VISION, NEEDS, AND PROPOSED ACTIONS FOR DATA FOR THE BIOECONOMY INITIATIVE

Product of the

INTERAGENCY WORKING GROUP
ON DATA FOR THE BIOECONOMY

of the

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

DECEMBER 2023

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the federal government. More information is available at https://www.whitehouse.gov/ostp.

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the federal research and development enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at https://www.whitehouse.gov/ostp/nstc.

About Data for the Bioeconomy Initiative Interagency Working Group

The Interagency Working Group on Data for the Bioeconomy, under the auspices of OSTP, coordinates the Data for the Bioeconomy Initiative (Data Initiative) outlined in the Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy to ensure that high-quality, wide-ranging, easily accessible, and secure biological datasets can drive breakthroughs for the United States bioeconomy.

About this Document

This document provides a U.S. government-wide vision, needs assessment, and action plan to enable a national data-driven bioeconomy. This effort will engage members of the federal, industry, nonprofit, and academic communities and build on existing data sources, infrastructure, training opportunities, and public-private partnerships. The action plan supports U.S. leadership in biotechnology and biomanufacturing and enhanced scientific and technological innovation, economic growth, commercial development, and science, technology, engineering, and mathematics (STEM) education and workforce development.

Copyright Information

NATIONAL SCIENCE & TECHNOLOGY COUNCIL

Chair

Arati Prabhakar
Assistant to the President for Science and Technology and Director, White House Office of Science and Technology Policy

Executive Director (Acting)

Kei Koizumi
Principal Deputy Director for Policy
White House Office of Science and Technology Policy

INTERAGENCY WORKING GROUP ON DATA FOR THE BIOECONOMY

Co-Chairs

Jason Albert, White House Office of Science and Technology Policy (through June 2023)
Gayle Bentley, Department of Energy
Sylvia Spengler, National Science Foundation

Executive Secretary

Jared Dashoff, National Science Foundation

Members

Council of Economic Advisors
Margaret Loudermilk (through June 2023)

Department of Agriculture
Anastasia Bodnar
Jeffrey O'Hara
Cyndy Parr

Department of Defense
Katherine Sixt

Department of Energy
Ramana Madupu

Department of Health and Human Services
Kristin DeBord
David Hassell

Department of Veterans Affairs
Sumitra Muralidhar

Environmental Protection Agency
Carolina Penalva Arana

National Aeronautics and Space Administration
Jonathan Galazka

National Institutes of Health
Amy Hafez
Taunton Paine

National Institute of Standards and Technology
Anne Plant

National Science Foundation
Andrew DeSoto
Steven Ellis
William Miller
Theodore Morgan
Kirsten Schwarz

Office of Management and Budget
Eileen Baca
Farnoosh Faezi-Marian
Laurel Fuller

Office of Science and Technology Policy
Sarah Glaven
Maryam Zaringhalam

Smithsonian Institution
Carol Butler

Table of Contents

Executive Summary

In September 2022, President Biden signed the Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy.[1] The bioeconomy refers to economic activity derived from the life sciences, particularly in the areas of biotechnology and biomanufacturing, and includes industries, products, services, and the workforce. The goal of this Executive Order was to advance biotechnology and biomanufacturing towards innovative solutions in health, climate change, energy, food security, agriculture, supply chain resilience, and national and economic security. Action towards these goals requires high-quality, wide-ranging, easily accessible, and secure biological datasets to drive breakthroughs for the U.S. bioeconomy.

Section 4(a) of the Executive Order establishes, as part of the National Biotechnology and Biomanufacturing Initiative (NBBI), a Data for the Bioeconomy Initiative (Data Initiative) for which this report lays the groundwork. The report describes existing federal data infrastructure, data gaps, and data and computational infrastructure needs and highlights needed strategic investments in building and maintaining a robust data infrastructure to serve as the foundation for the Data Initiative. The report also outlines a vision for transformative outcomes across the sectors of the bioeconomy, including health, agriculture, the environment, and biomanufacturing.

The federal government supports a vibrant community of researchers who generate enormous amounts of data and rely on existing data infrastructure and computational resources needed to develop technologies that underpin the bioeconomy. In addition to these existing resources, there remain opportunities to modernize and better integrate diverse and rapidly evolving sources of data. At present, many current mechanisms to support the existing federal data infrastructure are not designed for long-term continuity and findability, increased data storage needs, or the agility to respond to rapidly evolving data types and computational capabilities.

To realize a thriving bioeconomy, the Data Initiative identifies strategic investments and opportunities to leverage and build upon existing resources. The goal is to create an interwoven data fabric that connects data with the infrastructure and computational resources necessary to analyze, synthesize, and use those data for the widest audience. This vision depends on creation and adoption of community-driven standards, both for data and for repositories to enable interoperability and integration; training and education to build the bioeconomy data workforce of tomorrow; efforts to limit and mitigate security risks; and ongoing coordination to ensure efforts keep pace with transformations in data science, computing, biotechnology and biomanufacturing. While additional data are needed, government coordination and investment in infrastructure are also needed to make best use of the existing and anticipated data.

[1] The White House. *Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy*. 12 Sept. 2022, www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/.

The Data Initiative requires consistent whole-of-government coordination and investments in the following seven Core Actions:

1. **Dedicated long-term funding mechanisms for data and computational resources and infrastructure.**
2. **Standards** to establish common best practices that foster and strengthen a shared U.S. bioeconomy data ecosystem.
3. **Biodata Catalog** to identify extant data and metadata.
4. **Security** practices and policies that secure the data landscape while supporting innovation.
5. **Workforce** to drive U.S. leadership in the bioeconomy of the future.
6. **Strategically Targeted Areas for Rapid Transformation (STARTs)** to determine viability and impact and chart a course for larger investments.
7. **Coordination of intergovernmental investments, efforts, and resources.**

Data accessibility will determine the success of the U.S. bioeconomy. The activities and investments proposed here provide a clear path to fill acute gaps and are designed to secure the position of the United States as a leader in biotechnology and biomanufacturing, helping the nation achieve the Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

[2] White House Office of Science and Technology Policy. *Bold Goals for U.S. Biotechnology and Biomanufacturing: Harnessing Research and Development to Further Societal Goals*. Mar. 2023, www.whitehouse.gov/wp-content/uploads/2023/03/Bold-Goals-for-U.S.-Biotechnology-and-Biomanufacturing-Harnessing-Research-and-Development-To-Further-Societal-Goals-FINAL.pdf.

Introduction

The Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy[1] lays the groundwork for a range of new and persistent investments in biotechnology and biomanufacturing as part of a National Biotechnology and Biomanufacturing Initiative (NBBI). Part of the NBBI is a Data for the Bioeconomy Initiative (Data Initiative) that will streamline biotechnology developers access to high-quality, secure, and wide-ranging datasets and cutting-edge computing resources. Ultimately, this will enable a thriving bioeconomy that can drive solutions in health, climate change, energy, food security, agriculture, supply chain resilience, and national and economic security.

Data for the Bold Goals

A series of Bold Goals for U.S. Biotechnology and Biomanufacturing[2] (Bold Goals) chart the path to increase U.S. domestic capacity to advance the bioeconomy, all while reflecting necessary biosecurity and safety measures. These Goals cut across climate, food and agriculture, supply chain, and health and include specific aims such as:

- Producing 35 billion gallons of sustainable aviation fuel with at least 50% reduction in greenhouse gases by 2050;
- Spurring a circular economy for materials in which 90% of today's plastics can be replaced with biomaterials that will not pollute the environment;
- Increasing agricultural productivity while reducing nitrogen and methane emissions;
- Developing new food and feed sources to eliminate global hunger by 2030;
- Sustainably and renewably producing at least 30% of the chemicals we rely on;
- Enabling precision medicine and development of new therapeutic treatments; and
- Developing the foundational knowledge and technologies to enable these applications.

Achieving these goals holds the potential to build a new, sustainable economy that leverages renewable resources and advances equity, in alignment with the Justice40 Initiative.[3]

This report provides information on existing federally supported data sources (see Appendix A: Existing Data Sources by Sector) and gaps in the nation's data inventory. Much public investment has been made in data for the bioeconomy to date, including in federally operated or maintained datasets and resources. A thriving U.S. bioeconomy requires strategic investments and coordination to ensure the data and computational infrastructure has sustained support, the supported data are accessible, and the infrastructure is flexible to respond to rapidly evolving data sources and applications. Investment and coordination around in this data infrastructure is a critical component to enable the Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

[3] The White House. *Executive Order on Tackling the Climate Crisis at Home and Abroad*. 27 Jan. 2021, www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/executive-order-on-tackling-the-climate-crisis-at-home-and-abroad/.

Data Challenges and Opportunities

The bioeconomy is built on a wealth and breadth of data, including multi-omics data (i.e., genomic, epigenomic, transcriptomic, proteomic, phenomic, and metabolomic data), biological image data, health data, environmental and satellite data, nutrition and agricultural data, bioreactor data and data that captures the inputs and outcomes of biomanufacturing, and more. This report will outline many of the key data sources and repositories actively supported by federal investments. A diverse set of researchers, institutions, and companies across the U.S and internationally can leverage these data to advance the promise of biology to solve societal challenges and promote expansion of the U.S. bioeconomy, maintaining the nation's global leadership.

Data and their associated metadata (i.e., a basic, controlled description of the data that can include information about how, where, and when it was collected, how it may have been processed, usage and access rights, and structural information that make finding and working with the data easier), are not always findable, accessible, interoperable, and reusable (FAIR).[4] Federally funded research and development (R&D) has created a significant body of data, often held in individual research datasets. The diversity and rapid expansion of the biosciences has led to a range of data collected by varied researchers, from various locations and organisms, for various endpoints or uses, at different times, in different ways, at different scales, and with different metadata associated. The issues created by the diversity of data and the lack of FAIR-ness are further impacted by the lack of sustained data and computing infrastructure to support data collection, usage, and curation to make the data findable and accessible. Data and computing infrastructure are becoming more critical as more data are created and/or annotated.

Efforts to make data resulting from federally funded research freely available and publicly accessible are ongoing. The 2022 OSTP Memorandum for the Heads of Executive Departments and Agencies on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research[5] outlines expectations for making federally funded scientific data[6] available in public access repositories, unless subject to legal, privacy, ethical, technical, intellectual property, or security limitations. Agency policies implementing OSTP's expectations should go into effect no later than December 31, 2025. This

[4] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, no. 1, Mar. 2016, https://doi.org/10.1038/sdata.2016.18.

[5] White House Office of Science and Technology Policy. *Memorandum for the Heads of Executive Departments and Agencies on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*. 25 Aug. 2022, www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

[6] The 2022 OSTP Public Access Memorandum defines "scientific data" as the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings. Such scientific data do not include laboratory notebooks, preliminary analyses, case report forms, drafts of scientific papers, plans for future research, peer-reviews, communications with colleagues, or physical objects and materials, such as laboratory specimens, artifacts, or field notes. The definition of "scientific data" is similar to but broader than the term "research data" defined by 2 CFR 200.315 (e) and 45 CFR 75.322 (e).

memorandum builds on and strengthens expectations laid out in the 2013 OSTP Memorandum.[7] Agency public access plans should outline expectations for making data associated with peer-reviewed publications immediately available and provide timelines or approaches for sharing other federally funded data not associated with publications. With all these new data becoming publicly available, it will be critical to ensure data are FAIR and to support aggregation, analysis, and synthesis of these data.

Any data initiative must carefully consider the ethics associated with data access, sharing, ownership. The ethical, legal, and societal implications (ELSI) of research are central to many of the proposed recommendations in this report. Transparent and inclusive dialogue is needed in order to reconcile how advancing data for the bioeconomy aligns, diverges, or is in conflict with ELSI considerations.

In addition to the ethical, equity, and safety needs outlined in the Cross-Cutting Advances section of Bold Goals for U.S. Biotechnology and Biomanufacturing,[2] efforts focused on the bioeconomy must consider issues consistent with legal and policy requirements related to a range of issues. These include maintaining privacy and confidentiality (e.g., use of anonymization/pseudo-anonymization as a means to mitigate privacy or other concerns), respecting consent, Tribal sovereignty,[8] national security, and protection of sensitive data. Building the most effective bioeconomy enterprise requires full consideration of these issues. This includes areas such as understanding how data are used for scientific and societal good, sustaining and updating methods of security and privacy protection in alignment with societal expectations, and enhancing the culture of data sharing and curation. These issues will continuously need to be evaluated and studied as methods and approaches to advancing the bioeconomy change, including the expansion of co-produced and engaged research methodologies, to ensure data efforts continue to meet societal expectations.

