Sustaining the ‘Big Data’ Ecosystem

Eric Green, M.D., Ph.D.

National Human Genome Research Institute

National Institutes of Health, USA


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
go raibh maith agat
mochchakkeram
dziękuję
bedankt
спасибо
Баярлалаа
рахмат
sukriya
kop khun krap
terima kasih
감사합니다
ありがとう
arigatô
takk
dakujem
мерси
grazie
tapadh leat
dank je
hvala
maururu
sagolun
taiku
ευχαριστώ
spas
mersi
barka
welalin
tack
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
djiere dieuf
tau
дякую
mamnun
chnorakaloutioun
gracies
sulpay
trugarez
dhanyavadagalu
shukriya
merce
rahmet
tanemirt
diolch
xiexie
najis tuke
kam sah hamnida
rahmat
didi madloba
mèsi
sobodi
dekuji
enkosi
bayarlalaa
gracie
kiitos
dankie
nandri
nanni
dhanyavad
köszönöm
taafetai lava
спасибі
kia ora
dankon
aciú
akun
তোমাকে ধন্যবাদ

The Largest Bottleneck In Biomedical Research…
…Pick Your Favorite Data-Related Metaphor!

TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

Nature **2013**

SCIENCE | biology

Biology's Big Problem: There's Too Much Data to Handle

BY EMILY SINGER, QUANTA MAGAZINE 10.11.13 9:30 AM

***WIRED* (2013)**

For My Talk…

- **Goals:**
  - **Help to provide context for the workshop**
  - **Begin to frame the discussion**

- **Two Overlapping Roles:**
  - **Represent the U.S. National Institutes of Health (NIH)**
  - **Director, National Human Genome Research Institute (NHGRI)**

- **Articulate themes from the 2015 Bourne/Lorsch/Green *Nature* Perspective**

Challenges Of Sustaining Data Resources

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne, Jon R. Lorsch** *and* **Eric D. Green.**

Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries (including the United States), and data are being generated at unprecedented rates, so the research community must find more efficient models for storing, organizing and accessing biomedical data. Simply putting more and more money into the current systems is unlikely to work in the long term.

To better understand this situation, we are examining the current and projected costs of managing biomedical data at the US National Institutes of Health (NIH). Our initial analyses indicate that even if we leave out the National Center for Biotechnology Information, which is a special case, the 50 largest NIH-funded data resources have a collective annual budget of US$110 million. And this figure represents just the tip of the iceberg for future needs.

UNDERSTANDING USAGE

Today's biomedical data resources typically treat all items in their collections equally. This does not always make sense, given that the usage patterns of the data vary. But how do we decide which data get more attention? As larger and larger data sets are generated more easily, and the cost of maintaining and annotating these data continues to rise, this question is becoming increasingly important.

Answering it requires a better understanding of how research data are used. This has rarely been thoroughly explored. Historically, funders have been interested primarily in knowing how the data resources that they support are used and by whom. They tended not to look closely at the details of how and why individual items and types of data within a collection are used.

Analyses of these details can be revealing. Preliminary studies suggest that typically a small subset of the data is used frequently, whereas most of the data are rarely accessed. However, the exact subset of data that is used heavily may change over time, and most of the data access may be performed after the data are downloaded, so this is not recorded. All of this means that absolute numbers are hard to interpret.

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

When we have a better understanding of data usage, we can develop business models that consider supply and demand, and develop sustainable practices. In addition, finding economies of scale and harnessing market forces will be essential.

For a typical biomedical data resource, the cost of simply keeping the data is only a small fraction of the total cost of data management. The remainder is largely the cost needed to support the finding, accessing, interoperating and reusing (the FAIR principles; see go.nature.com/axkjiv) of the data — a cost that is widely underappreciated.

Is the FAIR fraction of the cost justified? Are services from different data resources redundant? Are resources subject to 'feature creep' — the addition of costly 'bells and whistles' that are of limited value? Do our funding mechanisms contribute to these problems? And most importantly, is the way we currently maintain biomedical data optimal for the science that needs to be done both today and in the future?

