Sustainability During Instability: Long-Lived Life Science Databases and U.S. Funding Outlook

bioRxiv, Oct 2025 https://doi.org/10.1101/2025.10.08.680785

Heidi Imker, Avner Professor of Interdisciplinary Research

University Library University of Illinois Urbana Champaign

Context and Methods

Old Problems

The demise of public data on the web?
Ellis & Kalumbi, 1998
Nature Biotechnology
https://doi.org/10.1038/4296

The demise of public data on the web?

Data rich but cash poor, many free biological databases are on the verge of financial collapse.

Lynda B.M. Ellis and Doyle Kalumbi

Hundreds of public biological databases exist on the Internet; in mid-1998, over 400 were listed in DBCAT, the INFOBIOGEN biological database catalog[1]. These are significant information resources that represent the fundamental infrastructure for future biological research and biotechnology application. The rapid expansion of this infrastructure and has outpaced the scientific and business communities' ability to establish economically viable models for its long-term survival. For example, many funding agencies that support the creation of new databases are less willing to fund their further development and continued maintenance[2].

Immediate efforts must therefore be undertaken to develop economic models to describe this new infrastructure, quantitatively measure its cost and value, and propose mechanisms to ensure its long-term economic viability. In May, 1997, the European Biotechnology Information Strategic Forum (BTSF) sponsored a workshop on Financing biotechnology databases[3]. The workshop concluded that a framework was required to guide and support database producers in their quest for finan-

records was downloaded from DBCAT. Only primary—not mirror—sites were included. Of these 397 biological database records, 8 had no e-mail contact addresses, 1 was not accessible to the public, and 67 were duplicates, leaving 321 databases with e-mail contacts.

A presurvey e-mail message was sent to each contact two days before the survey.

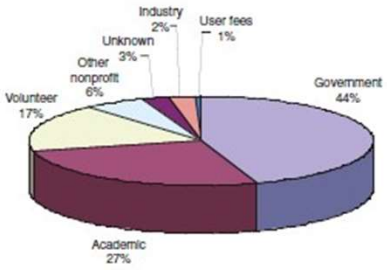


Figure 1. Present funding sources of 153 public biological databases, based on results of the authors' survey.

Two reminders, which also included the survey, were sent to nonresponders 10 and

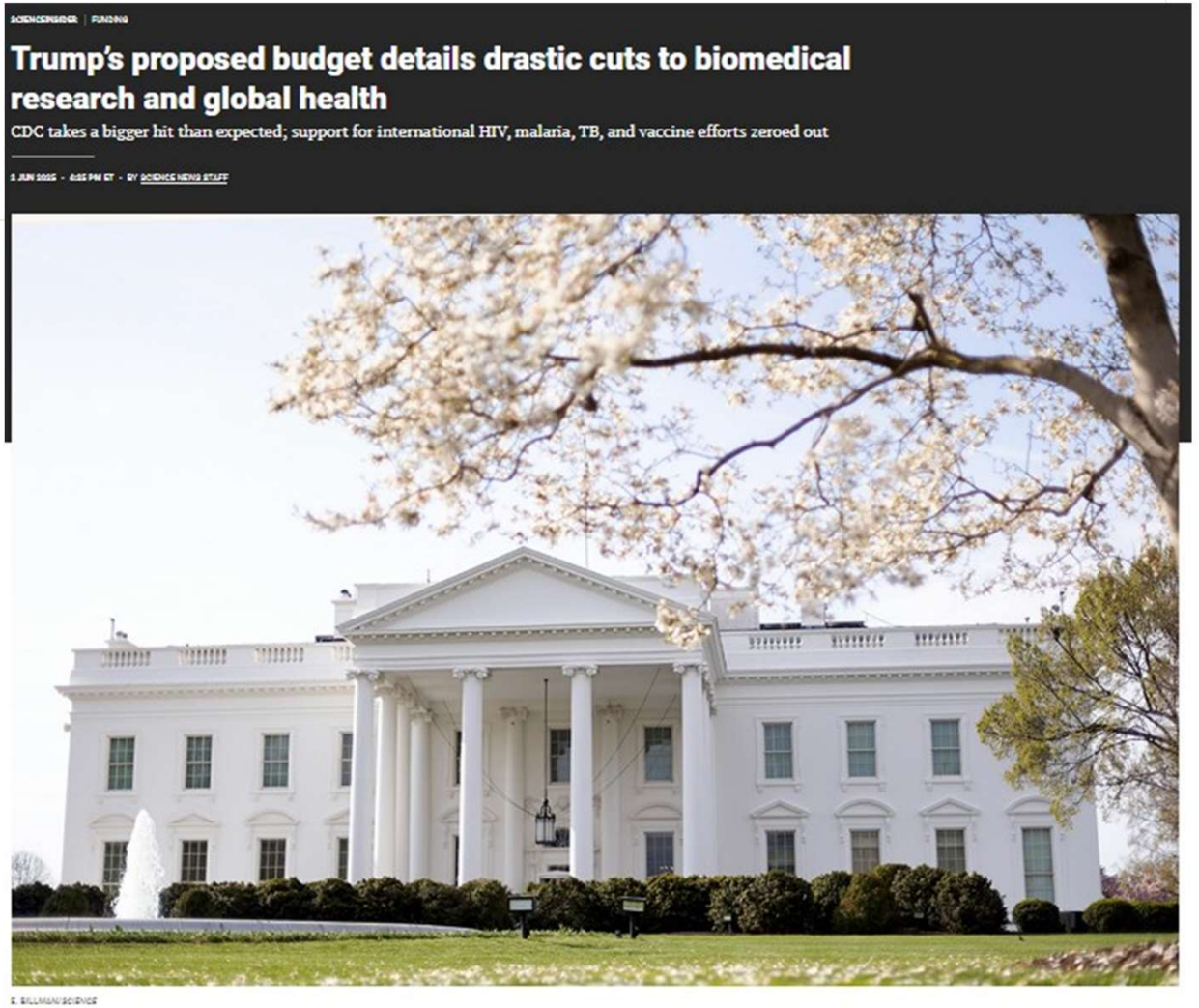
We received survey responses covering 153 (out of 321) databases for a 48% response rate. Based on the e-mail domain of the respondent, 48% of our responses were from the US and Canada; 33% from Western Europe; 14% from Australasia; and 5% from other countries (Mexico, Israel, India, Russia, and South Africa). No significant difference in geographic distribution was found between the 153 respondents and the 168 nonrespondents.

