quantify the costs. To do so, he can use the data set characteristics, activities, and cost drivers described above and in the template in the cost driver workbook. Many of the activities and cost drivers in the template may not be directly applicable to every information resource, but the forecaster needs to remain aware of potential future cost drivers so that decisions might be made that could keep life cycle costs low.

Another important consideration is the reliability of a cost forecast. Placing greater emphasis on cost forecasting at the development of the data management plan and the award process does not mean that the forecasts will become more precise estimates. However, they could be considered accurate reflections of uncertainties. Quantitatively assessing uncertainties may be difficult, but decision-makers should expect cost forecasters to communicate concerns, even if they cannot be precisely characterized.

Decision-makers must also keep an eye towards emerging disruptors, which could radically change how research is conducted and data are collected, used, archived, or preserved. For example, the most costly resource required for making biomedical data useful for science is often human labor. A major challenge to the biomedical research community, both now and into the future, is the continual training and education of a workforce that can effectively process and manage data. Additionally, changes in legislation and policy related to data may present potential disruptions. Disruptors may be positive, negative, or mixed, and could raise or lower the cost of data management and preservation. There is no way to anticipate the impacts of potential disruptors, but building flexibility into data planning can help to mitigate their effects.

Community Next Steps

The report makes note of several infrastructure systems and services that support broad research activities, which are challenging for data generators and primary data collectors to attempt to incorporate into their cost forecasts. In particular, the organizations, governance, standards, systems, and common knowledge structures are often viewed as being within the purview of a research "community"; thus, funders may not want to support their solutions as part of an individual research project. Funding program managers may want to consider investing in standards, knowledge structures, and common tools that can help research communities. Funding bodies may want to consider how best to support all parts of this infrastructure, particularly operations and maintenance.

In addition to implementing the cost-forecasting framework, the following actions can help expand the capacity of data producers and managers to make sound management decisions and cost forecasts:

- **Explicitly recognize the value of active data resources (i.e., repositories) to the enhanced curation, discoverability, and use of data.**

  This recognition is absent among the funding entities, researchers, and institutions supporting research, most of which apply the more traditional data management approach of transitioning data directly from the primary research environment to long-term archiving. As mentioned above, data can be multiplicatively integrative when stored in the proper location, providing benefits beyond the original research. The biomedical research community would benefit from recognizing that the long-term benefits of properly supporting active data resources often outweigh the costs and short-term burdens of establishing the resource and preparing data for them.
- **Structure cost forecasts for active data resources around communities and research programs rather than individual research efforts.**

  Because active data resources serve communities of researchers, it may not be appropriate to allocate the costs of managing data in an active data resource back to the individual data contributor.
- **Support standardization efforts, including developing tools and methodologies to estimate the cost of standards development, encouraging the use of those tools and standards as part of the funding programs where appropriate, and explicitly supporting metadata preparation.**

  Data that do not comply with standards or that have not been documented with appropriate metadata are of lesser value, and grants are not structured to allow money to be "held aside" until standards are established. Even when standards exist, the current incentives for researchers to deposit data in useful formats are weak, and requirements to do so lack enforcement. Funding agencies can assist by contributing to

tools for estimating the cost of standards development and metadata preparation, by explicitly funding metadata preparation, and by issuing clarifying language about the use of federal funds to preserve data beyond the end of the grant.

- **Identify incentives, tools, and training for adopting good data management practices, including cost-forecasting practices, which facilitate sustainable long-term data preservation, curation, and access.**

  Researchers may be incentivized to more accurately account for the uncertainties associated with sharing data and future reuse if funders place a greater emphasis on such accounting in data management plans in grant proposals. Clear guidance for researchers is also necessary for data management plans to be meaningful. Incorporating better-directed guidance and training of individuals in data management, including undergraduate and graduate students, would be beneficial as well. Additionally, research on the normative outcomes of any increase in benefits resulting from improved data management skills could inform future training efforts. Such activities would benefit the entire biomedical research community, including the institutions and funding entities that support research. To strengthen these endeavors, funding entities need to better understand research-community needs, help the community to define desired outcomes, support training, develop realistic and actionable metrics for success, and provide near-term incentives for success.

- **Understand the charges associated with storage and computation in a data resource, regardless of who "pays the bill," when making decisions about data and workflows.**

  Researchers are often unaware of costs associated with data management in part because they typically are not responsible directly for those costs. Mechanisms are needed to inform researchers of the actual costs paid for the services rendered to them, even if they are not directly charged, and institutions supporting research might also encourage researchers to limit those costs. This is especially applicable to agencies that provide storage for the data generated by research they have funded.