Current practices typically use many disparate sources of data to conduct a study. These data are located in a variety of repositories, often with different modes of access. This lack of centralization and commonality may hinder their ease of use and reduce productivity. We need a better understanding of usage patterns across multiple data resources to use as a basis for redesigning such resources to preserve valuable expertise and curation, and for improving how the data are found, accessed, integrated and reused.

The nature of curation and the quality assurance for biomedical data must also change. Complete and accurate automated or semi-automated extraction of literature is needed to provide metadata and annotation. We should consider crowdsourcing curation, with appropriate validation and incentives. Additionally, the role of professional curators must be better appreciated by data users, by the institutions where the curators work, and by the funders.

> THE RESEARCH COMMUNITY MUST FIND MORE **EFFICIENT** MODELS FOR **STORING, ORGANIZING AND ACCESSING** BIOMEDICAL DATA.

Nature **(2015)**

Overarching Realities

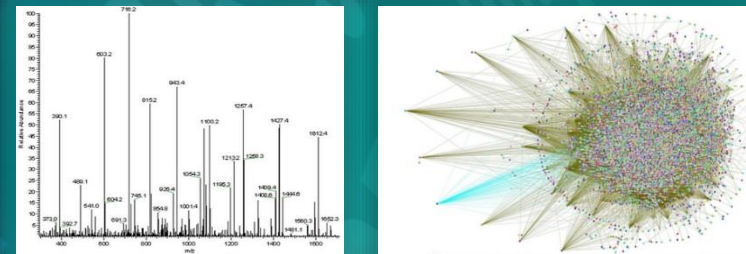
1. We are victims of our own success

2. Genomics is a 'poster child' for the problem

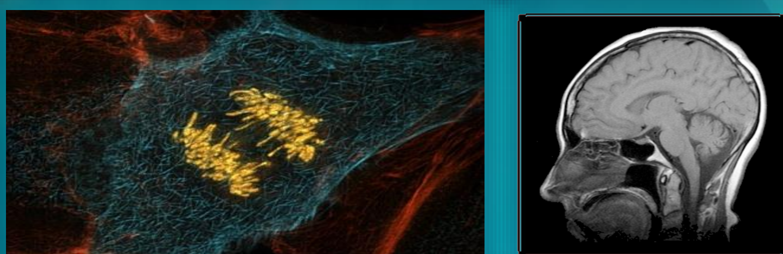
3. But other data types are quickly becoming similarly problematic

