The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Forecasting Costs of Biomedical Data Preservation

A User Guide for Storage Providers

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 data storage entities 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]

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 repositories, repository planners, and program managers can contribute to the successful implementation of the framework, as well as reduce the opacity of data storage options and pricing.

The biomedical data landscape is diverse, distributed, and dynamic, characterized by an array of data repositories, databases, and platforms, which host data and make them available for reuse. These repositories are the database infrastructures where long-term stewardship, preservation, and access to research data are made possible. For the purposes of its report, the committee uses the terms “data repository” and “data archive” to refer to data infrastructure

[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.

that host primary research data rather than to refer to knowledge bases that extract and aggregate analyzed data from scientific literature.

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.

There will naturally be overlap in some activities in all the data states. However, the distinction between States 2 and 3 helps focus on the different issues that arise as one moves from facilitating active use to long-term retention. Drawing this boundary helps to ensure that decision-making processes consider the challenges of long-term data preservation and their associated costs.

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. The biomedical research data environment is dynamic; new data are

constantly being generated and new data infrastructures are continually being developed while older ones may migrate, merge, grow stale, or be taken down.

In the absence of specific requirements coming from the funder or journal, community data repositories specialized for a particular type of data can be helpful in promoting data sharing. Specialist repositories generally enforce community standards and have software available to help researchers comply with these standards. They also generally provide visualization and analysis tools that work with these specialized data types. Another storage option for data generators are institutional repositories, which are often hosted by research libraries and generally provide services for private management or public sharing of research data only for researchers within their home institutions. Many journals also allow authors to publish small data sets that live on the journal website as supplemental materials to their papers.

The variety of expertise and types of infrastructures and services required to work with diverse data make it unlikely that a single large data resource can serve all communities. Multiple archives and data repositories enable the creation of specialized tools and services as well as innovation in the ecosystem. If these repositories use the same standards, federated search across them becomes possible. Nevertheless, multiple repositories impose a cost in that separate infrastructures, staff, and tools must be maintained at each site and may, in some cases, result in less value, and less data discovery, than might otherwise accrue from a more unified resource, particularly if different standards and formats are imposed by different repositories.

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.

The value of a data resource compounds if it sparks connections among diverse users. This compound value reflects factors such as the distribution of user backgrounds, geographic origins, and purposes. In the long term, the greatest value may be realized through the multiplier effect as heterogeneous data sets are aggregated and linked on novel computational platforms in ways that are impossible to predict at the time a data set is created. Therefore, 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. 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. Collecting early data on what specific activities will be required in any of the data steps will help to improve estimates. Institutions may also need to consider changes in the composition of their direct labor forces. 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.

While repositories will have additional considerations for cost factors that are external to the data itself, the table on page 8 provides an overview of steps to direct a cost forecaster's efforts throughout the forecasting process. 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[2] 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 her host institution, project funders, and perhaps the broader research community to identify or develop appropriate metrics to better understand and manage costs. It may be especially helpful for forecasters outside of the biomedical data repository domain to consult with information technology professionals, metadata librarians, software engineers, and many others in order 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 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.

[2] To watch the video think-aloud of the cost drivers workbook, visit https://vimeo.com/444647256

- **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 data manager 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 cost driver workbook[3]. Many of the activities and cost drivers in the template may not be directly applicable to every active and long-term 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. While repositories need to understand the costs for which they must budget today, they should also be aware of the total cost of repository responsibilities, should they eventually have to cover additional lifetime data costs.

It may be especially difficult to forecast costs for long-term data preservation platforms. The forecaster may be making decisions about the format in which the data should be preserved and the nature of access to be supported years in advance of the actual transfer of data to a long-term preservation environment. Community guidance and standards may be helpful in making such decisions, with due allowances for future evolution. Above all, decisions should be documented, since they constitute the assumptions on which the forecast rests. If they are clearly stated, it will be easier to adjust for changed circumstances as the transition to long-term preservation nears.

The characteristics of the data sets are often important predictors of storage costs and information technology services; these will likely dominate the long-term forecast. In a sense, the investment in long-term data preservation could be viewed as an option on the future availability of the data set. While there is no market for such options, that intellectual construct could help guide decisions regarding the preservation platform. Enhanced preservation options could make a data set more discoverable or more easily reconstructed and used, therefore

[3] Access the cost driver workbook at https://www.nap.edu/resource/25639/Cost_Driver_Template_Word_0715.doc

increasing the data's potential value. Conversely, decisions that make data harder to discover, reconstruct, and use can negatively impact the preserved data's value.

Reliability of cost forecasts is a critical issue, especially for long-term preservation platforms with their high degrees of uncertainty. 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. While assessing uncertainties may be difficult, cost forecasters should communicate concerns, especially with users, even if they cannot be precisely characterized.