A vision for a data-driven bioeconomy

The specific outcomes listed below serve as exemplary vignettes of the possibilities the Data Initiative may enable. The selected outcomes are inclusive of the broader range of Bold Goals, all of which are predicated on a more robust data infrastructure, as outlined in Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

Cross-cutting outcomes

Biosurveillance is “the process of gathering, integrating, interpreting, and communicating essential information and indications related to all-hazard threats or disease activity affecting human, animal,

[7] White House Office of Science and Technology Policy. *Memorandum for the Heads of Executive Departments and Agencies on Increasing Access to the Results of Federally Funded Scientific Research*. 22 Feb. 2013, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

[8] White House Office of Science and Technology Policy, and White House Council on Environmental Quality. *Guidance for Federal Departments and Agencies on Indigenous Knowledge*. 30 Nov. 2022, https://www.whitehouse.gov/wp-content/uploads/2022/12/OSTP-CEQ-IK-Guidance.pdf.

plant, and environmental health to … enable better decision-making at all levels."[9] Interconnections between federal and commercial computational and data resources can support biosurveillance across settings, allowing researchers to understand host-pathogen characteristics to treat and prevent disease. At the same time, researchers can work with health care providers to mitigate the impact of disease and support the work of public health agencies. Other outcomes include enabling researchers to create new forms of sustainable energy and enabling farmers to enhance food security by regulating the soil microbiome in fields to ensure crop health. Connecting these data with other modalities such as Global Positioning System (GPS), weather, environmental, and chemical data could help predict changes in species distribution and warn of natural disasters, such as crop failures, toxic invasive species, and the spread of vector-borne disease. A wide range of dramatic impacts would be possible based on this one example of a new technology facilitated by an improved data infrastructure. These potential outcomes align with Themes 1, 2, and 3 of the cross-cutting section of Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

Health

Understanding of disease mechanisms, such as the role of viruses in cancer, is quickly improving. In this rapidly evolving landscape, an improved data infrastructure may unlock a deeper understanding of the development and progression of diseases and create the potential to prevent or treat complex diseases like cancer or heart disease. Making data interoperable with information on known disease vectors and environmental factors could help identify and mitigate pandemics or identify environmental sources of carcinogens. To identify appropriate treatments or drugs, disease information, such as molecular and cellular mechanisms of neural degeneration, could be combined with computing resources and artificial intelligence (AI). When employing AI technologies, there is enormous potential to harness these technologies for the public good, but to do so, it is necessary to also identify and mitigate the risks.[10] Curating, validating, indexing, aggregating, and analyzing the diverse and siloed sets of data from patient care and clinical research programs and basic, pre-clinical research could result in knowledge that can drive improvements in healthcare; such outcomes would support all Themes, and most directly Themes 1 and 2, outlined in the health section of Bold Goals for U.S. Biotechnology and Biomanufacturing.[2] Together, these outcomes can improve human health and well-being, and spur a new industry of preventive medicine and targeted treatments informed by AI.

Food and agriculture

Amid climate change, the agriculture sector is facing increasing pressure to feed the planet and responsibly manage waste residues to serve as feedstocks for renewable fuels and chemicals. A new era of control in agriculture may be possible, in which farmers and breeders can increase productivity,

[9] The White House. *National Biodefense Strategy and Implementation Plan*. Oct. 2022, https://www.whitehouse.gov/wp-content/uploads/2022/10/National-Biodefense-Strategy-and-Implementation-Plan-Final.pdf.

[10] See Fact Sheet on the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

nutrition density, and resilience of food stocks, while reducing fertilizer needs and emissions. These outcomes would be enabled by an interoperable, networked data fabric that allows for aggregating broad multi-omics data from crops, wild relatives, livestock, pests, and pathogens.

Adding biophysical and economic data can foster innovation and improvement of alternative sources of protein, fat, and biofuel feedstocks that are increasingly climate smart and ready for adoption. This can be done with minimal impact on the environment. For example, multi-omics data could be leveraged to develop new metabolic models using machine learning and AI to predict optimal production pathways of model biosystems, anticipate the impact of pests or climate change on a critical crop, and predict effective mitigation strategies with low environmental and health impact. Achieving these outcomes would address Themes 1, 2 and 3 outlined in the food and agriculture section of Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

Environment

Healthy functioning ecosystems, including clean air, water, and soil, form the foundation of any bioeconomy. Ecosystems are complex and varied, and factors that impact functioning in one system may not translate to change in another. Connecting GPS satellites, weather, and environmental data with data on species distribution, as noted above, could help predict changes in species distribution and warn of natural disasters such as crop failures, invasive species, and the spread of vector borne disease. Increased access to multi-scale data on ecosystem structure and function could inform land management practices to maximize resilience to climate change, the design of green infrastructure to reduce flooding, provide urban cooling, promote biodiversity, provide recreational opportunities, and enhance environmental justice. The Data Initiative provides the framework to support the demand for biotechnology solutions to support a changing environment. Achieving these outcomes would support Theme 3 of the food and agriculture section and Theme 1 of the cross-cutting section of Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

Biomanufacturing

The biomanufacturing sector can support the deployment of a wide library of products and their respective nascent markets, including: biobased fuels and chemicals, materials and bioproducts, biohybrid, bioinspired or biomimetic systems; petrochemical-free biomanufacturing methods; and agricultural bioproducts that require less fertilizer or water or improve soil health. New products could be brought to market faster if existing high-quality data from high-throughput experimentation were made available for AI and machine learning. AI and machine learning could further be used to predict the scalability of a renewable biochemical process that will drive a new chemical market and ensure that production of that chemical will be sustainable and cost-effective. Shared ontologies and sustained data resources may increase the speed and scale of biomanufacturing, along with automation and downstream analysis and characterization needed for regulatory consideration.

The new data fabric would advance our fundamental understanding of the biology of plants, animals, microbes, and the environment, providing the basis for developing innovative processes for bioenergy and bioproducts. Durable goods may be produced in the future from carbon dioxide as a direct feedstock, but this first requires data-dependent process development, economic and energy modeling, and incorporating the science of scale-up to facilitate large-scale deployment of technologies based on optimizations performed in the laboratory. Achieving these outcomes would support Themes 1 through 4 of the Climate Change section of Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

The Data Initiative

This report will describe the existing data infrastructure, map the current needs for data generation and curation, and introduce the case for actions to be taken as part of the Data Initiative. The Data Initiative will ensure high-quality, wide-ranging, easily accessible, and secure biological data resources are supported through an interwoven fabric of federal and commercial computational and infrastructure resources.

Filling gaps in the nation's data inventory and ensuring the existence of a sustained, robust data and computational infrastructure will achieve the vision of a data-driven bioeconomy in which data fuels discovery, innovation, and advances that grow the U.S. bioeconomy (see Figure 2), and which are critical to investing in the Bold Goals for U.S. Biotechnology and Biomanufacturing[2] themselves.

Doing so would require:

- Creation of new mechanisms, provision of funding, and coordination around existing resources to enable long-term, sustained support of critical data and computational resources as part of an interwoven fabric;
- Efforts to develop and increase the use of data and metadata standards;
- Development of a catalog of existing data sources to increase findability;
- Data security and protection;
- A more diverse and larger biotechnology, biomanufacturing, computing, and data science workforce; and
- Ongoing coordination between agencies and key communities.

Existing data landscape

There is a large and growing body of data for the bioeconomy generated by government agencies, federally funded academic or industry scientists, and privately funded scientists. These generators, along with data users, actively rely on the data infrastructure and computational resources to accelerate development of technologies that underpin the bioeconomy. Due in part to the policies developed in response to the 2013 OSTP Public Access Memorandum,[7] federally supported researchers are sharing increasing amounts of data, including through intramural research programs within agencies and extramurally-funded research supported by grants and contracts.[5] Sharing of federally-supported data will further increase due to strengthened agency policies around data sharing, which go into effect by December 31, 2025 in response to the 2022 OSTP Public Access Memorandum.[5] Combined, these data provide information on a range of organisms and biological processes, disease and health, the environment, managed and urban lands, and more, all of which has relevance across sectors of the bioeconomy.

Additionally, much public investment has been made in federally operated or maintained data and resources. Appendix A contains a partial reference to some of the existing data sources, including those that cut across sectors and those that are sector specific. The listing includes a brief description of the resource and the federal agencies that provide funding for or maintain the resource. Sources include those supported or maintained by the:

- Department of Agriculture (USDA), including the Agricultural Research Service (ARS)
- Department of Commerce (DOC)
- Department of Defense (DOD), including the Uniformed Services University of the Health Sciences (USUHS)
- Department of Energy (DOE), including the Advanced Research Projects Agency-Energy (ARPA-E)
- Department of Veterans Affairs (VA)
- Food and Drug Administration (FDA)
- National Aeronautics and Space Administration (NASA)
- National Institutes of Health (NIH)
- National Institute of Standards and Technology (NIST)
- National Oceanic and Atmospheric Administration (NOAA)
- National Science Foundation (NSF), including several directorates within the Foundation
- U.S. Geological Survey (USGS)

There are also ongoing and recent investments in computing and networking infrastructure and services investments across many agencies, such as the NSF Advanced Infrastructure Coordination Ecosystem: Services & Support (ACCESS)[11] system, NIH's Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) initiative,[12] and DOE's high-performance computing (HPC) user facilities at the national labs and high-performance network user facility (ESnet).[13] The Data Initiative should leverage these existing resources and best practices, coupled with strategic investments and coordination efforts, as it works to meet the needs of a data-driven bioeconomy.

Needs to support the data-driven bioeconomy

Many reports from federal agencies and other organizations and intitiatives,[14] including those highlighted in Figure 1, note that the current state of data supporting the bioeconomy and the set of

[11] See NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support: https://access-ci.org/

[12] See NIH's Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative: https://cloud.nih.gov/

[13] See DOE's Energy Sciences Network (ESnet): https://www.es.net/

[14] Reports include: *Artificial Intelligence and Machine Learning for Bioenergy Research: Opportunities and Challenges* (DOE, https://doi.org/10.2172/1968870); *Designing for Deep Decarbonization: Accelerating the U.S. Bioeconomy* (DOE, https://biosciences.lbl.gov/wp-content/uploads/2021/12/21-BAO-3054-Designing-the-Bioeconomy-for-Deep-Decarbonization-Report_v5.pdf); *Genome Engineering for Materials Synthesis Workshop Report* (DOE, https://science.osti.gov/-/media/ber/pdf/community-resources/2019/GEMS_Report_2019.pdf); *Breaking the*

infrastructure and computing resources are not sufficiently organized to help drive the U.S. bioeconomy. There is an insufficient amount of data available, owing in part to lack of incentives and support for their generators to make data available; data that are available often cannot be easily found or combined, lack standards, are of uncertain quality, and may not reliably be available in the future. This is particularly true for data created and/or maintained by academic institutions and researchers with federal support.

National Strategy for Modernizing the Regulatory System for Biotechnology Products

Product of the Emerging Technologies Interagency Policy Coordination Committee's Biotechnology Working Group

September 2016

★★★

REPORT TO THE PRESIDENT

Biomanufacturing to Advance the Bioeconomy

Executive Office of the President

President's Council of Advisors on Science and Technology

December 2022

The U.S. Bioeconomy: Charting a Course for a Resilient and Competitive Future

EBRC Engineering Biology Research Consortium

Moonshots for the 21st-Century Bioeconomy

A Policy Paper by the Engineering Biology Research Consortium

Compiled and edited by Emily R. Aurand, EBRC Director of Roadmapping and Education

December 2022

Engineering biology has the immense potential to revolutionize our economy, further our societal goals, and ensure products, technologies, and solutions that benefit all Americans. A number of consumer products and solutions that we see everyday integrate engineering biology-based technologies, helping Americans and people worldwide to live healthier and happier lives. For example, in the healthcare sector, many biologics, like insulin, have been produced via biomanufacturing for over 40 years and more recently, CAR T-cell immunotherapy is being used to treat complex cancers with improved effectiveness. Engineering biology was also responsible for the rapid development of the vaccines to fight the COVID-19 pandemic. Engineering biology has the potential to make these healthcare solutions more equitable and accessible, and to treat many of the diseases and illnesses that impact hundreds of thousands of lives. Engineering biology will also be pivotal to tackling the climate crisis and global sustainability challenges. Over the last 30 years, ethanol, biofuels, and sustainable chemicals have been successfully produced through the engineering of biology and expanding the use of these technologies will be critical for energy and supply chain security. Innovations in engineering biology are making possible carbon-negative manufacturing using microbes to convert carbon from the atmosphere into the products we use every day, from packaging materials to clothing and laundry detergent. The agriculture sector has long employed engineering biology tools and technologies to accelerate breeding and develop agronomically valuable traits, and is gaining steam with sustainable, climate-friendly fertilizers and crops resilient to drought and pests. And for the dinner table, numerous companies are providing alternative and plant-based meats, dairy, and other sustainable foods developed with engineering biology to consumers, helping to secure the food sector and combat climate change.