Myriad Data Types

Genomic

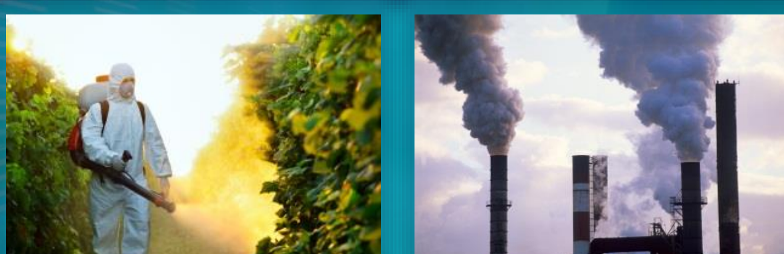


Other 'Omic

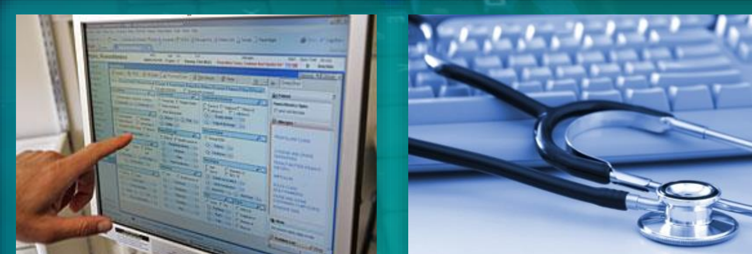


Imaging

Phenotypic



Exposure



Clinical

Overarching Realities

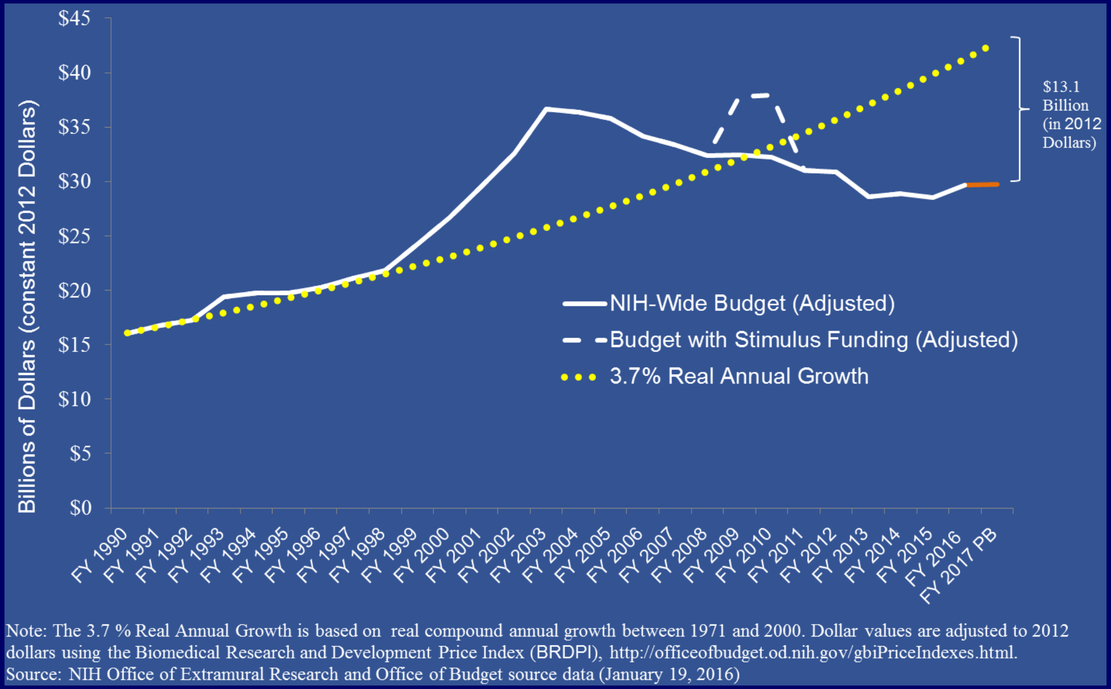
1. We are victims of our own success

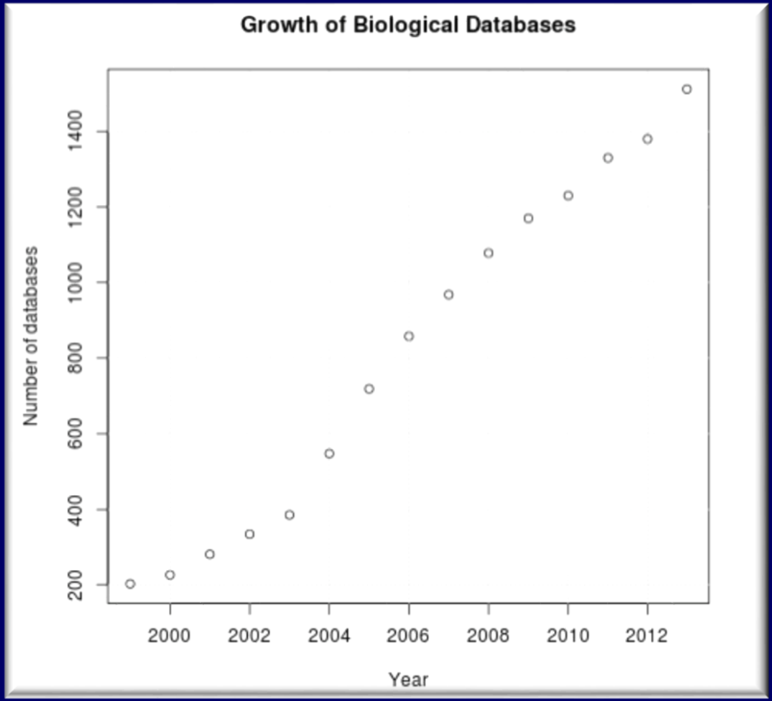
2. Genomics is a 'poster child' for the problem