The lack of visibility regarding the true costs of data storage and access in individual laboratories, institutions, and community resources often hampers reliable cost forecasting. A price reflects the amount of money needed to purchase a product or service, but it may not always accurately reflect the true cost to provide those inputs. The issue of cost versus price is especially important to consider when projecting the cost of commercial services. Service providers may benefit from much greater economies of scale and thus lower cost than an individual institution or researcher, but their lower costs will not necessarily translate into lower prices for the science community. Even if prices accurately reflect past (marginal) costs, there is no guarantee that they will do so in the future.

When accounting for the costs of the biomedical data repository, it is important to consider the differences between sunken, marginal, investment, operational, and underlying information infrastructure costs. Sunken costs have already been sustained and cannot be recovered, while marginal costs represent future costs, including costs for the next increment of effort. Decisions are best informed based on marginal costs because the incremental resources required for a project may be better understood. This dictum is true even if an institution requires a budget to be prepared otherwise. Investment costs are often underestimated at inception, in part owing to the cost of developing necessary new technology, the procurement costs of which may thus be greater than anticipated.

Organizations charged with constructing and operating a biomedical information resource can expect to bear costs in several major categories, which are external to the properties of the data. These are labor, IT infrastructure, IT services, media, licenses and subscriptions, facilities and utilities, outside services, travel, institutional overhead, and other soft costs. The relative costs of hardware and storage media for long-term preservation need to be compared in a systematic way, especially in light of how quickly options evolve. The discrepancy between true costs and price can be particularly important for institutions acting in the public interest, which may be instructed to include the full cost of producing a result or to avoid practices that impose social costs not reflected in market prices. They may be instructed to finance some of the subsidies. They may also be directed to reduce the impact of imperfections through how they procure an item.

Steps for Forecasting Costs of a Biomedical Information Resource

1. Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle.	• Decide the goals and objectives for the data resource. • Consider how the resource is likely to be used now and in the future. • Identify available guidance that defines the type of resource to be created or managed. • Compare the above with the activities defined for each of the data states and decide which data state(s) best align(s).
2. Identify the characteristics of the data, the data contributors, and users.	• Fill in the cost driver template in order to o Identify the size, complexity, replaceability, and depth versus breadth of the data; metadata requirements; and processing levels and fidelity; o Identify the life cycle issues; and o Identify data contributors and users.
3. Identify the current and potential value of the data and how the data value might be maintained or increased with time.	• Consult with the institution hosting the data resource, the project funders, and the broader research community to develop appropriate metrics for assessing the value of the data. • Identify decisions that affect data value in the shorter and longer terms. • Consider how data generation methodologies affect short- and long-term data value in terms of data contributors and users and the data life cycle.
4. Identify the personnel and infrastructure likely necessary in the short and long terms.	• Identify the major activities and sub-activities associated with the information resource, including activities related to potential transitions between data states. • Identify short- and long-term staffing requirements for the current state and transition between states. • Identify the infrastructure requirements and available resources.
5. Identify the major cost drivers associated with each activity based on the steps above, including how decisions might affect future data use and its cost.	• Identify the major cost drivers and associated uncertainties for each of the activities identified above by completing the cost driver template (download here). • Identify likely relative costs. • Consult with institutional experts and determine available personnel and infrastructure resources. • Work with experts at the host institution to quantify short-term costs and to bound uncertainties in longer-term forecasts.
6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource.	• Identify which cost drivers are important for each cost component of the information resource (e.g., labor, information technology infrastructure and services, media, licenses and subscriptions, facilities and utilities, outside services, travel, and institutional overhead). • Estimate costs for the current funding period. • Estimate costs and cost uncertainties for future funding periods, including costs to transition data to other states.

Disruptors

It is essential that decision-makers and cost forecasters keep an eye towards emerging disruptors, which could radically change how research is conducted and data are collected, used, archived, or preserved. 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.

The biomedical research community will likely continue to experience spurts in data growth that will tend to either (1) add a dimension to the data space or (2) extend a dimension by an order of magnitude. The size and complexity of those data sets are overwhelming existing repository structures and are pushing the boundaries of the current capabilities of technologies to access, manage, integrate, and analyze them at scale. Increasingly, biomedical data are too voluminous for a single platform, too unstructured for a traditional database system, or too continuous to store for analysis at a later time. More than ever, such challenges or possible cost increases associated with data must be considered in the context of additional value and novel opportunities for scientific understanding at different scales.

Increases in data volumes have been accompanied by the increased use of machine learning and artificial intelligence techniques, which have the potential to lower costs by automating processes. However, they also can give rise to new challenges, including data identifiability and security.