Commercial Products and Solutions	*Engineering Biology Technology*
COVID-19 vaccines (e.g., Pfizer, Moderna)	mRNA designed to trigger an immune response to the SARS-CoV-2 spike protein.
Alternative meats (e.g., Beyond Meat, Impossible Foods)	Engineered proteins and enzymes to look, taste, and feel like meat.
Sustainable, climate-friendly fertilizers (e.g., Pivot Bio)	Engineered microbes designed to increase nitrogen availability for crops.
Bio-based, renewable chemicals and materials (e.g., BASF, LanzaTech)	Biobased chemicals, enzymes, and proteins for nutrition, personal and home care, medical and dietary nutrition, and cleaning detergents.
Synthetic DNA (e.g., Twist Bioscience)	Customized synthesis of genes and genomes at various scales for use in research and industry.

5885 Hollis Street, 4th Floor | Emeryville, CA 94608 | 510-871-3272 | 510-245-2222 | www.ebrc.org

NIH STRATEGIC PLAN FOR DATA SCIENCE

Introduction

As articulated in the National Institutes of Health (NIH)-Wide Strategic Plan[1] and the Department of Health and Human Services (HHS) Strategic Plan,[2] our nation and the world stand at a unique moment of opportunity in biomedical research, and data science is an integral contributor. Understanding basic biological mechanisms through NIH-funded research depends upon vast amounts of data and has propelled biomedicine into the sphere of "Big Data" along with other sectors of the national and global economies. Reflecting today's highly integrated biomedical research landscape, NIH defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

NIH supports the generation and analysis of substantial quantities of biomedical research data (see, for

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

SAFEGUARDING the BIOECONOMY

Figure 1. Community input on data, data and cyber infrastructure, and data security needs can be found in a number of recent reports, including: The U.S. Bioeconomy: Charting a Course for a Resilient and Competitive

Bottleneck of Genomes: Understanding Gene Function Across Taxa Workshop (DOE, https://genomicscience.energy.gov/breaking-bottleneck-of-genomes/); *NIH Strategic Plan for Data Science* (https://datascience.nih.gov/nih-strategic-plan-data-science); *National Strategy to Advance Privacy-Preserving Data Sharing and Analytics* (NSTC, https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf); *Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies* (NIH, https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf); *Report to the National Library of Medicine Director on the State of Data Science Workforce Development* (NIH, https://www.nlm.nih.gov/pubs/reports/state_of_data_science_training_report_final2.pdf); *Department of Defense Biomanufacturing Strategy* (https://media.defense.gov/2023/Mar/22/2003184301/-1/-1/1/BIOMANUFACTURING-STRATEGY.PDF); *National Biodefense Strategy and Implementation Plan* (The White House, https://www.whitehouse.gov/wp-content/uploads/2022/10/National-Biodefense-Strategy-and-Implementation-Plan-Final.pdf); *Plan to Advance Data Innovation* (NSTC, https://www.whitehouse.gov/wp-content/uploads/2022/02/02-2022-Plan-to-Advance-Data-Innovation.pdf); *National Strategy for Modernizing the Regulatory System for Biotechnology Products* (Emerging Technologies Interagency Policy Coordination Committee, https://usbiotechnologyregulation.mrp.usda.gov/biotech_national_strategy_final.pdf); and *Biomanufacturing to Advance the Bioeconomy* (President's Council of Advisors on Science and Technology, https://www.whitehouse.gov/wp-content/uploads/2022/12/PCAST_Biomanufacturing-Report_Dec2022.pdf).

Future, Safeguarding the Bioeconomy*,* Moonshots for the 21st-Century Bioeconomy, the NIH Strategic Plan for Data Science, the National Strategy for Modernizing the Regulatory System for Biotechnology Products, and Biomanufacturing to Advance the Bioeconomy from the President's Council of Advisors on Science & Technology.

Data and computational infrastructure needs

Federal, academic, and industry researchers seeking to develop new biotechnologies and biomanufacturing processes have an array of unmet needs relating to data and computational infrastructure that hamper the overall research-driven bioeconomy, including insufficient means and capacity for storing, sharing, and publishing raw and analyzed data. These challenges are compounded by a lack of sustainable resources, a lack of streamlined access to existing data and computing resources, and limited knowledge and understanding within the user community about those resources and how to use them.

These challenges directly impact the ability to efficiently discover, develop, and translate scientific discoveries out to industry across R&D sectors. A national strategic investment in data and computational infrastructure is needed. For instance, the National AI Research Resource Task Force Report,[15] NSTC National Strategic Overview on R&D Infrastructure,[16] and the NSTC Future Advanced Computing Ecosystem Strategic Plan FY2022 Implementation Roadmap[17] cite similar needs for a broad based computational and data infrastructure ecosystem.

Long-term infrastructure funding mechanisms and support for modernization of existing infrastructure

Historically, data infrastructure has been decentralized and lacks interconnection, and is often supported with time-limited funding. Nearly all the data sources listed in Appendix A are supported by relatively short periods of funding before competitive renewal or the end-of-grant periods. The vast majority of federally supported data sources are funded under this paradigm. Consequently, these data sources, however presently valuable, cannot currently be relied on for the long-term realization of a thriving bioeconomy without the sustaining investments proposed below. Some key resources and tools supported by term-limited federal grants include: Protein Databank,[18] the DOE Systems Biology Knowledgebase,[19] the National Microbiome Data Collaborative,[20] and CyVerse.[21] Some of the

[15] National Artificial Intelligence Research Resource Task Force. *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem an Implementation Plan for a National Artificial Intelligence Research Resource*. Jan. 2023, www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

[16] NSTC Subcommittee on Research and Development Infrastructure. *National Strategic Overview for Research and Development Infrastructure*. Oct. 2021, www.whitehouse.gov/wp-content/uploads/2021/10/NSTC-NSO-RDI-_REV_FINAL-10-2021.pdf.

[17] NSTC Subcommittee on Future Advanced Computing Ecosystem. *Future Advanced Computing Ecosystem Strategic Plan FY2022 Implementation Roadmap*. May 2022, www.nitrd.gov/pubs/FACE-SP-FY22-Implementation-Roadmap.pdf.

[18] See RCSB Protein Data Bank: https://www.rcsb.org/

[19] See DOE's Systems Biology Knowledgebase (KBase): https://www.kbase.us/about/

[20] See National Microbiome Data Collaborative: https://microbiomedata.org/

[21] See NSF's CyVerse: https://cyverse.org/

more persistent exceptions are those hosted by the National Library of Medicine (NLM)[22] and the DOE Joint Genome Institute,[23] though existing funding may be insufficient for modernizing infrastructure. A model or system for long-term funding, with appropriate oversight for this critical data fabric is needed to ensure its existence, viability, and usability, and the availability of the data.

Further, there is a need to enhance data storage capabilities and modernize existing systems to align with cutting-edge technologies. Many existing federal and community data repositories and other federally supported data resources discussed in this report were originally deployed without the advantage of current technologies that enable seamless sharing and interoperability. Consequently, coordinated and strategic efforts at the national level are needed to ensure trusted and interoperable data infrastructures that enable accessibility to and utilization of a variety of data sources from many providers to address complex research questions. Capacity expansion is also key. The rapidly increasing volume and heterogeneity of data have made it critical to increase both the overall capacity for data storage and curation and the ability for sharing among multiple communities.

Index of open science data and analysis

To facilitate awareness of existing and future data, increase access to data sources, and enhance data sharing and interoperability, there is a need to index information around data and metadata from across the federal government and federally funded research. In addition, such an index must have stable funding for its implementation.

Indexing could be aided by use of unique persistent identifiers (PIDs) and associated metadata, leveraging expectations around PIDs and metadata outlined in the 2022 OSTP Public Access Memorandum.[5,24] An approach to data security for such an index that engages the federal, private, nonprofit, and academic sectors and enables indexing of publicly accessible data with necessary controls, unauthorized access to data, or manipulation of data or computational systems is also required.

Workforce development

Training activities are required to expand the cadre of workers who can contribute to the acquisition, use, sharing, and analysis of data for the bioeconomy in private, public, academic, corporate, and federal contexts. Because of the high demand for these skilled workers, additional federal hiring mechanisms are needed to ensure the federal government can hire developers and engineers who can often attract significant salaries in the private sector.[25] Additional workforce needs include enhanced career paths for computer and data science support staff that are essential to maintaining, upgrading,

[22] See resources from the National Library of Medicine at the NIH: https://www.ncbi.nlm.nih.gov/

[23] See resources from DOE's Joint Genome Institute (JGI): https://jgi.doe.gov/

[24] See the National Security Presidential Memorandum 33 Implementation Guidance for definition: A digital identifier that is globally unique, persistent, machine resolvable and processable, and has an associated metadata schema, https://www.whitehouse.gov/wp-content/uploads/2022/01/010422-NSPM-33-Implementation-Guidance.pdf

[25] See discussions in *NIH Strategic Plan for Data Science* (https://datascience.nih.gov/nih-strategic-plan-data-science) and *Report to the National Library of Medicine Director on the State of Data Science Workforce Development* (NIH, https://www.nlm.nih.gov/pubs/reports/state_of_data_science_training_report_final2.pdf).

and increasing usability of data sources, as well as finding paths to support building a diversified workforce. These staff are particularly needed in academic settings. These needs and training activities align with those outlined in the Plan for Building the Bioworkforce of the Future, released in June 2023.[26] The Plan was developed by an Interagency Working Group to coordinate and use relevant federal education and training programs, while also recommending new efforts to promote multidisciplinary education programs, as directed in Section 4 of the Executive Order.[1]

Data needs

Issues may also exist with the data themselves. These issues include missing metadata creating data quality concerns, limited alignment between existing standards and substantial amounts of unstandardized data, lack of persistent metadata, restricted ability to integrate across data types and access analog data housed in collections, and inadequate data in key areas of relevance to the bioeconomy. Also limited are considerations for the social and human behavioral implications of data creation and use, including privacy, Tribal sovereignty, ethical, and national security concerns.

As such, there is a need to increase the development and use of data and metadata standards, promote greater integration and linkage, address quality and interoperability issues, and create new data that is created as part of collaborations between experimentalists, data experts, and end users while considering the societal implications and potential use in predictive forecasting.

Integration

The wide variety of research methods used across the bioeconomy to provide a holistic view of a sample or system provides more information than an approach that uses only a single data type. Additionally, as technologies advance measurement is possible at new scales and dimensions. To combine these diverse types of data at different scales into a comprehensive system view, data integration across platforms, levels, scales, and types is critical to gain a more complete perspective. For the bioeconomy, the major challenges include data heterogeneity and model integration, both of which depend critically on data curation and data quality.

Linkage

Tools and new or revised policies are needed to enable the federal government to connect diverse datasets together to enable the secure and privacy-protecting leveraging of different datasets together to draw new understandings to inform health and other areas of priority. Linkage (i.e., the ability to serve us data on the same entity from different datasets) requires both agreement on the meaning of data and the ability to share it in a networked information space. Linking genomic and other information with geographic range or disease incidence can help inform efforts to identify useful phenotypes of plants, while integration of multi-omic data and joint modelling and analyses reveal the systems biology for healthy and sustainable production of animals.

[26] White House Office of Science and Technology Policy. *Building the Bioworkforce of the Future*. June 2023, www.whitehouse.gov/wp-content/uploads/2023/06/Building-the-Bioworkforce-of-the-Future.pdf.

Connecting diverse datasets together, either directly or through privacy preserving record linkage, will support new understandings to inform health and other priority areas. For example, nutrition data and environmental/military exposure data are critically needed to inform whole health and are currently lagging in terms of availability and/or readiness for use in research and medical care. Crossing between the health and agricultural sectors, it is necessary to integrate food data across nutrition, diets of understudied populations and the associated health outcomes, individual health, production, consumer, and environmental impact axes, which will improve the health of all.