3. But other data types are quickly becoming similarly problematic

4. The 'trends' are particularly troubling

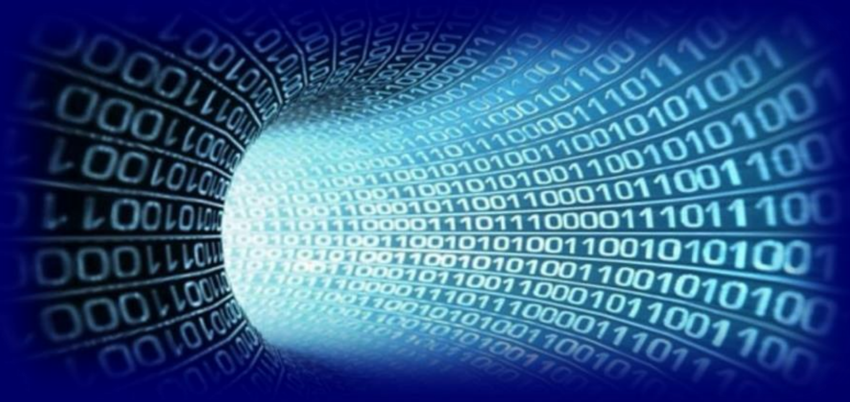
Budgets Are Mostly Flat (or Worse)…
Data Growth Is Anything But Flat!