The original source of these database listings, DBCAT, though open to all biological databases, primarily contains those focused on molecular biology, as opposed to organismal or ecological biology[1]. It would be interesting to expand the survey to these other biological databases. Also, DBCAT listings and our survey respondents were primarily from developed countries. Although this is a valid indication of who hosts these particular databases, their users span the entire globe. For example, in 1997, the University of Minnesota Biocatalysis/Biodegradation Database[4,5] was accessed by people from 90 countries on 6 continents[6].

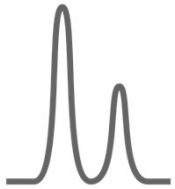
Who hosts and who funds these

New Problems

Trump's proposed budget details drastic cuts to biomedical research and global health
June 2, 2025
Science
https://doi.10.1126/science.ze5w9si



Prior Experience



Background:

Biochemist by training

User and creator of data resources

Director:

Research Data Service

Data management and sharing across campus

Prior Work

25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance.
Frontiers Research Metrics and Analytics, 2018
https://doi.org/10.3389/frma.2018.00018

→ 100 **new** databases in annual *NAR* Database Issue each year

Who Bears the Burden of Long-Lived Molecular Biology Databases?
Data Science Journal, 2020
https://doi.org/10.5334/dsj-2020-008

→ U.S. government has a massive role in funding and hosting databases

A machine learning-enabled open biodata resource inventory from the scientific literature.
PLoS ONE, 2023 (with Schackart, Istrate, and Cook)
https://doi.org/10.1371/journal.pone.0294812

→ Large, globally-distributed infrastructure that's hard to track and quantify

Study Overview

Research Objective

investigate vulnerability of long-lived life science databases amid funding cuts in the US

Database Criteria

available on the web >15 yrs, based in US with life science focus, and have received funding from US federal agency

Methodology

multiple mini case study, including interviews and document triangulation for verification and comprehensive analysis