Best practices and policies related to the use of these tools are also likely needed to ensure consistent and appropriate adoption while also addressing ethics, consent considerations, and Tribal sovereignty, as well as privacy and security concerns. Privacy concerns include the use and disclosure of data protected under the Health Insurance Portability and Accountability Act (HIPAA) that often cannot be combined with other datasets.

Quality

High-quality data and notation of data quality are both critical to provide appropriate benchmark data across sources and to enhance reproducibility, as well as for translation to products. Data quality assurance can include annotation of a dataset and the associated metadata to meet existing or selected standards and removal of data once quality issues are identified. Improving the quality of data, including annotation and curation in line with standards, also improves the FAIR characteristics of those data.

Interoperability and reusability

Data for the bioeconomy must be FAIR to the greatest extent possible to foster aggregation, analysis, and synthesis of existing and future data that supports new advancements and cross disciplinary collaborations. Siloed approaches hamper data sharing, integration, and utilization among agencies and across disciplines. Achieving the needed integration across domains requires coordination across agencies and with the broader community to support biodata stewardship and promote the use of data standards for rigorous data reuse and collaboration. Areas of need include enabling clear descriptions of data ontologies and representation for the vast diversity of data sources and assuring smooth and sustainable collaboration and partnerships that produce rapid research outcomes and economic benefits.

When done appropriately and consistent with consent and privacy obligations, ethical standards and Tribal sovereignty, such integration would increase the reusability of data, extending the return on investment in generation and storage. In addition, such integration would also enhance the statistical power of data from clinical trials, which can often be small due to their targeted nature. The ability to accurately predict the behavior of designed biological systems, which requires models that incorporate emerging biological concepts, disparate measures, phenotypic data, and geographic information, would also be enhanced. Such increased predictive capabilities could lead to bio-based fuels, chemicals, materials and bioproducts needed by a modern society while limiting risks associated with the production thereof and securing supply chains.

Generation

While there is a wide range of existing federally maintained or supported data pipelines, and data continue to be generated in the process of answering a primary research question as part of a time-limited grant, there is a need for increased and more diverse data generation, particularly in areas that support the research and development needs outlined in Bold Goals for U.S. Biotechnology and Biomanufacturing.[2] These include strategic large scale non-human genome sequencing, bioreactor fermentation data, high-throughput phenotyping, and multi-omic characterization of non-model organisms and nontraditional host organisms. Biological materials, such as those in natural history collections, provide a source of annotated data for digitization, including current annotated samples and those analog data generated prior to the advent of online digital data sharing.[27,28] Persistent identifiers could be used as a mechanism to enhance the FAIR-ness of biological materials. Targeted generation of data or digitization of analog collections will require enhanced infrastructure, specifically storage capacity for increased amounts of data.

There is also a need to encourage other data generators such as nongovernmental organizations and industry to share data while maintaining necessary security and privacy protections and honoring Tribal sovereignty. This includes bioreactor data, human and animal -omics data. Like the silos that may exist in the research landscape, data kept privately may never reach their full potential. Use of precompetitive spaces, such as those that occur among pharmaceutical companies, academia, and the federal government through public-private partnerships, or other creative economic models such as tax incentives for donating data, may be needed to encourage sharing of these data while also addressing any security concerns.

Monitoring and Assessment

There is also a need to collect data on the broader set of behavioral, social, and economic factors of the sort studied by behavioral and social scientists. Many of these scientific areas focus on the human side of bioeconomy data and its use. For example, social scientists study the culture of data sharing and use; the "science of science," or metascience; and the science of innovation (i.e., fields which seek to understand how data are used for scientific and societal good and how these efforts can be strengthened). Areas of study may include: determining the benefit to science, economy, and society of publicly available data assets and infrastructure; generating community buy-in to data sharing; connecting with communities in other ways to encourage data use and reuse; understanding individual or community preferences regarding data use and governance; and understanding and managing issues around collection and application of bioeconomy data.

More generally, there remains a culture gap between experimentalists who generate data via general scientific inquiry and data scientists who seek to use data in computational approaches. Due to the structure in which experimental data are generated, these data are often not generated with data

[27] Schindel, David E., and Economic Study Group of the Interagency Working Group on Scientific Collections. *Economic Analyses of Federal Scientific Collections: Methods for Documenting Costs and Benefits*. Smithsonian Scholarly Press, 20 Nov. 2020, https://doi.org/10.5479/si.13241612.

[28] NSTC Interagency Working Group on Scientific Collections. *Scientific Collections: Mission-Critical Infrastructure of Federal Science Agencies*. 2009, https://iwgsc.nal.usda.gov/sites/default/files/IWGSC_GreenReport_FINAL_2009.pdf.

curation and downstream computation in mind. It is critical to understand what gaps remain in enabling the generation of high-quality curated data.

Data security needs

In addition to the infrastructure needs and those associated with the data themselves, there is a need to address security concerns and create the appropriate protections for dual-use information data with personally identifiable information, and proprietary data. This is especially true with regard to potential biothreats, health data, and proprietary data. To address these concerns, there is a need for enhanced security across the data lifecycle and support for new and existing infrastructure capable of controlling access.

Lifecycle Security

There are inherent risks in holding all data, including biotechnology and biomanufacturing data. As such, there is a need to monitor and mitigate risks, including those associated with data integrity, data management, disclosure, aggregation, reidentification, and release of proprietary or sensitive data. More standard approaches to data lifecycles would be beneficial, including appropriate duration of retention and criteria and processes for deaccessioning of publicly available data.

Data-agnostic infrastructure capable of controlling access for sensitive data

There is also a need for more broadly useable, data-agnostic infrastructure capable of controlling access to sensitive human or other data, when warranted or required, in a cost-effective and sustainable manner. This infrastructure is in place for some data and research types, such as the Database of Genotypes and Phenotypes (dbGaP),[29] but needs further development, support, and low-cost solutions.

Actions needed to support the data initiative

The next generation of the U.S. bioeconomy will be driven by the interconnection, integration, synthesis, analysis, and use of a wide range of data, including those produced by the domains of omics, imaging, engineering production, environmental sciences, and beyond. This data will be integrated with persistent high-performance computing and cloud resources to capitalize on and accelerate beyond current innovations in agriculture, manufacturing, health, energy, and the environment. While there are specific gaps in data collection to be identified and filled, a persistent need is improved data infrastructure, standards, and computational resources that can bridge, protect, and connect data sources from across the research landscape. This expansion and democratized access to infrastructure will stimulate innovation to enable the creation of novel products and new understanding of complex phenomena, as described in Bold Goals for U.S. Biotechnology and Biomanufacturing.[2]

Below are actions required as part of a Data for the Bioeconomy Initiative to achieve those necessary improvements. These actions aim to provide the infrastructure and data resources (see Figure 2)

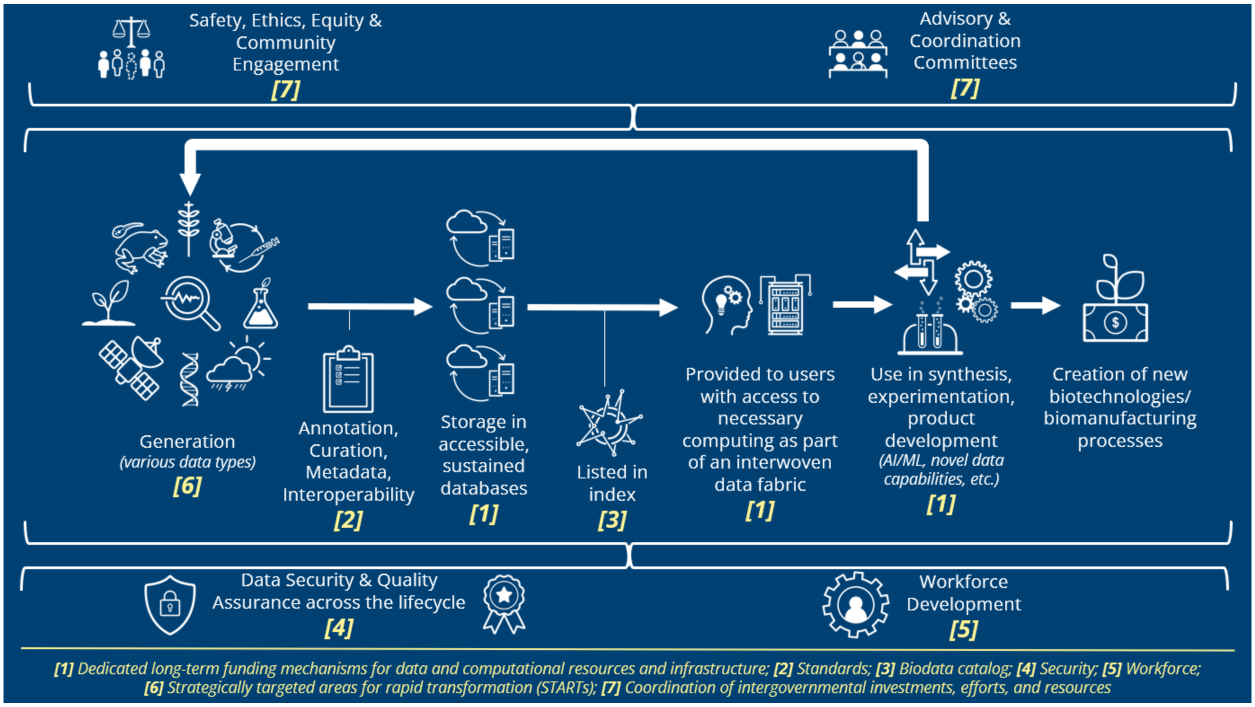
[29] See Database of Genotypes and Phenotypes (dbGaP): https://www.ncbi.nlm.nih.gov/gap/

needed to achieve the Bold Goals. Realization of these proposals will require prioritization of R&D investments and efforts across the U.S. government, and engagement with expertise in government, academia, industry, and nonprofit organizations at all levels throughout this process.

To date, resources have been the limiting factor in achieving these goals. Federal investment in and interagency coordination around the Initiative will be essential to achieve a discovery-enabling ecosystem that supports and accelerates the broader biotechnology initiative. Agencies should dedicate appropriate funding to be able to take the proposed actions.

The Initiative requires concerted, continuing whole-of-government coordination around and investments in each of the seven Core Actions:

1. **Dedicated long-term funding mechanisms for data and computational resources and infrastructure.** Design and support a trusted, sustainable data ecosystem that connects both data and computational resources, seamlessly integrating these to enable and accelerate bioeconomy outcomes.
2. **Standards.** Develop and adopt consensus standards to establish common best practices that foster and strengthen a shared U.S. bioeconomy data ecosystem of integrated resources spanning agencies, the research enterprise, and other key community members to assure smooth and sustainable collaboration and partnerships.
3. **Biodata Catalog.** Provide a new persistent resource that identifies extant data and metadata and adds new data and metadata coupled with unique persistent identifiers (PIDs).
4. **Security.** Adapt practices and develop policies that secure the data landscape while supporting innovation pathways.
5. **Workforce.** Include new federal hiring mechanisms and education, training, and engagement efforts to broaden and expand participation in data science and related STEM fields and to drive U.S. leadership in the bioeconomy of the future.
6. **Strategically Targeted Areas for Rapid Transformation (STARTs).** Consider focused investments in key areas of need to determine viability and impact and chart a course for larger investments.
7. **Coordination of intergovernmental investments, efforts, and resources.** Leverage NSTC frameworks for cross-government collaboration and agency support to create an advisory committee of external experts.


Safety, Ethics, Equity & Community Engagement
[7]
Advisory & Coordination Committees
[7]
Generation
(various data types)
[6]
Annotation, Curation, Metadata, Interoperability
[2]
Storage in accessible, sustained databases
[1]
Listed in index
[3]
Provided to users with access to necessary computing as part of an interwoven data fabric
[1]
Use in synthesis, experimentation, product development
(AI/ML, novel data capabilities, etc.)
[1]
Creation of new biotechnologies/ biomanufacturing processes
Data Security & Quality Assurance across the lifecycle
[4]
Workforce Development
[5]

[1] Dedicated long-term funding mechanisms for data and computational resources and infrastructure; [2] Standards; [3] Biodata catalog; [4] Security; [5] Workforce; [6] Strategically targeted areas for rapid transformation (STARTs); [7] Coordination of intergovernmental investments, efforts, and resources

Figure 2. Graphical representation of the vision of a data-driven bioeconomy wherein FAIR data and metadata supported by a nationwide infrastructure spur innovation across and within sectors that address societal challenges and ensure the nation's economic leadership. Bracketed numbers align with the efforts as listed in the numbered list above and in the key at the bottom of the graphic.

