Securing the U.S. Bioeconomy: A Managed Access Framework for Biotechnology Innovation

Jon Judd, EBRC Science Policy Postdoctoral Fellow

May 2026

The U.S. bioeconomy is a growing sector driving technological innovation and global competitiveness. This innovation relies on diverse biological information, spanning genomic sequences, molecular profiles, and functional data across human populations, crop and livestock species, and pathogens. While the U.S. is a leader in producing this data, it is highly siloed and lacks interoperability, undermining scientific innovation, national security, and global competitiveness. To overcome these challenges and build a modern data ecosystem, a new framework for biological data governance is necessary. However, some biological data carries inherent risks if accessed without appropriate safeguards. Therefore, a new biological data framework must prioritize federal managed-access infrastructure to simultaneously: (1) maximize data accessibility for innovation, and (2) secure sensitive information against misuse.

Federal agencies are already mobilizing biological data for AI-driven discovery under initiatives like the Genesis Mission, but without coordinated governance, these efforts risk reproducing the same fragmentation they are meant to solve. The National Security Commission on Emerging Biotechnology (NSCEB) recognized this challenge in its 2025 final report, recommending the creation of a Web of Biological Data (WOBD) as a single access point for federally funded datasets and providing a useful model for the broader managed-access infrastructure the federal government needs now. However, operationalizing this infrastructure to be a secure and viable tool that achieves both goals requires more than aggregating data; it requires a governance architecture that opens access broadly while establishing proportionate controls for sensitive data. **To do this, policymakers must: (1) establish a managed-access infrastructure with a federated data access model to unify fragmented repositories, (2) create a standardized biological data ontology to enable machine-readable interoperability across datasets, and (3) develop a unified data-use agreement to replace the current patchwork of incompatible access requirements.**

These recommendations are complementary: the ontology (Recommendation 2) provides the standardized metadata that makes data interoperable within the infrastructure (Recommendation 1), while the unified data-use agreement (Recommendation 3) establishes the access rules governing who can use it. Implementing this policy will modernize biological data governance and enable a future that maximizes scientific discoveries, improves national security, and reinforces the United States' leading position in biotechnology.

Global Leadership Through Biological Data

The U.S. bioeconomy is an increasingly important facet of the U.S. GDP due to the growing role of biotechnology in economic sectors, including defense, agriculture, energy, and manufacturing, with the total market size of biotechnology expected to be close to $2-4 trillion in 2030-40. However, in order for Americans to see the most benefit from the growing bioeconomy, the U.S. needs to remain a global leader in the responsible development and deployment of biotechnologies.

U.S. biotechnological innovation has been fueled by the collection of massive datasets of biological data, including genetic, molecular, and protein information across human, animal, plant, and microbe specimens (**Box 1: Federal Biological Data and the Case for Interoperability**). This data is particularly crucial for training biological Artificial Intelligence (AI) models that rely on large and diverse datasets to provide novel biological insights and minimize bias in these models. The U.S. federal government has promoted data collection and curation directly and through grants to support large-scale data repositories of individuals, species, and molecules from varied geographic, demographic, and historical backgrounds. This data is used in fields ranging from medical research to manufacturing. However, even though these datasets are individually large, disorganization has led to a lack of interoperability of biological datasets and limited future innovations.

Box 1: Federal Biological Data and the Case for Interoperability

Federal biological data spans dozens of repositories across agencies and data types, each developed for distinct scientific purposes and subject to different regulatory regimes. NIH maintains genomic sequence repositories including GenBank; USDA manages agricultural genomics data across crop and livestock species; DOE operates the Joint Genome Institute, one of the world's largest environmental metagenomic databases. These are illustrative examples of a far larger federal data landscape, not a comprehensive inventory. Each repository carries a distinct risk profile and operates under different oversight requirements: human genomic records are subject to HIPAA and the Genetic Information Nondiscrimination Act; agricultural and environmental datasets typically carry minimal access restrictions. Sequences associated with select agents present biosecurity sensitivities, but the Federal Select Agent Program governs physical possession of those agents, not access to their genetic sequences– many of which remain publicly available in repositories like GenBank.