Furthermore, pursuing the following activities could advance practices and drive future improvements in the ability to forecast costs:

- **Recognize explicitly that scientific data constitute an asset and that data stewardship requires support.**

  Data represent more than just the end product of research. Unlike physical infrastructure, biomedical research data and the resources that house them are assets that contribute to the delivery of good science and, ultimately, the public good. The universities and institutions that support or enable research and host data resources, in turn, benefit from the recognition of that support.

- **Systematically collect data on costs associated with the biomedical research data enterprise to allow the translation of the framework outlined in this report into resources and methodologies that would benefit individual researchers and repository institutions.**

  The true costs of preserving, archiving, and accessing biomedical research data need to be investigated in a systematic way at the level of the funding program manager rather than at the individual researcher or project level. Some federal agencies treat cost estimation as a profession and invest in training, recognizing success, critiquing failures, and encouraging assembly of cost-related data. The biomedical research data preservation enterprise has become an undertaking that warrants a similar cadre of cost analysts to augment domain expertise and expertise in data science.

- **Develop easier mechanisms for creating and maintaining data management plans, automatically incorporating data and metadata into resources, and improving citations for data to work together with other research products.**

  Data management plans are typically static documents prepared as a mandatory—but not necessarily influential—part of the funding process. Placing more emphasis on quantified cost forecasts during the development of the data management plan and the award process may be one way to incentivize early planning and communication, even if cost forecasts are uncertain. Cost forecasts and data management plans need to evolve and be updated as research progresses, and as associated data and the resources and technologies available to manage those data evolve. Monitored evolution of a data management plan might inform eligibility for future funding. By providing these mechanisms, funders and research institutions could help improve efficiency, return value for stakeholders, and increase the likelihood that stakeholders will make sound data-related decisions.

Lastly, the current system for funding research cannot accommodate data life cycle cost forecasting. Planning horizons are dictated by funding streams and thus extend only for the life of the project, excluding post-project data-preservation issues. Although some funding organizations may require researchers to have data management plans specified in their grant proposals, there are currently no requirements for how such plans are to be formulated.

Proper data preservation is a complex endeavor requiring dedicated resources over the long term. Many agencies have traditionally attached less importance to data preservation, and increased efforts on that front may require major adaptations within those agencies. Additional challenges are that the planning horizons for agencies may be tied to annual budget appropriations, and there may even be legal prohibitions against planning expenditures beyond the appropriation period.

It is increasingly important to develop the rules governing data life cycle cost forecasting and to educate the community about the value of implementing them. In doing so, cost forecasting can become an integral part of responsible conduct of research, as opposed to a bureaucratic chore. Figure 1 outlines important objectives for achieving such a paradigm shift. Although data management practices in the laboratory are at the front line of eventual data sharing and long-term data access, there is often a lack of incentive for researchers to think about long-term curation and preservation needs, as they do not recognize a personal benefit. The biggest limitation is the amount of time it takes, followed by lack of best practices, and lack of training. There may also be other motives that prevent researchers from sharing their data, which could become apparent and possibly better-understood if funders begin to enforce data-sharing policies. To the greatest extent possible, it should be made easier for researchers and other stakeholders to make good data-related decisions from the onset.

Implementing this cost forecasting framework into the research funding system and the broader research community will require a cultural shift, which needs to be driven by community engagement. Oversight entities are in an exceptional position to offer incentives for this change. However, the process must be led by researchers so as to better meet their needs and so that they can fully understand and agree to the value returned to them for their efforts. Ultimately, this will benefit the scientific enterprise as a whole, as well as individuals whose well-being biomedical research seeks to advance.

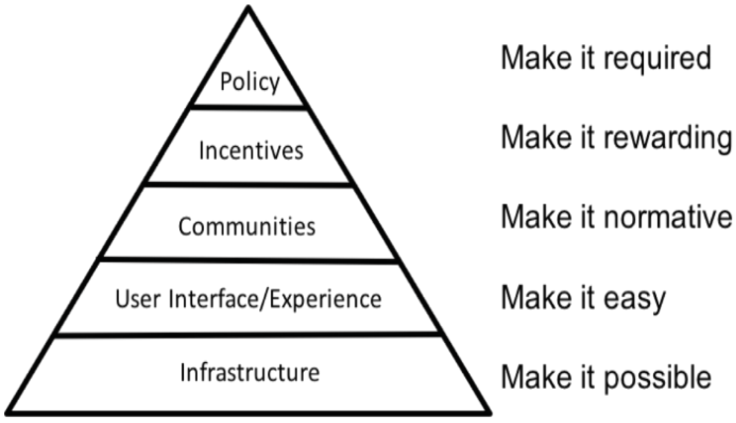

Figure 1 Key steps needed to change a research culture. SOURCE: Lucy Ofiesh, Center for Open Science, presentation to the workshop, July 12, 2019. Image available courtesy of CC-BY attribution license.