These investments and coordination efforts, further detailed below, are crucial actions needed to drive improvements in the data landscape. These should be taken as priorities for the Data Initiative. These actions directly support aspects of the vision for a data-driven bioeconomy outlined above (see Figure 2), the Bold Goals outlined by federal agencies,[2] and the goals of the NBBI listed in the Executive Order.[1]

Dedicated long-term funding for data and computational infrastructure

Long-term funding and support to sustain, modernize, and expand biodata repositories and data and computational infrastructure are fundamental in establishing and advancing a data-driven bioeconomy. For funding to sustain existing infrastructure and platforms and to safeguard investments, funding must keep pace with: increased data volumes and storage costs; changes in needs regarding security, access, and maintenance; and advances in data generation and analytic capabilities to enhance the competitiveness of the U.S. bioeconomy. Such investments require hardware and software systems, computational resources, and high-performance network infrastructure, and additional staffing expertise. The research and user community, along with funding agencies, industry, and resource managers should be consulted in the development of a strategy to ensure success of this critical investment. These resources should also align with existing best practices where they exist, such as those outlined in the NSTC Desirable Characteristics of Data

Repositories for Federally Funded Research.[30] Dedicated and appropriate funding are needed to take the proposed actions.

Investments in new data presentation approaches to aggregate, summarize, and visualize data at scale also are needed. In addition to coordinating and leveraging existing best practices, agencies should provide funds for data and computational infrastructure, including repositories supported both intramurally within the government and extramurally by DOE, NIH, NSF, USDA, and other agencies, which may require specific authorities.

A well-connected open ecosystem of bioeconomy data resources, analysis and visualization tools, and compute environments requires modernizing existing infrastructure and systems. These improvements can enable interconnectivity and interoperability at scale with appropriate annotation and curation plans. In so doing, researchers will be better equipped to find and utilize data to advance discovery and drive pathways to product development. A key set of coordinated investments will focus on creating data-driven computational approaches that develop and support the complex workflows required by bioeconomy research, such as seamlessly linking AI models with computational analysis and cloud computing environments at scale. An example in another domain is the planned NSF National Discovery Cloud for Climate[31] that will federate and enable democratized access to a range of climate data resources and their connection to the national cyberinfrastructure ecosystem. Such a coordinated resource can, in turn, engage and leverage the National AI Research Resource[15] as it develops. Seamless access to advanced infrastructure, coupled with access to data, models, and tools can optimize workflow integration for discovery, sharing, and integration.

Dedicated funding to expand the advanced infrastructure needed for the bioeconomy will ensure that these resources and capabilities scale with the technological advances and needs of the user communities. These investments will leverage and expand ongoing computing, data, and networking infrastructure and services investments across many agencies, such as the NSF ACCESS[11] system of advanced computational resources system of advanced computational resources; NIH's STRIDES[12] available to NIH and NIH-funded researchers; and DOE's HPC user facilities at the national labs and high-performance network user facility (ESnet).[13] The investments will also leverage DOE's Integrated Research Infrastructure (IRI) initiative, including the High Performance Data Facility (HPDF) project, which seeks to develop seamless interconnectivity and integration of instrumentation, data, and computing for federal R&D activities. Increasing accessibility of all computational and data resources will create opportunities for a wide range of users, not just resource-rich institutions.

The aggregation, harmonization, and analysis of existing data can lead to novel, productive and profitable advances. As such, biodata synthesis centers should be developed, building on the model of NSF synthesis centers. This model invests in a form of scientific organization that catalyzes and supports research that integrates diverse theories, methods, and data across sectors, spatial or temporal scales to increase the generality, parsimony, applicability, or empirical soundness of

[30] NSTC Subcommittee on Open Science. *Desirable Characteristics of Data Repositories for Federally Funded Research*. 18 May 2022, https://doi.org/10.5479/10088/113528.

[31] Martonosi, Margaret. *Supporting Computing & Networking Research for a National Discovery Cloud for Climate (NDC-C)*. National Science Foundation, 9 May 2023, www.nsf.gov/pubs/2023/nsf23101/nsf23101.jsp.

scientific explanation.[32] Synthesis Centers are dedicated to facilitating synthesis of available data by multidisciplinary research teams to address compelling scientific questions. These new centers could be developed with cross-agency support, as appropriate. The centers would also provide needed education and workforce training in data science, team science, computation, biological science, and related subjects.

Given the rapid pace of technology innovation, the Data for the Bioeconomy Initiative must include support to leverage and incorporate new technical approaches, data models and system architectures, and needed resources into the bioeconomy data ecosystem. Infrastructure as a Service, Software as a Service, and Platform as a Service approaches could create efficiently shared resources and improve prospects for re-use and discovery. Identifying, incorporating, and leveraging such tools could be accomplished through public-private partnerships, Small Business Innovation Research (SBIR)/Small Business Technology Transfer (STTR) funding, or other means.

Needs addressed: Long-term infrastructure funding mechanisms and support for advancement; Availability, accessibility, and interoperability of data; Sequencing capacity; Lifecycle Security; Data-agnostic infrastructure capable of controlling access for sensitive data

Standards

There is a need to enhance efforts to make data FAIR across domains, removing unnecessary siloes that limit data sharing, integration, and utilization especially as more data from federal agencies and federally funded research becomes publicly accessible.[5,7] Increased uptake and harmonization of standards as well as human and machine-assisted annotation and curation are critical to achieve this.

The Biodata Interagency Working Group (Biodata IWG) described below should coordinate the development of a holistic strategy for the advancement of biological data and metadata standards. This should include: (1) the community coordination needed to develop and adopt various new and existing standards and best practices to conform to FAIR principles, and (2) the creation and maintenance of a compendium of standards that will increase awareness and accessibility for the data-generating and data-using communities. The Biodata IWG should establish or collaborate with and participate in existing community-driven standards bodies that include representatives from across the biological and computer sciences. The IWG should have regular public meetings to obtain community input from academia, government, industry, and non-profits, should consider privacy, Tribal sovereignty, ethical, and security concerns, and should have dedicated resources to sustain and evolve their efforts over time.

Federal representatives should work within the community-driven standards bodies to assess existing international standards for biotechnology and biomanufacturing, such as International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) standards, including ISO

[32] Hackett, Edward J., et al. "Do Synthesis Centers Synthesize? A Semantic Analysis of Topical Diversity in Research." *Research Policy*, vol. 50, no. 1, Jan. 2021, p. 104069, https://doi.org/10.1016/j.respol.2020.104069.

20691:2022 — Biotechnology,[33] and others, and engage other international groups.[34] The Biodata IWG and standards bodies should also engage less formal focused efforts in standards development such as all-volunteer discipline-centric groups and groups that are recipients of short-term agency funding to work with standards. The Biodata IWG and community-driven standards bodies should consider the wide breadth of data types needed for the bioeconomy, including established and newly emerging data types, such as 3D and time-series images. Consideration should also be given to how to best utilize advanced computing and emerging technologies for standards and vice-versa.[35] Additionally, efforts should be taken to expand automated curation of data and screening for quality metrics.

The Biodata IWG and community-driven standards bodies should also support the creation of methods for recognizing and crediting data generators when their data are used, in an effort to incentivize greater data sharing. Tools for sharing and communicating standards, including the engagement of journal publishers as a means of communication, should be leveraged to ensure that the research community remains informed about the development of standards and benefits they offer.

These efforts will capitalize on guidance from OSTP to federal agencies on public access policies, including the Memorandum for the Heads of Executive Departments and Agencies on Increasing Access to the Results of Federally Funded Scientific Research[7] and the Memorandum on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research.[5] This effort should be conducted in close coordination and consultation with the NSTC Subcommittee on Open Science,[36] as well as the Federal Chief Data Officers Council and the Federal Committee on Statistical Methodology.

Needs addressed: Integration; Linkage; Quality; Interoperability; Generation; Predictive capabilities; Societal and human behavioral implications

Biodata catalog

Making data findable is the first step in making that data FAIR. To achieve this, the data should, to the extent possible, be given a unique digital persistent identifier and further described with metadata that covers a range of attributes. Both data and metadata should be registered and indexed in a searchable persistent resource. To ensure the maximum value of such a catalog, it is important that steps are taken to ensure data indexed in the catalog are well curated and publicly accessible to the

[33] See ISO 20691:2022 — Biotechnology Requirements for data formatting and description in the life sciences: https://www.iso.org/obp/ui/#iso:std:iso:20691:ed-1:v1:en

[34] Groups to be engaged may include the International Nucleotide Sequence Data Collaboration (https://www.insdc.org/), Global Alliance for Genomics and Health (https://www.ga4gh.org/), International Nucleotide Sequence Data Collaboration (https://www.insdc.org/), Public Health Alliance for Genomic Epidemiology (https://pha4ge.org/), Research Data Alliance (https://www.rd-alliance.org/), Committee on Data International Science Council (https://codata.org/), GO FAIR (https://www.go-fair.org/), and Standards Coordinating Body on Regenerative Medicine (https://www.standardscoordinatingbody.org/).

[35] The White House. *United States Government National Standards Strategy for Critical and Emerging Technology*. May 2023, www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf.

[36] See National Science and Technology Council Subcommittee on Open Science charter: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-SOS-NSTC-CHARTER.pdf

extent possible, and safe from external manipulation and control, damage, or access denial. Both data and metadata should adhere to established standards and formatting and be accessible in a sustained resource for storage, handling, integration, analysis, and facilitate ethical usage and data privacy and security policies. Associated metadata should be shared as broadly as possible, especially for data that have access controlled, consistent with guidance from OSTP to federal agencies on public access policies for federally funded research.[5,30]

Especially as agencies implement OSTP's public access guidance and more data becomes available, having a centralized, searchable, and accessible listing of extant resources and metadata will maximize the return on federal research investment. This resource will automatically index federal and federally supported data with the appropriate associated metadata, including persistent dataset identifiers. It could also index shared commercial data and associated metadata. This national resource will enable researchers and other key communities to find useful existing data for the bioeconomy and can also foster actions to provide credit to those who generated the data. In time, functionality provided by supporting computational resources could be added to allow for analysis across multiple data sources simultaneously.

All components of a biodata catalog require persistent funding. It also relies on agency, academic, and industry data sharing policies that make data shareable through existing repositories. The utility of these resources can be maximized by integration with computational infrastructure.

Needs addressed: Index of open science data and analysis; Integration; Linkage; Quality; Interoperability; Generation; Predictive capabilities; Societal and human behavioral implications

Security

Investments and adoption of best practices in cybersecurity and trustworthiness of these infrastructures, including authentication systems that enable controlled access to data, will be crucial to safely support the bioeconomy. While the U.S. government is committed to making data as open as possible and data sharing is a key tool, sharing certain classes of data may be limited by legal and policy requirements such as privacy and respect for consent, Tribal sovereignty, and the need to protect national security. Additional security issues and risks include those associated with data integrity, data management, dual use potential, disclosure, aggregation, reidentification, manipulation, exfiltration, and deletion. Monitoring and mitigation of these risks requires concerted effort across the data lifecycle from generation through use and application.

The Biodata IWG mentioned above and described in detail below should remain up to date on guidance and policies from U.S. government agencies and outside organizations.[37] Federal agencies

[37] Guidance and policies from U.S. government agencies include: Circular A-130 on Managing Information as a Strategic Resource (OMB, https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf); Acquisition & Assistance Policy Directive 16-02 (U.S. Agency for International Development, https://www.usaid.gov/sites/default/files/2022-11/AAPD16-02-Revision3.pdf); How-to Note: Conduct a Data Quality Assessment (U.S. Agency for International

such as NIST can recommend security requirements for federal databases and update the coordination body on these recommendations and should take into account concerns and needs of different communities and industries. There are additional considerations for security standards based on national security. Stress tests for federal databases and databases that store federally funded research data should be considered. Additional information and guidance may be included in reports issued in response to Sections 4b (cybersecurity best practices for biological data stored on federal government information systems), 4c (bio-related software), and 9 (Reducing Risk by Advancing Biosafety and Biosecurity) of the Executive Order.[1]

Needs addressed: Lifecycle security; Data-agnostic infrastructure capable of controlling access for sensitive data

Workforce

Scientific research data has grown rapidly in volume and diversity of data type, and as a result, researchers work with increasingly large and complex datasets. Researchers across many scientific disciplines will need substantial training resources to adapt and learn how to utilize advanced computational tools. This need was identified in the State of Data Science Workforce Development,[38] as well as recent National Academies of Science, Engineering, and Medicine studies.[39] Researchers must be educated and empowered to do so through incentives or guidelines on how to responsibly share their data so that it is FAIR and used appropriately by the research community and others who may have access.