The UK Biobank demonstrates what integrated biological data governance can produce. By linking genomic sequences, electronic health records, physical measurements, and imaging data from 500,000 participants under a unified access framework, the program has enabled researchers worldwide to generate over 18,000 peer-reviewed publications. The integration of genetic and imaging data has produced significant discoveries regarding the relationship between genetic variants and brain structure and function, particularly in Alzheimer disease and Parkinson disease. The U.S. holds far more raw biological data than the UK Biobank; what it lacks is the governance infrastructure to make that data similarly productive.

This fragmentation stems from several factors: institutional focus on agency-specific data needs, a lack of standardized formats across repositories, and an unresolved tension between maximizing data accessibility and protecting sensitive information such as identifiable human genomic records. Even federal data management mandates, including NSF's requirement for data management plans as a condition of award, have not resolved this fragmentation. The emergence of AI models capable of synthesizing data across repositories creates new opportunities and risks that the original system was never designed to manage, making coordinated access controls necessary for the first time. This is particularly crucial to ensure that the U.S. remains innovative in the biotechnology space and so that we avoid a national security data liability. **A new framework for biological data governance that addresses these factors is necessary to build a modern data ecosystem.**

Managed-access infrastructure makes this framework feasible. Federal managed-access infrastructure does more than open data; it makes opening data safe. Without tiered controls, aggregating federal biological datasets creates a system that bad actors can treat as a single target. Combining individually low-risk records produces re-identification risks for human subjects or reconstruction pathways for pathogen characteristics that no individual repository enables alone. Managed access addresses this

aggregation problem by ensuring that most data remains freely accessible while the narrow subset presenting genuine hazards receives proportionate controls calibrated to the specific risk. By developing a uniform standard for accessing federal biological data, this information can remain safe, security systems can be uniformly updated, and innovation enabled by this infrastructure can be used to address other national security challenges.

This is the moment to build infrastructure that treats biological data as a unified strategic asset: one where most data flows freely to researchers, and the small fraction carrying privacy, biosecurity, or intellectual property risk receives proportionate, standardized protection. Without these built-in coordinated access controls, the same AI models that accelerate drug discovery can also be trained to identify vulnerabilities in agricultural systems or reconstruct sensitive pathogen characteristics. Modern governance must account for both possibilities.

The Model for Biological Data Managed-Access

Congress established the National Security Commission on Emerging Biotechnology (NSCEB) through the 2022 National Defense Authorization Act to assess how the bioeconomy affects national security and to recommend policies that preserve American competitiveness. Recognizing the negative economic and national security implications of fragmented biological data governance, one of the NSCEB's 49 recommendations in their 2025 final report is the creation of a Web of Biological Data (WOBD), a single point where researchers across academia, government, and industry can readily access federally-funded biological data.

The NSCEB's WOBD recommendation is a useful starting model for the managed-access infrastructure that is necessary for a modern biological data ecosystem. However, no legislation has implemented the WOBD, and any federal effort to integrate biological data must address critical questions in two domains: accessibility (how users interact with data) and security (how risk is managed).

Domain 1: Accessibility

Federal biological data infrastructure must prioritize the user experience of researchers and their ability to utilize critical data to drive innovation. Mere data aggregation is insufficient; the infrastructure requires an architecture that integrates diverse datasets while minimizing barriers for legitimate users and adheres to FAIR data principles. To succeed, federal implementers should operationalize three core capabilities for accessibility:

- **Streamlined Onboarding.** The current fragmented vetting process slows research. A researcher studying microbe diversity currently navigates ten or more databases maintained by different institutions, each with distinct credentialing timelines, formats, and approval bodies. Managed access replaces this fragmented landscape with a single credentialing process and a unified query interface. In a managed-access infrastructure, most data will be accessible with minimal-to-no verification, comparable to creating an account on a federal research portal. For the small subset of sensitive datasets requiring controlled access, the new infrastructure replaces fragmented, agency-specific applications with a single credentialing process.
- **Dataset Interoperability.** Disparate federal databases currently use incompatible formats. The new data ecosystem should implement standards that allow machines to ingest and correlate data across human, agricultural, and microbial repositories. Data interoperability and the creation of AI-ready data standards is necessary for future innovation.
- **Ease of Use.** Researchers should be able to deploy user-friendly tools and query biological datasets through intuitive interfaces without needing to manage the underlying storage or compute infrastructure.