Current models for data sustainability do <u>not</u> scale

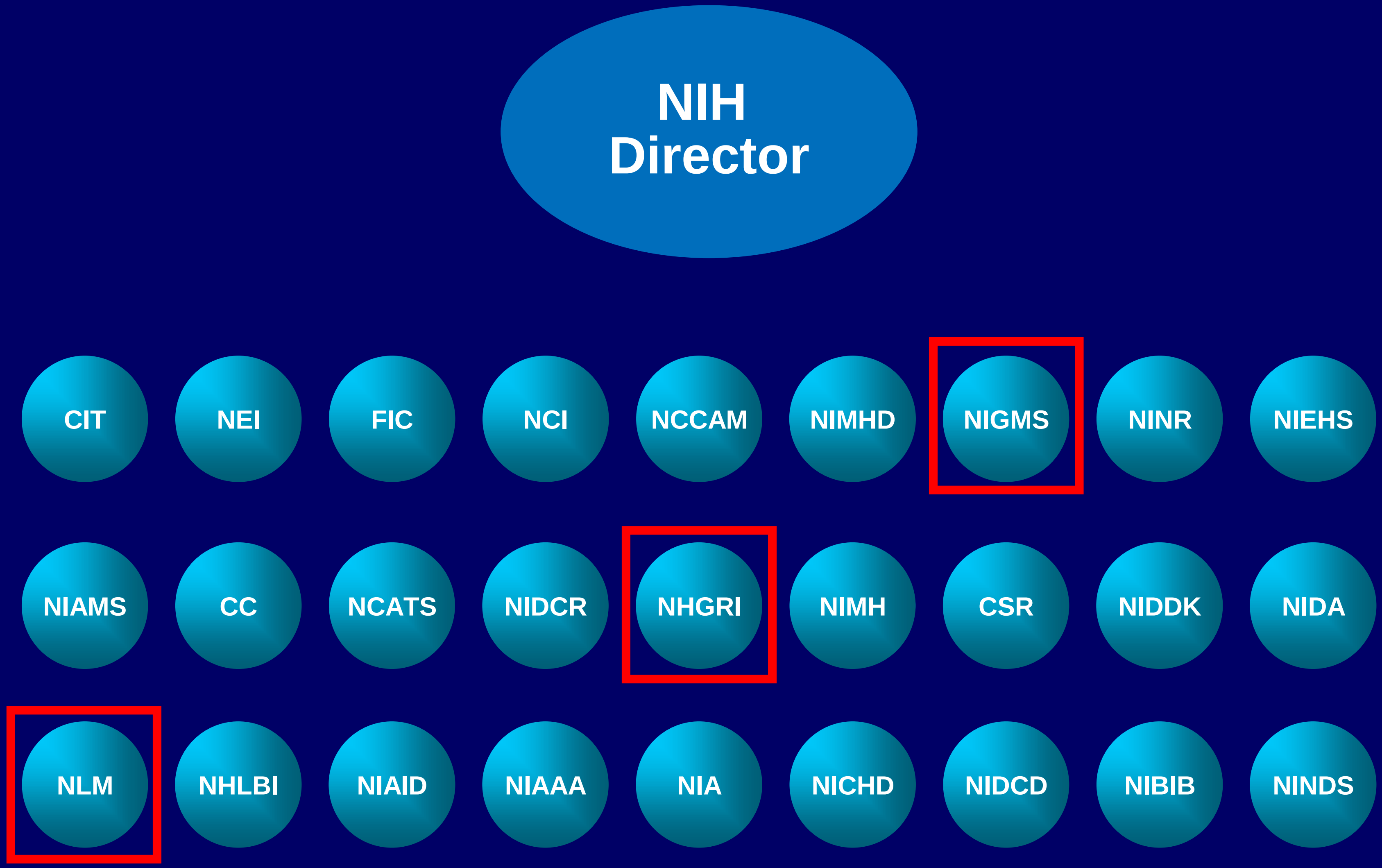
Big Data & Data Sustainability



Analyzing the problem

Self-reflection

NIH: 27 Institutes & Centers



Many Aspects Of The Problem 'Slip Between The Cracks'

Relevant Working Group Reports

ADVISORY COMMITTEE TO THE DIRECTOR

National Institutes of Health

Data and Informatics Working Group

Draft Report to
The Advisory Committee to the Director

June 15, 2012

National Institutes of Health
Advisory Committee to the Director

National Library of Medicine (NLM) Working Group

FINAL REPORT – JUNE 11, 2015

MEMBERS: Eric Green (co-chair), Harlan Krumholz (co-chair), Russ Altman, Howard Bauchner, Deborah Brooks, Doug Fridsma, Steven Goodman, Eric Horvitz, Trudy MacKay, Alexa McCray, Chris Shaffer, David Van Essen, Joanne Waldstreicher, James Williams, II, Kathy Hudson (ex officio), Lyric Jorgenson (executive secretary) *(titles and affiliations listed in Appendix A)*

EXECUTIVE SUMMARY

The NIH Director charged the National Library of Medicine (NLM) Working Group, hereafter referred to as the Working Group, with articulating a strategic vision for NLM to ensure that NLM remains an international leader in biomedical and health information. Over the course of five months of deliberations, the Working Group reviewed numerous documents and reports pertaining to NLM's mission and activities, consulted with NLM leadership and staff, and solicited public comments and suggestions. The Working Group recognizes that NLM has an important opportunity to play a key leadership role in one of the most exciting periods of biomedical history: data science is increasing rapidly, computational power is expanding at a breathtaking pace, the breadth and depth of digital health data are undergoing unprecedented and accelerating growth, a movement towards more interdisciplinary work and team science continues to gain momentum, a broad commitment to open science is becoming increasingly adopted, and the demand for services to support an ever more engaged and informed public is expanding. To leverage these historic changes, the Working Group, with respect for the outstanding history of NLM and its potential for the future, formulated a series of recommendations to guide the future of NLM:

acd.od.nih.gov/diwg.htm

acd.od.nih.gov/nlm.htm

NIH: 27 Institutes & Centers

NIH Director

Associate Director for Data Science (Dr. Phil Bourne)

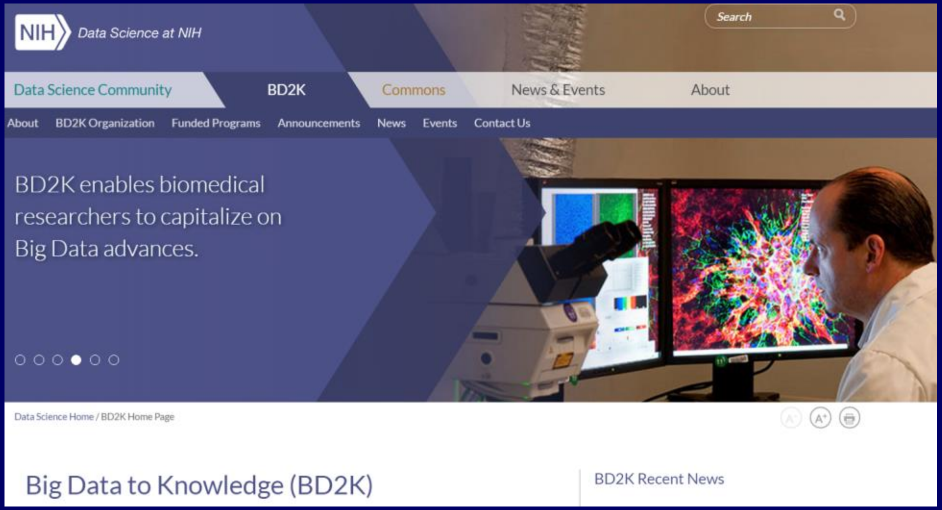
CIT NEI FIC NCI NCCAM NIMHD NIGMS NINR NIEHS