As such, funding for and coordination of education and training opportunities should be considered. Advice and guidance on how this could best be accomplished could be led by a Biodata Education and Training working group of the Biodata IWG described below and should be in line with the Biotechnology and Biomanufacturing Workforce report written in fulfillment of Section 7a of the Executive Order.[1] Efforts of this working group should align with the relevant goals and recommendations outlined in the June 2023 Report on Building the Bioworkforce of the Future to

Development, https://usaidlearninglab.org/sites/default/files/resource/files/how-to_note_-_conduct_a_dqa-final2021.pdf); NIST Cybersecurity Framework (https://www.nist.gov/cyberframework); NIST Privacy Framework (https://www.nist.gov/privacy-framework/privacy-framework); NIST Privacy Framework Guidelines and Tools (https://www.nist.gov/privacy-framework/resource-repository/browse/guidelines-and-tools); Data Protection Toolkit (Federal Committee on Statistical Methodology, https://nces.ed.gov/fcsm/dpt); Designating Scientific Data for Controlled Access (NIH, https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/designating-scientific-data-for-controlled-access); The Standard Application Process (Interagency Council on Statistical Policy, https://ncses.nsf.gov/about/standard-application-process); NSTC Desirable Characteristics of Data Repositories for Federally Funded Research (https://doi.org/10.5479/10088/113528); and the Office of the Director of National Intelligence-sponsored report Safeguarding the Bioeconomy (National Academy of Sciences, Engineering, and Medicine, https://doi.org/10.17226/25525).

[38] Federer, Lisa, et al. *Report to the NLM Director: The State of Data Science Workforce Development*. 8 Jan. 2018, www.nlm.nih.gov/pubs/reports/state_of_data_science_training_report_final2.pdf.

[39] National Academies of Sciences, Engineering, and Medicine. *Safeguarding the Bioeconomy*. National Academies Press, 2020, https://doi.org/10.17226/25525.

coordinate and use relevant federal education and training programs, as well as recommend new efforts to promote multidisciplinary education programs.[26]

Training in data processing, analysis, management, and sharing should leverage existing programs, include public-private partnerships to enhance transition to industry positions, engage a diverse set of academic institutions, and include investment in expansion of course-based research experiences with a bioeconomy focus will be used to promote data literacy and integration of bioeconomy concepts. Similarly, Research Experiences for Undergraduates (REUs) could address both diversity and adoption of new technologies. Such courses will be located across institutional types, from major universities to community colleges. Trainees will use discovery and synthesis research tools to identify or evaluate prospects for contributing to the bioeconomy.

An emphasis on diversity, equity, and inclusion will be foundational for all investments to avoid unintentional amplification of implicit biases in our computational infrastructure and tools and to build a representative workforce that benefits from diverse perspectives and backgrounds.[40] Developing and providing additional education resources, including enhanced curricula for development of AI and machine learning modeling expertise, will allow researchers and data processing and management staff to develop new capabilities to analyze and visualize data allowing researchers to extract the most information from data generated and understand best practices in data management to ensure quality and rigor of data. Leveraging such tools could be accomplished through public-private partnerships, SBIR/STTR support, or other means.

For the federal workforce, additional hiring authorities, building on the recent efforts such as the special salary rate for federal information technology (IT) employees,[41] may need to be created to achieve the necessary talent mix. Consideration should also be given to an expansion of the U.S. Digital Service or establishment of a biodata corps.

Needs addressed: Workforce development

Strategically targeted areas for rapid transformation (STARTs)

To pilot implementation of the actions recommended in this report, the Biodata IWG described below should consider tractable priority projects in at least one of the following areas, each of which is an identified data gap: (1) biodata corps, (2) bioreactor fermentation, (3) non-human genome sequencing, (4) digitizing biobank collections, (5) multi-omic characterization of non-model organisms and non-traditional host cells, and (6) image phenotyping. Federal agencies could also independently choose to embark on one of these tractable projects. Many of these areas currently exist as early-stage work at agencies, presenting readily available opportunities for increased whole-of-government coordination and investment to advance. As highlighted above, each of these potential foci will require investment in the proposed infrastructure actions to be successful and provide a springboard for addressing the larger bioeconomy data landscape. In addition to building

[40] Marinucci, Ludovica, et al. “Exposing Implicit Biases and Stereotypes in Human and Artificial Intelligence: State of the Art and Challenges with a Focus on Gender.” *AI & Society*, vol. 38, May 2022, https://doi.org/10.1007/s00146-022-01474-3.

[41] See: https://www.opm.gov/special-rates/2023/IndexByOccupations.aspx

on existing resources and initiatives, they will also require planned and dedicated funding, whether they are advanced by the Biodata IWG described below or by a single federal agency or group of agencies.

None of these projects on their own will achieve the vision above, but focusing on any or several of these areas will enable community engagement and coalescence, development of best practices, and understanding of scope of need that can be used to expand the broader range of bioeconomic data.

The short-term projects (e.g., three years) on the chosen STARTs, should enable understanding of best practices in infrastructure development, interagency coordination, community engagement, incentives, standards creation and dissemination, and workforce development. These insights can be applied to other STARTs and the broader data for the bioeconomy ecosystem (see Figure 3).

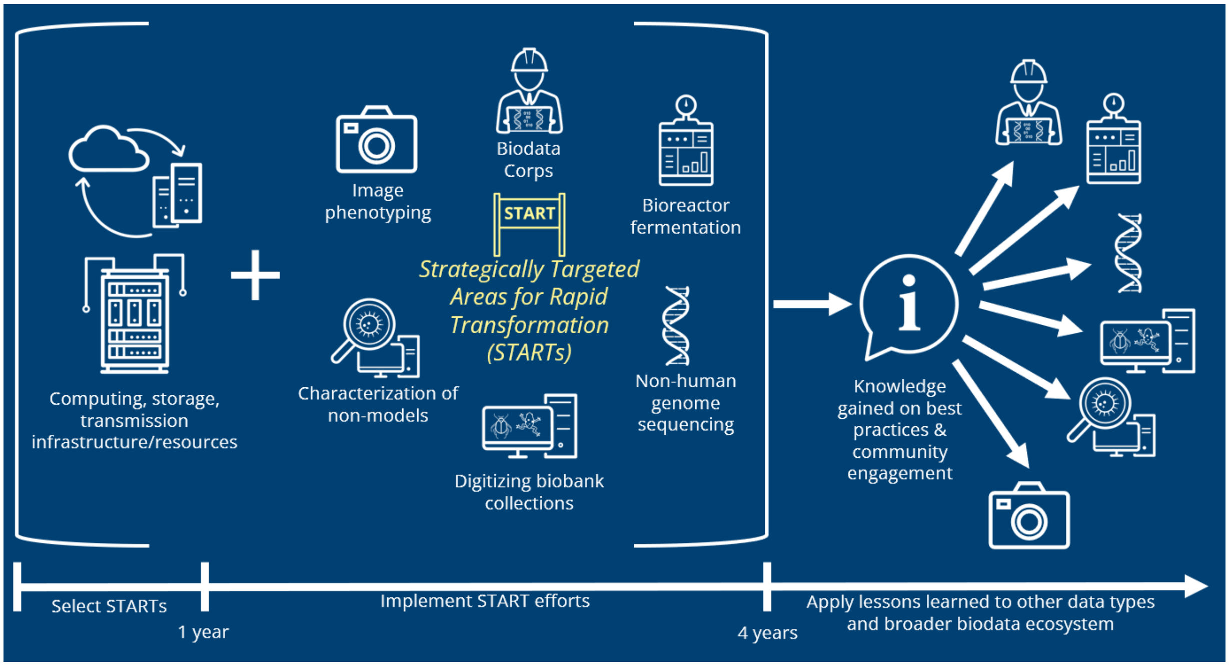

Computing, storage, transmission infrastructure/resources
+
Image phenotyping
Biodata Corps
START
Bioreactor fermentation
Strategically Targeted Areas for Rapid Transformation (STARTs)
Characterization of non-models
Non-human genome sequencing
Digitizing biobank collections
Knowledge gained on best practices & community engagement
Select STARTs
1 year
Implement START efforts
4 years
Apply lessons learned to other data types and broader biodata ecosystem

Figure 2. Graphical representation of the timeline for selecting and implementing STARTs and applying lessons learned to other parts of the biodata ecosystem.

Biodata corps

All of the Bioeconomy Initiative goals depend on the availability of a data-knowledgeable workforce that can apply advanced data science methods to organize, analyze, visualize, and provide reports about lessons learned from the bioeconomy data. The increased demand for data scientists also creates an opportunity to broaden participation and diversify the bioeconomy participants. It also presents a chance to enhance connections between data generators and data users. The biodata corps would provide federal funding for training efforts at all academic levels and across institutional types, including community colleges, primarily undergraduate institutions, and minority serving institutions. These training efforts would include networking opportunities with data generators to increase understanding of varied aspects of generation and how they impact usability and vice versa.

Bioreactor fermentation

Biotechnology and biomanufacturing efforts in federal facilities, academia, and industry generate and store vast amounts of data on each fermentation batch, but these data are often not synthesized or reported. Such data could contain valuable insights to improve fermentation conditions and practices. Anonymized sharing of these data could be used to train machine learning models, AI controllers, and other next generation biomanufacturing applications. Improved use of sensors and online sampling could further expand the available data. The federal government could also support standard fermentation runs and make the data publicly available as a training set for new companies. These efforts could tap into innovation institutes such as BioMADE and the Advanced Biofuels and Bioproducts Process Development Unit and should be in coordination with the efforts relating to Building a Vibrant Domestic Biomanufacturing Ecosystem under Section 5 of the Executive Order.[1]

Non-human genome sequencing

In line with the cross-cutting Bold Goal to leverage biodiversity to create new products for the bioeconomy and the wealth of biodiversity among eukaryotic organisms,[2] efforts to generate annotated genome sequences for organisms with high potential for creating new products for the bioeconomy should be undertaken. The initiative would require enhanced specimen collection and data storage, sequencing capability, interoperability and standards, and cyberinfrastructure and computing resources to enable connections between various partners, including international, and analysis and use of the data at scale. Tapping into this biodiversity could yield insights for all sectors of the bioeconomy. For example, exploring non-crop plant species, especially from marginal lands could be used to enhance biomass production for a viable and sustainable biomass, biofuels, and bioproducts industry.

Digitizing biobank collections

Museums, universities, and other institutions across the nation hold large collections of biological samples. Though significant strides have been made, many of these collections remain undigitized and unknown to researchers outside of academia. Imaging and digitizing these collections, using the appropriate metadata standards and inclusion of specimen data and incorporating Tribal guidance and perspectives when appropriate, would give scientists all over the country access to these data, increasing speed and efficiency of research by ensuring that existing data are not unnecessarily re-collected. Such an effort could entail infrastructure connections between collections and/or connections to other sorts of data such as phenotypic data, genomic sequences, and multi-omic characterizations.

Ensuring that the data are FAIR and use appropriate standards such as Darwin Core[42] and guidelines such as ISO 8000 — Data Quality[43] can support interoperability of these resources and enhance their utility; incentivizing such enhancements could be required as part of grant processes. Democratizing access to these data and integrated analytics could enable studies of human impact on species

[42] See Darwin Core: https://dwc.tdwg.org/

[43] See ISO 8000: Data Quality: https://www.iso.org/standard/81745.html

diversity, such as impacts of overfishing, or trends in insects that provide important ecosystem services, such as pollinating crop plants, or those that damage and destroy crops.