Domain 2: Security

Federal biological data infrastructure must implement robust cybersecurity protections and establish clear policies that govern how researchers access different datasets to develop novel technologies that improve security in the long-term. A functional managed-access infrastructure requires three security capabilities:

- **User Monitoring and Compliance.** Federal datasets currently operate under incompatible data-use and consent frameworks, creating ongoing challenges for researchers who must track separate compliance requirements across each dataset they access. The new infrastructure should adopt a unified monitoring protocol that tracks research activity across datasets and detects anomalous access patterns consistent with both accidental misuse and intentional data exfiltration.
- **Data Provenance Verification.** The integrity of managed data depends on tracking its lineage from collection through preprocessing. A new infrastructure should implement provenance verification mechanisms to ensure that datasets have not been contaminated, manipulated, or incorrectly attributed before incorporation.
- **Data Risk Stratification.** The vast majority of biological data, including environmental samples and de-identified research datasets, will be openly accessible under a base data-use agreement. For the small subset of data that presents specific hazards, including identifiable human genomic records, select agent sequences, and proprietary datasets with significant IP value, the infrastructure architects should: (1) identify the hazards associated with different types of biological data, (2) assess the risk posed by each data category based on likelihood and consequences of accidental data leakage or intentional misuse, and (3) operationalize data risk tiers to enable dynamic "Data Passports" that permit controlled and modifiable access to sets of data.

If properly constructed, a federal managed-access infrastructure can also serve as a model for other nations and industries participating in the global bioeconomy. Streamlining access to existing resources maximizes their utility and promotes interoperability across borders and industries. However, a consolidated framework and policy for data governance is necessary during implementation to optimize data access and security and to maximize the benefits of biotechnology in a new age driven by AI and large-scale biological data.

Biological Data Governance Framework

Three structural problems prevent the federal government from managing biological data as a coherent asset. Fragmented repositories and incompatible credentialing systems block coordinated researcher access. The absence of a shared metadata standard renders datasets non-interoperable and incompatible with AI training pipelines. And the proliferation of agency-specific data-use agreements creates compliance burdens that slow research without improving security. The recommendations below resolve each problem: a federated pilot program addresses infrastructure fragmentation, a biological data ontology addresses interoperability, and a unified data-use agreement addresses access governance.

Recommendation 1: Congress Should Establish a Federated Managed-Access Pilot Program for Federal Biological Data

Congress should authorize and fund the Department of Energy (DOE) to create a managed-access pilot program that tests federated data integration across agencies and identifies storage, compute, and security challenges that arise when biological datasets of different

types and risk profiles are combined. The WOBD, as proposed by the NSCEB, provides the architectural model for this pilot and the broader federal managed-access infrastructure the Genesis Mission now requires.

This authorization should mandate management by a coalition of National Laboratories and direct the DOE to designate a lead National Laboratory for the pilot, prioritizing laboratories with integrated biological data generation and high-performance computing capabilities. The Lawrence Berkeley National Laboratory, which houses both the Joint Genome Institute and the National Energy Research Scientific Computing Center, represents a strong candidate due to their existing combination of capabilities.

As part of the pilot program, the DOE should identify federally-funded biological datasets available for voluntary contribution to the preliminary repository, prioritizing collections that span varied biological sources, data types, agency origins, and risk profiles. The DOE should specify participants from NIH and USDA that manage biological datasets spanning different risk profiles: identifiable human genomic data subject to HIPAA and Common Rule protections, and de-identified agricultural or environmental datasets with minimal access restrictions. The DOE should initiate parallel conversations with the Department of Defense and the Department of Veterans Affairs to assess pathways for incorporating their highly-sensitive data in subsequent phases. While identifying preliminary datasets to incorporate in the pilot repository, the DOE should engage with interested public and private stakeholders to identify barriers to data sharing and new data sharing incentives that would promote participation in a novel infrastructure.