NIAMS CC NCATS NIDCR NHGRI NIMH CSR NIDDK NIDA

NLM NHLBI NIAID NIAAA NIA NICHD NIDCD NIBIB NINDS

Data Science @ NIH

Valentina Di Francesco
NHGRI, NIH



Model Organism Databases (MODs)

BIOMEDICAL RESOURCES

Funding for key data resources in jeopardy

NIH genome institute wants to scale back support of human and model organism databases

By **Jocelyn Kaiser**

As the world's most authoritative catalog of human disease-related genes approaches its 50th birthday, it faces unsettling change. Over the next few years, the National Human Genome Research Institute (NHGRI) expects to bow out as sole funder for the granddaddy of genomic databases, known as Online Mendelian Inheritance in Man (OMIM). Who will pick up the tab is not yet clear. Other, newer databases supported by NHGRI are facing a similar threat as the National Institutes of Health (NIH) takes stock of all its data resources.

Users are concerned. These free-to-use resources, which cover everything from yeast genomics to proteins, are "critical for our daily life as geneticists and biomedical researchers," says University of California, Berkeley, geneticist Jasper Rine, president of the Genetics Society of America. Ada Hamosh of Johns Hopkins University in Baltimore, Maryland, who oversees OMIM, adds: "If NIH is going to develop new funding models, they need to make sure they don't compromise the integrity of existing, heavily used resources."

NHGRI Director Eric Green says that nothing has been decided and that rumors that his institute plans to phase out all of its funding are incorrect. But he and other NIH leaders are searching for ways to make the databases more efficient, and are urging databases to consider charging for use.

Biology databases have long had funding woes. Science agencies often complain that database support diverts resources from their mission of funding research. Philip Bourne, NIH's first associate director for data science, estimates that the 50 largest NIH-supported resources—not counting GenBank and other databases at the National Library of Medicine (NLM)—require $110 million of the agency's $30 billion annual budget. An explosion in data is making them ever more costly to run. "There is a sustainability issue. We need to do something," Bourne says.

In addition to OMIM, NHGRI supports five databases for model organisms and others such as UniProt, which holds data on protein function. All are troves of molecular data annotated with information that curators have gleaned from the literature. OMIM, which began 50 years ago as a paper resource and moved to the Web in 1995, draws more than 23 million page views a year. Clinicians use it to diagnose patients with rare diseases, while basic researchers rely on OMIM and model organism databases as go-to references for genes and their protein products.

Last May, Green called together leaders of these databases to tell them that by 2020 they need to find new options for funding. Grantees across many NIH institutes use the NHGRI databases, he said, and the nearly $30 million a year NHGRI now provides isn't enough as the databases expand beyond genomes to biological data. "We're not a good long-term home," Green says. "We need to think about new ways to do business."

Bourne's office plans to compile data on database usage across NIH, although he notes this can't be the only measure of value: "You can have a relatively small number of users, but it's absolutely critical for those users," he says. He and Green wonder whether some databases could be combined to lower costs. Further automating curation might also help. But humans still need to read papers and pull out data, in part because formats and nomenclatures vary.

Shifting some databases to other institutes could cut NHGRI's costs, as could adopting a subscription model. The Arabidopsis Information Resource (TAIR), the central database for a model organism in plant science, started charging fees in 2013 after the National Science Foundation phased out funding. "We resisted very strongly," says TAIR Director Eva Huala in San Francisco, California, but in the end, "we were converted."