Multi-omic characterization of non-model organisms and non-traditional host cells

While model organisms, like mice, rats, and zebra fish, and traditional cell lines like Chinese hamster ovary cells and human embryonic kidney 293 are thoroughly characterized in the health sector and organisms such as *Saccharomyces cerevisiae* and *Escherichia coli* in the biomanufacturing sector, multi-omic data on non-model organisms and non-traditional cell lines are lacking. Building out multi-omic characterization of a wider variety of cell lines, animals, plants, and microbial production hosts could, for example, lead to a bridge between orphan crops and modern agriculture and aquaculture and provide more robust hosts to meet the demand for renewable and sustainable chemical production. The new data could also lead to new materials, pathways, and products for rapid translation to the marketplace.

Image phenotyping

The analysis of multicellular and single-celled organisms and integrating both genomic and environmental impacts on an organism are critical to predict outcomes of genetic manipulation or understand patterns in evolution and other complex biological processes. Traits are often visible and can be characterized by the use of 3D imaging. In this area, standards, including those for phylogenies coupled with image data, could provide the information needed to develop basic analytics to identify the key outcomes through knowledge guided machine learning. The work can help classify species, describe complex phenotypes, and support fields as diverse as agronomy and conservation. Recently, single cell analytics has uncovered disparate single cell behavior in fermentations of microorganisms to produce chemical and fuel products, when these cultivations were previously only understood at the bulk level. 3D imaging with sub-single cell resolution further builds on single cell analytics and will provide the essential data to inform improved engineered systems with performance that can meet the demand for biomanufactured products.

As images are increasingly 3D and their generation is enabled by remote sensing methods, efforts to advance such methods, data storage capacity, and connections to computing resources will be necessary. End user and community engagement will also be critical to ensure collection of images such users need for their efforts.

Needs addressed: Integration; Linkage; Quality; Interoperability; Generation; Predictive capabilities; Societal and human behavioral implications; Lifecycle Security

Coordination of intergovernmental investments, efforts, and resources

These and future actions of the Data Initiative will require continuous coordinated engagement from the whole of the federal government in the short- and long-term and will best serve the American bioeconomy if informed by a wide array of key communities and experts. As such, to facilitate these actions, a Biodata Interagency Working Group (Biodata IWG) or similar body should continue to exist.

The Biodata IWG will oversee the proposed Data Initiative programs and policies, advise on establishment of future programs, policies, and data governance approaches. To guide its work, the group should develop a timeline for action to implement the recommendations outlined in this report and identify any federal resources and legal authorities needed to advance its efforts. Avoiding duplication of efforts and coordinating prioritization of activities will maximize use of human capital and funds. The Biodata IWG should be led by key departments and agencies and should be supported by NBBI and/or OSTP staff.

A body of external experts should also be organized to better inform and advise on these efforts. Capturing the breadth of the biodata landscape will be crucial and, as such, this body should include leaders spanning academia and industry (startups, midsize and large companies); expertise from across the biological sciences, computer and information sciences, and ESLI; expertise across the breadth of the data lifecycle, including data security; and data generators and data users. This external expert body could be convened, for example, under the Federal Advisory Committee Act at a participating department or agency.

Needs addressed: Long-term infrastructure funding mechanisms and support for advancement; Availability, accessibility, and interoperability of data; Sequencing capacity; Index of open science data and analysis; Workforce development; Integration; Linkage; Quality; Interoperability; Generation; Predictive capabilities; Societal and human behavioral implications; Lifecycle Security; Data-agnostic infrastructure capable of controlling access for sensitive data

Conclusion

The bioeconomy drives the wider economy and is a prerequisite for global leadership in the 21st century. It has and will continue to advance the ability to solve grand challenges facing the planet in climate, clean energy, food and agriculture, supply chain, and health. Data coupled with computational infrastructure are the fuel that will allow solutions to these grand challenges and bold goals for the bioeconomy. However, the foundational infrastructure upon which the bioeconomy is currently built requires significant investment to realize these possibilities while mitigating the hazards. This report outlines the current landscape of data and infrastructure supporting the bioeconomy and highlights the current needs and proposed actions to build a robust foundation to fuel the bioeconomy. These actions include coordination around and investments in persistent data and computing infrastructure, a comprehensive biodata catalog, data standards, data security, a data workforce, infrastructure coordination via federal and community oversight, and potential application via strategically targeted areas for rapid transformation.

A whole-of-government approach is necessary to achieve the Bold Goals[2] for the bioeconomy, maintain global leadership in research and innovation, and provide rapid responses to dynamic global needs, including pandemics, famine, and geopolitical conflict.

Appendix A: Existing Data Sources by Sector

This contains a list of some of the existing federally created, supported, or maininted data sources, nearly all of which are fragile as they are funded for relatively short periods of time before competitive renewal or the end of grant periods. Consequently, these data sources, however presently valuable, cannot currently be counted on for the long-term realization of a thriving bioeconomy without the sustaining investments proposed in the report. Creating long-term funding mechanisms to support these data sources and others like them, as well as providing funding for advancing existing resources, is a critical risk management effort.

Resource listings contain the name of the resource, a short description, and the departments and agencies that support or maintain the resource. Where individual subcomponents provide the support or maintain the resource, they are listed in parenthesis below the agency. This is not an exhaustive list of all federally supported or maintained data resources with relevance to the bioeconomy, but highlights select resources with data critical to the vision laid out in the report.

All the data infrastructure components noted below have significant value-added aspects beyond the bioeconomy including basic biological research and engineering and projects that leverage the computational and analytical pipelines developed by federal investments. As an example, CyVerse's data systems were utilized to combine imagery to create visual evidence of a black holes.[1]

Cross-cutting

Table 1 lists and describes select data resources that support innovation across sectors of the bioeconomy along with the federal agencies that provide funding for those resources.

Table 1. Select data resources that support innovation across the bioeconomy

Resource	Description	Federal Support
Biomining[2]	Program that seeks to develop novel approaches in microbiology, synthetic biology, and process engineering to address mining industry challenges and ensure a robust mineral supply chain for clean energy applications.	DOE (ARPA-E)
CyVerse[3]	National infrastructure for life sciences research.	NSF (Directorate for Biological Sciences)

[1] See "Donuts, Data and DOIs: Using CyVerse to Analyze and Publish the Second Black Hole Image": https://cyverse.org/eht-sagA

[2] See Biotechnologies to Ensure a Robust Supply of Critical Materials for Clean Energy: https://arpa-e.energy.gov/technologies/exploratory-topics/biomining

[3] See CyVerse: https://cyverse.org/

Comparative Genomics Resource (CGR)[4]	An ecosystem to facilitate reliable comparative genomics analyses for all eukaryotic organisms to maximize their impact and their genomic data resources to meet emerging human health research needs.	NIH (NLM)
DOE-funded bioproducts-focused initiatives	Performance Advantaged Bioproducts Consortium[5] generates bioprocess data and microorganism data to improve the ability to identify promising molecules and produce them from biomass and renewable feedstocks.	DOE
DOE Systems Biology Knowledgebase (KBase) Error! Bookmark not defined.	Powerful integrated data science and software platform that enables transparent biological discovery through secure sharing of data, tools, methods, and conclusions in a unified, extensible system and is designed to perform large-scale analyses on scalable computing infrastructure to meet the key challenges of systems biology: predicting and designing biological function to improve the environment and provide sustainable routes for energy production and security.	DOE (Office of Science)
ECOSynBio[6]	Program that aims to promote the use of advanced synthetic biology tools to engineer novel biomass conversion platforms and systems. These systems will be designed to use external energy inputs to substantially increase carbon use, versatility, and efficiency while achieving economies of scale for industrial applications. Successful platforms will offer new capacities for the bioeconomy by enabling fully carbon-optimized renewable fuel and chemical synthesis with maximum carbon and resource efficiency.	DOE (ARPA-E)
GenBank®[7]	GenBank® collects, preserves, and provides public access to assembled and annotated nucleotide sequence data from all domains of life, which are shared among the members of the International Nucleotide Sequence Database Collaboration (INSDC).[8]	NIH (NLM)

[4] See Comparative Genomics Resource: https://www.ncbi.nlm.nih.gov/comparative-genomics-resource/

[5] See Systems Biology Knowledgebase (KBase): https://www.kbase.us/about/

[6] See ECOSynBio: https://arpa-e.energy.gov/technologies/programs/ecosynbio

[7] See GenBank: https://www.ncbi.nlm.nih.gov/genbank/

[8] See International Nucleotide Sequence Database Collaboration (INSDC): http://www.insdc.org/

Harnessing Emissions into Structures Taking Inputs from the Atmosphere (HESTIA)[9]	Supports the development of technologies that cancel out embodied emissions while transforming buildings into net carbon storage structures.	DOE (ARPA-E)
NASA Life Sciences Portal (NLSP)[10]	Contains data collected on astronauts, as well as other supporting studies on human subjects and model organisms. NLSP provides appropriately controlled access to astronaut health data.	NASA (Human Research Program)
NASA Open Science Data Repository (OSDR)[11]	Consists of two interconnected space biological databases. The Ames Life Sciences Data Archive (ALSDA) is the official repository of non-human biology data spanning a broad range of approaches. GeneLab is an open science multi-omics repository.	NASA (Space Biology Program)
National Microbiome Data Collaborative (NMDC)[12]	Creating an integrated, enabling environment for FAIR microbiome data, with core capabilities in metadata standards and standardized bioinformatic workflows for omics analysis.	DOE (Office of Science)
NSF-funded Synthesis Centers	National Evolutionary Synthesis Center,[13] the National Institute for Mathematical and Biological Synthesis,[14] the National Socio-Environmental Synthesis Center,[15] and others, which work to combine and analyze existing data for use in new applications and to further new research.	NSF
NSF-funded Engineering Research Centers	Center for Cell Manufacturing Technologies[16], Center for Advancing Sustainable and Distributed Fertilizer Production,[17] and the Center for Precision Microbiome Engineering,[18] each of which supports convergent research and data creation, education, and technology translation.	NSF (Directorate for Engineering)

[9] See HESTIA: https://arpa-e.energy.gov/technologies/programs/hestia

[10] See NASA Life Sciences Portal: https://nlsp.nasa.gov/

[11] See NASA Open Science Data Repository: https://osdr.nasa.gov/

[12] See National Microbiome Data Collaborative (NMDC): https://microbiomedata.org/

[13] See National Evolutionary Synthesis Center: http://www.nescent.org/

[14] See National Institute for Mathematical and Biological Synthesis: http://www.nimbios.org/

[15] See National Socio-Environmental Synthesis Center: http://www.sesync.org/

[16] See Center for Cell Manufacturing Technologies: https://cellmanufacturingusa.org/

[17] See Center for Advancing Sustainable and Distributed Fertilizer Production: https://nsf.gov/cgi-bin/good-bye?https://www.casfer.us

[18] See Center for Precision Microbiome Engineering: https://nsf.gov/cgi-bin/good-bye?https://premier.pratt.duke.edu/

Biopreparedness Research Virtual Environment (BRaVE)[19]	Efforts building on the success of the national Virtual Biotechnology Laboratory,[20] which was a consortium of DOE National laboratories linking research talents with enabling data platforms, instrumental capabilities, and DOE User Facilities in a virtual format to accelerate and amplify Next-Gen Biology research	DOE (Office of Science)
RCSB Protein Data Bank (PDB)[21]	U.S. data center for the global PDB archive of 3D structure data for large biological molecules essential for research and education in fundamental biology, health, energy, and biotechnology.	DOE, NIH, NSF
Sequence Read Archive (SRA)[22]	Largest publicly available repository of high throughput sequencing data which accepts data from all branches of life as well as metagenomic and environmental surveys, and stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries.	NIH (NLM)
Office of Science and Technical Information (OSTI)[23]	A unit of the Office of Science that fulfills agency-wide responsibilities to collect, preserve, and disseminate both unclassified and classified scientific and technical information (STI) emanating from DOE-funded research and development activities at DOE national laboratories and facilities and at universities and other institutions nationwide. OSTI provides access to DOE STI through a suite of web-based, searchable discovery tools and through other commonly used search engines.	DOE (Office of Science)

Health

Table 2 lists and describes select data resources primarily within health and medicine along with the federal agencies that provide funding for those resources.