The legislation should direct the National Laboratories to develop the pilot as a federated data access architecture with access to expanded storage and compute resources across the National Laboratory network. This pilot would serve as a proof-of-concept for the managed-access infrastructure the NSCEB envisioned through the WOBD and that the Genesis Mission now demands. A federated architecture allows agencies to maintain control over their own datasets and security protocols while enabling cross-agency queries through a shared interface, shifting the locus of coordination from data custody disputes to interface governance and narrowing the administrative friction that deters agency participation. **This federated architecture should enable (1) a tiered data access system that calibrates researcher permissions to dataset sensitivity, (2) a compute-to-data model that allows researchers to analyze sensitive datasets in place rather than transferring them, and (3) a trusted-researcher environment that provides credentialed users with a monitored platform for accessing tools and data.**

The NIH All of Us Research Program provides a tested model for the pilot's architecture: it uses tiered data passports that grant progressively deeper access based on researcher credentials and training, and it operates a trusted-researcher environment where sensitive analyses occur without data leaving secure infrastructure (**Box 2: The NIH All of Us Research Program**).

Box 2: The NIH All of Us Research Program

The NIH All of Us Research Program has enrolled over 750,000 participants and governs researcher access through a tiered data passport system: a Registered Tier grants access to aggregate, de-identified data via click-wrap agreement with no prior institutional verification, while a Controlled Tier requires institutional affiliation verification, ethics training, and a signed data-use agreement before granting access to individual-level data. All sensitive analyses occur within a cloud-based Researcher Workbench where data cannot be downloaded outside secure infrastructure. The pilot program authorized under Recommendation 1 should replicate this architecture across federal biological data types, extending tiered access governance beyond human health data to environmental, agricultural, and pathogen datasets.

Recommendation 2: NIST Should Develop a Standardized Biological Data Ontology

Congress should, as companion legislation to the pilot program authorized under Recommendation 1, direct and fund the National Institute of Standards and Technology (NIST) to develop a standardized biological data ontology that incorporates biological data into a managed-access infrastructure, assigns it to risk tiers, and supports AI model training.

This ontology will serve as a basis of metadata to be assigned to all biological data included in a managed-access infrastructure, and should include key data structure categories:

- **Data Type:** Information regarding the sample/data species, source, cell type, and other biological annotations (e.g., human, whole-genome sequencing, blood tissue)
- **Data Provenance:** Information regarding the sample/data history and preprocessing as well as funding agency (e.g., NIH-funded, collected 2022, preprocessed with standard QC pipeline)
- **Data Quality:** Information regarding technical quality metrics for the dataset, including sequencing depth, error rates, and quality control pipeline outputs (e.g., mean coverage 30x, batch correction applied)
- **Data Analysis:** Information regarding computational analyses performed on the dataset and their outputs (e.g., variant calling completed, gene expression normalization applied)
- **Data Use-Context:** Information regarding the original research goal of the data (e.g. clinical pharmacogenomics research)
- **Data Risk:** Information regarding anticipated intellectual property, privacy, and biosecurity risk of the data (e.g., HIPAA-covered, identifiable, Tier 2 controlled access)

In developing this ontology, NIST should engage industry and academic stakeholders to identify crucial classification information necessary for AI model training and data risk stratification, to identify cost and technical barriers to consistent ontological annotation, and to co-develop automated pipelines that assign ontological classifications at the point of data generation wherever technically feasible. Policymakers should look towards current efforts to standardize biological design language and ontologies, such as the Biolink Model, Synthetic Biology Open Language, and European Viral Outbreak Response Alliance Ontology. NIST should treat ontology development as an iterative process: an initial base ontology should establish core metadata requirements across the six data structure categories, with successive revisions deepening classification specificity as new datasets enter the managed-access infrastructure and stakeholder feedback surfaces gaps. In this process, NIST should look towards recommendations from the NSCEB on developing this ontology to advance towards AI-ready biological data that can be incorporated into any federal managed-access infrastructure[1].