Limitations

depth, single author, generalizability

9 Diverse Databases

biodiversitylibrary.org

morphobank.org

wormatlas.org

ordb.biotech.ttu.edu/ORDB

rrndb.umms.med.umich.edu

omim.org

veupathdb.org

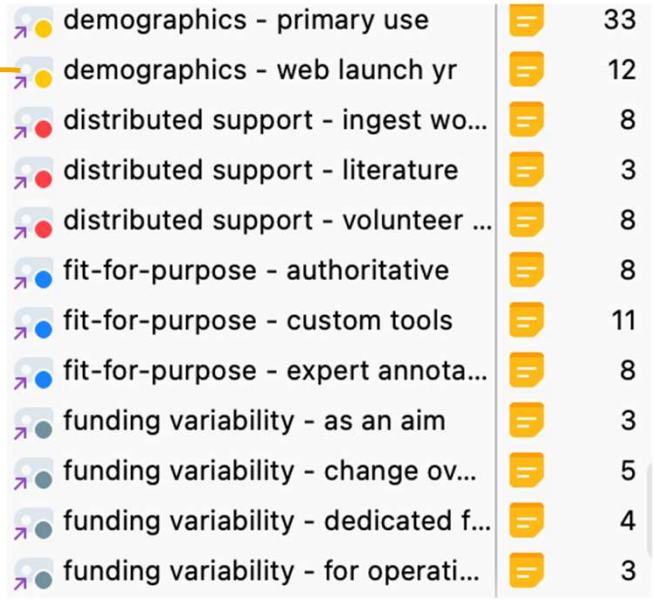
2 anonymous

Databases Profiles

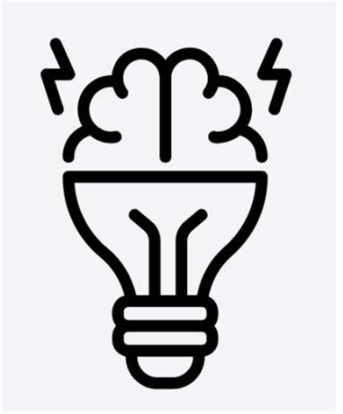
NAME	1ST YR ON WEB	PRIMARY DATA	PRIMARY USES
BHL	2007	Archival/contemporary literature/materials	Study historical trends, taxonomy verification
MorphoBank	2004	Phylogenetic matrices and data, 2D/3D morphology	Comparative studies, phylogenetics
OMIM	1995	Gene and phenotype entries, summaries/references	Exploration of genes and genetic disorders
ORDB	1994	Olfactory receptor gene/protein sequences	Study of OR and OR-Like genes
rrnDB	2001	16S/23S RNA genes, taxonomies, summaries/references	Microbial population studies
VEuPathDB	2004	Multi-omic data on pathogens, vectors, hosts	Centralized access for disease research
WormAtlas	2002	Electron microscopy images, maps, and education	Study of anatomy and development
DB1 and DB2	> 15 yrs	Data of biomedical relevance	Data and tools relevant to human health

Qualitative Method

Collect Sources



Iterative Coding



Interpret

Results

Common Purpose

Efficiency

- make scattered or otherwise inaccessible data discoverable and fit-for-purpose (reliable and relevant)
- reduces effort, capitalizes on past research investments, supports comparative studies, and makes research quicker and more comprehensive

★ **Data Resource = Expertise Resource**

Support

Direct Monetary Support	funds explicitly given to carry out the work of operating, maintaining, or upgrading the resource
Non-monetary support	resources, such as access to services or expertise, provided by 1) host or 2) distributed among a resource's community

Sidebar: On Funding Models

Best Overview of Potential Models

Organisation for Economic Co-operation and Development (OECD) "Business Models for Sustainable Research Data Repositories" 2018
http://www.oecd-ilibrary.org/science-and-technology/business-models-for-sustainable-research-data-repositories_302b12bb-en

Best In-depth Longitudinal Study

Eschenfelder, Shankar, and Downey, "The Financial Maintenance of Social Science Data Archives: Four Case Studies of Long-Term Infrastructure Work," 2022
https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24691

Direct Monetary Support

TYPE	EXAMPLES	NOTES
Grants and Contracts	NSF, NIH, private foundations	**Current norm;** time-limited and must align with funder priorities
Membership Fees	BHL, MorphoBank, Dryad	Needs high incentive to join via benefits, such as governance rights, deposit advantages
Subscriptions	VEuPathDB (newly), TAIR	Needs high incentive to pay for specific good/services
Donations	Fund drives, call-to-action buttons	Needs good will; less predictable

Non-Monetary Support

TYPE	EXAMPLES	NOTES
Host Support*	Fiscal sponsorship, shared IT resources, staffing	Needs local capacity and resources
Distributed Support	Data deposits, volunteer service	Needs good overall health of the community

* can be direct monetary, e.g., via institutional grants or a form of donation

Impacts and Strategies

Anticipated Impacts of Lost Funding

Higher Barriers
to fit-for-purpose data, reintroduction of inefficient and fragmented research

Loss of Expertise
loss of subject matter experts, impacting data quality, interpretation, and future scientific training

Interrupted Innovation
stalled infrastructure upgrades and adoption of new technologies like AI

Current Impacts
cuts and knock-on effects have forced databases to find alternative support

Sustainability Today

Prior discussions **assumed robust government support for science overall**

Diversified funding sources more imperative than ever, being aware of:

- Move towards philanthropic and industry funders
- Tightening university budgets will impact host support, stop gap funding, and the viability of organizational-level memberships and subscriptions

As much about **durability of and cushion in operations** to weather storms

Yet this is a very **data hungry** world (and soon, very expertise hungry)

Conclusion

Key Takeaways

Importance of Databases

essential for research efficiency, cohesion, and education

Funding Instability

threatens database sustainability, risking practical utility of data and loss of expert knowledge

Reinforces the Continued Need for Long-Term Strategies

balancing open access and financial realities

As open as possible, as closed as self-sustaining requires?

imker@illinois.edu