Because of the continuous growth of aggregate dataset sizes, storage technologies and practices will continue to evolve. This will be met with the physical and financial constraints of upgrading and expanding storage options. The scale and speed of the adoption of emerging technologies in the next 5 to 10 years is uncertain. Future computing technologies may also reshape how data are stored and reused, and the associated costs of storage and usage.

Lastly, changes in legislation and policy related to issues such as data sharing; data identifiability; permissible data collection, storage, and sharing; and as related to human-subjects research may require changes in the way data are stored, shared, and accessed.

Community Next Steps

In addition to utilizing the cost-forecasting framework, the following strategies can help enable efficient long-term data management and effective cost forecasting:

- **Create data environments that foster discoverability and interpretability through planning and investment throughout the data life cycle.**
  Making data discoverable, interpretable, and reusable requires a lot of forethought and sustained long-term investment from those who manage the data. However, doing so can help alleviate some of the challenges associated with generating, using, and storing growing volumes of complex data. After all, the objective is data reuse, not just data sharing.
- **Incorporate data management activities throughout the data life cycle to strengthen data curation and preservation.**
  Long-term data curation and data management needs are in every data state. Thus, the biomedical research community must expand the focus of data management and curation activities to include the entire life cycle, not just the end of the funding period. Up-front costs may be increased, but data value may also increase, and the overall cost of research may be reduced.
- **Incorporate the expertise and resources needed to create and curate metadata throughout the data life cycle, and in the transition between data states, into the cost forecast.**
  It is up to everyone in the biomedical research community to promote, support, and improve the understanding of the expertise and resources required for proper metadata that facilitates data discoverability and interpretability in and between all data states. This will likely be more successful if researchers are involved in decision-making and preservation efforts. It will also be more efficient if data librarians work with researchers, as they can help ensure adherence to community-accepted data and metadata standards.
- **Weigh the benefits, risks, and costs (both short- and long-term) of data storage and computation options before selecting among them**.
  Decision-makers should give substantial attention to several additional features of the data, regardless of the storage and computation options. Such features include confidentiality, ownership, and security; standards, regulatory, and governance concerns; access control; and the various disruptors listed in the report. The risk-management strategy of service providers, and of any evolution that strategy undergoes with time, needs to be understood and addressed. The institution managing an information resource is not absolved from information technology responsibilities if commercial vendors are chosen to provide services.

Furthermore, pursuing 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.**
  Developing and operating a sophisticated active aggregating platform requires an organization, developers, user-interface designers, training and documentation, help desks, and community building. Current storage costs (even total storage costs) are only one—and, in many cases, probably not the dominant—factor in total system costs. However, as mentioned above, data can be multiplicatively integrative when stored in the proper location, providing benefits beyond the original research. The biomedical research community needs to recognize that the long-term benefits of properly supporting active data resources 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 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.
- **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.**
  The biomedical research community needs to provide incentives for adopting good data management practices, including good cost-forecasting practices that facilitate sustainable long-term data preservation, curation, and access. Researchers often lack the skills needed for efficient and effective data management, which translates to a lack of meaningful management and good data stewardship, and little understanding of the real costs of effective management or of how to forecast them.

- **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.**

  Regardless of who provides the resources, there is a lack of visibility regarding storage costs in individual laboratories, institutions, and community resources. Understanding the charges associated with storage and computation in a data resource is vital for researchers making decisions about their own 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. Costs may be invisible to them if borne by their institutions or by a data-resource-platform manager. Purchased services (e.g., storage and computing) may be important, although the ability of individual researchers working in a primary research environment to forecast and manage those costs depends on the transparency of the information-technology environment. Mechanisms are needed to inform researchers of the actual costs paid for the services rendered to them, even if they are not directly charged.

An incentive for researchers to participate constructively in data management, and especially planning for active data resources, is providing them the opportunity to influence a superior computational environment. A data science platform could support complex research environments that free the researcher to focus on the science rather than on data collection and management. This capability could effectively reduce a primary research environment to data (i.e., signal) capture. This idea underscores the benefit of a closer interaction between the data curators of active data resources and individual researchers, recognizing that there are a variety of approaches to building and managing archives. This approach would co-locate the costs of supporting computing and analytics with an active repository.

Proper data preservation is a complex endeavor requiring dedicated resources over the long term. Many funding 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. Although data management practices in the laboratory are at the front line of eventual data sharing and long-term data access, many researchers do not see the inherent value in thinking about long-term data curation and preservation and lack external incentives to encourage them otherwise. To the greatest extent possible, it should be made easy for researchers and other stakeholders to make good data-related decisions from the onset.

Implementing this cost forecasting framework into the repository landscape 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.