Recommendation 3: OSTP & NSTC Should Create a Unified Data-Use Agreement for Biological Data

The President should direct the Office of Science and Technology Policy and National Science and Technology Council to form an interagency subcommittee charged with developing a unified data-use agreement (DUA) for federally funded biological data. This subcommittee should include representatives from agencies responsible for data generation and custody (DOE, National Laboratories, NIH, USDA), national security oversight (National Security Council, the Cybersecurity and Infrastructure Security Agency), and federal

[1] The AI-Ready Bio-Data Standards Act (S. 4069, 119th Congress) would direct NIST to establish definitions, standards, and frameworks for AI-ready biological datasets, operationalizing this NSCEB recommendation.

procurement and budget authority (the Office of Management and Budget), whose Federal Acquisition Regulation authority creates a mechanism for DUA compliance in federal contracting.

If Congress establishes a National Biotechnology Coordination Office within the Executive Office of the President as recommended by NSCEB, this office would assume direct responsibility for convening an interagency group to draft the unified DUA[2].

Through this subcommittee, federal agencies that fund, develop, and manage biological datasets will develop a unified DUA that: (1) clarifies the authentication and authorization process for researchers accessing federal biological data, (2) standardizes how researchers access different data types, tools, and models across agencies, and (3) defines the permitted use cases and research criteria for each access tier.

As part of this process, subcommittee members will identify criteria for classes and tiers by which data is structured and to which researchers can separately gain access. A model for this structure is the tiered passport system implemented within the All of Us Research Program. In practice, the subcommittee should establish three access levels proportional to data risk.

The base data classification should cover the majority of federally funded biological data, including sequence data, environmental samples, and de-identified research datasets. Researchers should access this classification through a click-wrap acknowledgment of standard terms of use, comparable to open-access repositories such as NIH's GenBank, with no prior application or institutional verification required.

A secondary classification should cover data with moderate access requirements, including aggregate datasets with re-identification risk or proprietary research outputs. Researchers should access this classification through a standardized institutional verification process confirming affiliation with an accredited research institution or registered biotechnology firm. The subcommittee should establish a standardized enrollment process for institutions not yet participating in the federal biological data system, ensuring that managed access does not reproduce existing disparities in who can conduct biological research.

Data carrying identified privacy, biosecurity, or intellectual property risks should be governed by category-specific addendums to the base DUA. For example, an addendum for identifiable human genomic data would require HIPAA compliance training and IRB approval; an addendum for select agent characterization data would require institutional biosafety credentials and usage monitoring. Each addendum should specify the additional credentials, training, and monitoring required beyond the base DUA. The unified DUA and addendum system should not collapse existing access distinctions but standardize the process for navigating them and enable researchers to navigate distinct access pathways calibrated to their credentials and use case.

The subcommittee should also assess how the unified DUA interacts with existing memoranda of understanding and material transfer agreements that govern biological data sharing; where these instruments duplicate or conflict with DUA requirements, the subcommittee should recommend

[2] The National Biotechnology Initiative Act (S. 1387, 119th Congress) would establish a National Biotechnology Coordination Office within the Executive Office of the President to coordinate biological data standardization across federal agencies; the Biosecurity Modernization and Innovation Act (S. 3741, 119th Congress) offers a parallel track by establishing a biotechnology governance office within the Department of Commerce.

standardized templates that reduce the compliance burden researchers face when navigating multiple agreement types for a single dataset.

Conclusion

The United States holds a global advantage in biological data collection and biotechnological innovation, but fragmented governance threatens to squander that advantage at the moment when AI makes integrated data most valuable.

The pilot program establishes the infrastructure for cross-agency biological data integration and tests the governance model at scale. The standardized ontology provides the metadata framework that makes datasets interoperable and AI-ready. The unified DUA replaces the current patchwork of incompatible access requirements with a single, tiered system that balances openness with proportionate security.

These policies, implemented together, provide the federal government a consolidated framework for biological data governance at the scale and speed that AI demands, and position the United States to lead in biotechnology for the next generation.

Jon Judd is supported by the National Science Foundation (NSF) under Award No 2534654. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Jon would like to thank and acknowledge all individuals who lent their expertise and insights through informal interviews that helped shape the commentary and recommendations.

Jon would also like to thank and acknowledge the following individuals for their careful review and feedback: Sebastian Rivera, Christopher Hoover, Emily Aurand, Garrett Dunlap, Rebecca Mackelprang, India Hook-Barnard, Sarah Carter, and Steven Moss.