TAIR tailors its prices to how much individuals, institutions, and companies use the database. "It's a great way to ensure that those who benefit the most from a resource also contribute the most," Huala says. As a bonus, she adds, because TAIR doesn't rely on federal grants, it no longer has to please peer reviewers and can focus instead on what users want, mainly up-to-date data.

The shift required some major changes, however. Huala and her staff left the Carnegie Institution for Science, which had hosted TAIR, to start a nonprofit, Phoenix Bioinformatics, to run it. They also had to set up accounting and business systems.

Several of the NHGRI-funded databases are wary of this model. "It isn't practical for many reasons," says Janan Eppig of the Jackson Laboratory in Bar Harbor, Maine, principal investigator for the Mouse Genome Database. One problem is that paywalls may prevent researchers from linking to genetic data in other databases. And researchers would need to use their grant money to subscribe. "NIH ultimately pays the bill anyway," says Monte Westerfield of the University of Oregon in Eugene, who heads the Zebrafish Model Organism Database.

Bourne says an NIH-wide committee hopes to begin considering new funding schemes later this year. The fate of these data troves may not be clear, however, until the agency has hired a new director for NLM—a possible new home for the beleaguered databases. ■

Data troves in transition

These databases supported by the National Human Genome Research Institute have 4 years to develop new funding models.

DATABASE	ORGANISM	UNIQUE USERS PER MONTH	2015 NHGRI FUNDING
FlyBase	*Drosophila*	51,300	$4.2 million
Gene Ontology Consortium	Multiple	36,000	$3.7 million
Mouse Genome Database	Mouse	53,100	$4.7 million
Online Mendelian Inheritance in Man	Human	300,000	$2.1 million (2014)
Reactome (biological pathways)	Human	19,400	$1.2 million
Saccharomyces Genome Database	Yeast	65,000	$2.7 million
UniProt (protein function)	Multiple	433,100	$4.9 million
WormBase	*Caenorhabditis elegans*	15,500	$2.9 million
Zebrafish Model Organism Database	Zebrafish	23,300	$3.1 million

DATA: FLYBASE/GO CONSORTIUM/MGD/OMIM/REACTOME/SGD/UNIPROT/WORMBASE/ZFIN

Valentina Di Francesco
NHGRI, NIH

Emerging Themes from NIH Experience

- **At a pivotal point:**
  - Risk failing to capitalize on technology advances
  - Failure to act would be devastating
- **Must implement new models to ensure a sustainable infrastructure for preserving, sharing, analyzing, and integrating data**
- **Cultural changes must be part of the solution**
- **Long-term commitment is required**

Challenges Of Sustaining Data Resources

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne, Jon R. Lorsch** *and* **Eric D. Green.**

THE RESEARCH COMMUNITY MUST FIND MORE **EFFICIENT** MODELS FOR **STORING, ORGANIZING AND ACCESSING** BIOMEDICAL DATA.

- **Understanding Usage**
- **Fair and Efficient**
- **Business Models**
- **Common Ground**
- **Uniting Funders**

What Do We Need To Understand?

- What is the relative value of data at different points in time?
- Where are redundancies?
- How to better match data curation to usage?
- How to make more informed decisions about changing data resources (e.g., merging, closing, expanding, etc.)?
- What is the nature of the major cost drivers?

Who Is Responsible For Sustaining Data?

- Indexed to data production vs. data usage?
- What component(s) within a funding agency?
- Government vs. other funders vs. industry?
- Individual countries vs. international group(s)?

International Workshop:
Data Resources in the Life Sciences

HFSPO Secretariat, Strasbourg, France
18/19 November 2016

Tackling the *international* dimensions of a problem that the NIH has been intensely working on for ~5 years (…and likely will be addressing for the next ~100 years!)

What We Need…

- To recognize that:
  - This is a significant and growing problem
  - This is not the problem of any one country/agency
  - This is not just the funders' problem
  - The data ecosystem is complicated and changing
  - The answer cannot be 'provide more $$$'
- A robust 'scoping out' of the problem
- A willingness to break down traditional boundaries and silos to create new collaborative solutions
- An openness to new 'new business models'
- A new framework for international solutions
- Most important … get to work!