Table 2. Select data resources that support innovation primarily within health and medicine

Resource	Description	Federal Support
***All of Us* Research Program**[24]	Effort to advance individualized health care by enrolling one million or more participants to contribute their health data over many years.	NIH

[19] See Biopreparedness Research Virtual Environment (BRaVE): https://science.osti.gov/Initiatives/Biopreparedness
[20] See National Virtual Biotechnology Laboratory (NVBL): https://science.osti.gov/nvbl
[21] See RCSB Protein Data Bank: https://www.rcsb.org/
[22] See Sequence Read Archive (SRA): https://www.ncbi.nlm.nih.gov/sra
[23] See Office of Science and Technical Information: https://www.osti.gov/
[24] See *All of Us* Research Program: https://allofus.nih.gov/

Cancer Research Data Commons[25]	Cloud-based data science infrastructure that connects datasets with analytics tools to allow users to share, integrate, analyze, and visualize cancer research data to drive scientific discovery.	NIH (NCI)
Center for Health Services Research (CHSR)[26]	Provides access to the Military Health Service (MHS) Data Repository and other MHS healthcare databases that can enable longitudinal studies associating risk factors with long-term outcomes as well as system-wide studies of cost drivers in healthcare.	DOD (USUHS)
Centralized Interactive Phenomics Resource (CIPHER)[27]	Library of curated phenotypes and associated metadata from Department of Veterans Affairs electronic health records.	VA (Office of Research and Development)
Database of Genotypes and Phenotypes (dbGaP)[28]	The dbGaP repository is the most comprehensive, free, NIH-supported controlled access archive of human genotype and phenotype data which was generated from a broad array of research studies exploring diverse human health questions.	NIH (NLM)
Genome in a Bottle[29]	Public-private-academic consortium hosted by NIST to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice and innovations in technologies.	NIST (Material Measurement Laboratory)
Million Veteran Program[30]	National research program looking at how genes, lifestyle, military experiences, and exposures affect health and wellness in veterans that currently hosts deidentified genotype and curated electronic phenotype data on the VA Data Commons platform. Pilot testing of broader access will begin in FY24.	VA (Office of Research and Development)

[25] See Cancer Research Data Commons: https://datascience.cancer.gov/data-commons

[26] See Center for Health Services Research (CHSR): https://chsr.usuhs.edu/

[27] See Centralized Interactive Phenomics Resource (CIPHER): https://www.research.va.gov/programs/cipher.cfm

[28] See Database of Genotypes and Phenotypes (dbGaP): https://www.ncbi.nlm.nih.gov/gap/

[29] See Genome in a Bottle: https://www.nist.gov/programs-projects/genome-bottle

[30] See Million Veteran Program: https://www.research.va.gov/mvp/

Food and agriculture

Table 3 lists and describes select data resources primarily within food and agriculture along with the federal agencies that provide funding for those resources.

Table 3. Select data resources that support innovation primarily within food and agriculture

Resource	Description	Federal Support
Ag100Pest[31]	Sequencing genomes of at least one hundred arthropod pests as part of the Sequencing Five Thousand Arthropod Genomes Initiative (i5K) and Earth BioGenome Project.	USDA (ARS)
Ag Data Commons[32]	Catalog and repository with openly available data from USDA-funded research.	USDA (National Agricultural Library)
Breeding Insight[33]	Comprised of tools and databases for assembling and analyzing genomic and phenomic data for breeders of alfalfa, blueberry, table grape, sweet potato, rainbow trout and North American Atlantic salmon, cranberry, cucumber, honeybee, lettuce, oat, pecan, and strawberry with additional organisms added biannually.	USDA (ARS)
Economic Research Service Data Products[34]	Catalog of openly available economic data from USDA agencies, on such topics as food security, international trade, farming practices, and bioenergy.	USDA (Economic Research Service)
FoodData Central[35]	Provides nutritional profiles of food, including foundational and standard reference foods, branded products, and experimental foods.	USDA (ARS)
GenomeTrackr Network[36]	Distributed network of public health and university laboratories to utilize whole genome sequencing for foodborne pathogen identification with data housed in public databases at the National Center for Biotechnology Information.	FDA
Germplasm Resources Information Network (GRIN)[37]	Provides information about USDA national collections of animal, microbial, and plant genetic resources (germplasm) important for food and agricultural production and biomanufacturing.	USDA (ARS)

[31] See Ag100Pest: http://i5k.github.io/ag100pest

[32] See Ag Data Commons: https://data.nal.usda.gov/

[33] See Breeding Insight: https://breedinginsight.org/

[34] See Economic Research Service Data Products: https://www.ers.usda.gov/data-products/

[35] See Food Data Central: https://fdc.nal.usda.gov/

[36] See GenomeTrackr Network: https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network

[37] See Germplasm Resources Information Network (GRIN): https://www.ars-grin.gov/

Long-Term Agroecosystem (LTAR) Network[38]	Eighteen established, long-term research sites focused on developing national strategies for more efficient agricultural production while improving the quality of the environment and the well-being of America's farming communities	USDA (ARS)
SMARTFARM[39]	Bridges the data gap in the biofuel supply chain by funding technologies that can quantify feedstock-related emissions at the field-level and enable new market incentives for efficiency in feedstock production and carbon management.	DOE (ARPA-E)
SCINet[40]	An effort by ARS to grow USDA's research capacity by providing scientists with access to high-performance computing clusters, high-speed networking for data transfer, and training in scientific computing.	USDA (ARS)

Environment

Table 4 lists and describes select data resources primarily supporting efforts on the environment and climate along with the federal agencies that provide funding for those resources.

Table 4. Select data resources that support innovation primarily within the environment and climate

Resource	Description	Federal Support
AmeriFlux[41]	Network of sites managed by primary investigators that measure ecosystem carbon dioxide, water, and energy fluxes in North, Central and South America established to connect research on field sites representing major climate and ecological biomes, including tundra, grasslands, savanna, crops, and conifer, deciduous, and tropical forests.	DOE (Office of Science)
Earth Data[42]	Full and open access to data from NASA's Earth Science Data Systems Program, including on topics within atmosphere, biosphere, cryosphere, human dimensions, land surface, ocean, solid earth, sun-Earth interactions, and terrestrial hydrosphere.	NASA (Earth Science Data Systems Program)
Fine-Root Ecology Database (FRED)[43]	More than 150,000 observations of more than 330 root traits, with data collected from more than 1400 data sources, including information on root anatomy, morphology, microbial associations, ancillary data on associated site, vegetation, edaphic, and climatic conditions.	DOE (Office of Science)

[38] See Long-Term Agroecosystem Research (LTAR) Network: https://ltar.ars.usda.gov/network

[39] See SMARTFARM: https://arpa-e.energy.gov/technologies/programs/smartfarm

[40] See SCINet: https://scinet.usda.gov/

[41] See AmeriFlux: https://ameriflux.lbl.gov/

[42] See Earth Data: https://www.earthdata.nasa.gov/

[43] See Fine-Root Ecology Database (FRED): https://roots.ornl.gov/

Forest Inventory Data & Tools[44]	Catalog of openly available data from the USDA Forest Service including data and geospatial information about forests across the United States.	USDA (Forest Service)
International Tree-Ring Data Bank (ITRDB)[45]	Data on raw ring width, wood density, isotope measurements, and site growth index chronologies from more than 5,000 sites on six continents and reconstructed climate parameters for some areas.	NOAA (National Centers for Environmental Information)
Landsat[46]	Data from the joint U.S. Geological Survey/NASA Landsat series of Earth Observation satellites providing uninterrupted data to help land managers and policymakers make informed decisions about natural resources and the environment.	USGS, NASA
Long Term Ecological Research (LTER) Network[47]	Long-term analysis of locally important variables (e.g., permafrost depth in the Arctic, neighborhood income for urban sites) that is made freely available and used and reused many times over, often to answer unexpected questions years after collection.	NSF (Directorates for Biological Sciences; Geosciences; and Social, Behavioral and Economic Sciences)
Marine Biodiversity Observation Network (MBON)[48]	Network integrating various sciences and themes across diverse geographies and stakeholders to collect, curate, analyze, manage, and communicate marine biodiversity data and related services to improve understanding of changes and connections between marine biodiversity and ecosystem functions.	NOAA, NASA, DOI (Bureau of Ocean Energy Management), and Office of Naval Research (ONR)
National Ecological Observatory Network (NEON)[49]	Network of 81 field sites across geographic and environmentally distinct areas that generate standardized and quality-controlled data products, including meteorological, soil, organismal, biogeochemical, and freshwater aquatic data using automated instruments, observational field sampling, and airborne remote sensing surveys.	NSF (Directorate for Biological Sciences)

[44] See Forest Inventory Data & Tools: https://www.fs.usda.gov/research/products/dataandtools/forestinventorydata

[45] See International Tree-Ring Data Bank: https://www.ncei.noaa.gov/products/paleoclimatology/tree-ring

[46] See Landsat: https://www.usgs.gov/landsat-missions

[47] See Long Term Ecological Research (LTER) Network: http://www.lterner.edu/

[48] See Marine Biodiversity Observation Network (MBON): https://marinebon.org/

[49] See National Ecological Observatory Network (NEON): https://www.neonscience.org/

Biomanufacturing

Table 5 lists and describes select data resources primarily within biomanufacturing along with the federal agencies that provide funding for those resources.

Table 5. Select data resources that support innovation primarily within biomanufacturing

Resource	Description	Federal Support
Advanced Biofuels and Bioproducts Process Development Unit (ABPDU)[50]	Facility that provides pilot scale process development resources to accelerate the deployment of biomanufacturing. Generates a wide range of bioprocess data and contributes to R&D on the science of scale-up.	DOE (Office of Energy Efficiency and Renewable Energy, Bioenergy Technologies Office)
Agile BioFoundry[51]	Develops the nation's synthetic biology infrastructure for production of renewable fuels and chemicals and collects data including multi-omics, sequencing, cultivation, strain engineering, fermenter performance data, and high throughput analytics for small molecules.	DOE (Office of Energy Efficiency and Renewable Energy, Bioenergy Technologies Office)
BioFabUSA,[52] **BioIndustrial Manufacturing & Design Ecosystem (BioMADE),**[53] **and National Institute for Innovation in Manufacturing Biopharmaceuticals (NIIMBL)**[54]	Manufacturing Innovation Institutes (MIIs) consisting of large-scale public-private collaborations that conduct advanced research and development, grow manufacturing ecosystems, and further education and workforce development.	DOD (BioFabUSA and BioMADE), DOC (NIIMBL)
BOTTLE (Bio-Optimized Technologies to keep Thermoplastics out of Landfills and the Environment)[55]	Consortium that conducts high-impact research and development on improved catalytic and biocatalytic recycling strategies to break down today's plastics into chemical building blocks for manufacturing higher-value products (upcycling) and the design of tomorrow's plastics to be recyclable-by-design, generating data related to polymers and processes around plastic recycling.	DOE (Office of Energy Efficiency and Renewable Energy, Bioenergy Technologies Office and Advanced Materials & Manufacturing Technologies Office)

[50] See Advanced Biofuels and Bioproducts Process Development Unit (ABPDU): https://abpdu.lbl.gov/

[51] See Agile BioFoundry: https://agilebiofoundry.org/

[52] See BioFabUSA: https://www.armiusa.org/biofabusa/

[53] See BioIndustrial Manufacturing & Design Ecosystem (BioMADE): https://www.biomade.org/

[54] See National Institute for Innovation in Manufacturing Biopharmaceuticals (NIIMBL): https://niimbl.my.site.com/s/

[55] See BOTTLE (Bio-Optimized Technologies to keep Thermoplastics out of Landfills and the Environment): https://www.bottle.org/about.html

Experimental Data Depot (EDD)[56]	EDD is a synthetic biology data repository designed to collect bioprocess, fermentation, and multi-omics data to feed into computational pipelines and AI and machine learning platforms.	DOE (Office of Energy Efficiency and Renewable Energy, Bioenergy Technologies Office)
Joint Genome Institute (JGI)[57]	User facility focused on advancing genomics in support of DOE priorities in clean energy and environmental remediation containing over 14 PB of high quality genomic and omics data and metadata on plant, algal, fungal, viral, and microbial and metagenomes in its Data Portal.	DOE (Office of Science)
nSoft[58]	Public-private consortium for the advancement of neutron-based measurements for manufacturing of soft materials, such as biodegradable plastics, efficient membranes for energy storage, high strength low density composites for transportation, and protein-based biopharmaceuticals for cancer treatment.	NIST

[56] See Experiment Data Depot: https://agilebiofoundry.org/capabilities/test/experiment-data-depot-edd/

[57] See Joint Genome Institute (JGI): https://jgi.doe.gov/

[58] See nSoft: https://www.nist.gov/nsoft