

# Invisible Infrastructure at Risk: Funding Crises Threaten Global Bioscience Research

## Authors

Valerie Wood<sup>1\*</sup>, Elspeth A. Bruford<sup>2</sup>, Steven J. Marygold<sup>3</sup>, Valerie B. O'Donnell<sup>4</sup>, Katja Röper<sup>3</sup>, Jamie A. Davies<sup>5</sup>, Christine Orengo<sup>6</sup>, Paul Nurse<sup>7</sup>, Neil Hall<sup>8</sup>

## Affiliations

<sup>1</sup> Department of Biochemistry, University of Cambridge, Cambridge, UK

<sup>2</sup> Department of Haematology, School of Clinical Medicine, University of Cambridge, Cambridge, UK

<sup>3</sup> Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK.

<sup>4</sup> School of Medicine, Heath Park, Cardiff University, Cardiff, UK

<sup>5</sup> Edinburgh Medical School, The University of Edinburgh, Edinburgh, UK

<sup>6</sup> Institute of Structural and Molecular Biology, University College London, London, UK

<sup>7</sup> The Francis Crick Institute, London, UK

<sup>8</sup> The Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK

\* Corresponding Author: Valerie Wood (vw253@cam.ac.uk)

## Running Head

Invisible infrastructure at risk

## Abstract

Modern biology depends on invisible infrastructure that is essential yet precariously funded. Millions of biological observations are interpreted through resources that sit beneath routine scientific practice, largely unseen in everyday research. Every experiment, genome analysis, AI model, and clinical interpretation relies on these biological knowledgebases, which most researchers rarely notice until they disappear. That infrastructure is now under threat. Across the world, the biological databases that underpin modern research, built over decades of public and charitable investment, face severe and increasing financial instability. This crisis arrives at a particularly damaging moment: many governments have identified life sciences as a strategic growth sector and are investing heavily in genomics, AI-enabled drug discovery, and precision medicine. Yet every one of these ambitions ultimately depends on curated biological data resources reliant on emergency bridge grants, donations, staff secondments, and ad hoc rescue efforts.

## Keywords

Knowledgebase; DataBase; Science Policy; Science Funding; Artificial Intelligence; Biocuration

## Why these resources matter

### Strategic value

The Global Biodata Coalition ([globalbiodata.org](https://globalbiodata.org)) is a forum for the world's research funders to exchange knowledge and share strategies for supporting biodata resources. Towards this aim, it identifies the most widely used and critical nodes in the world's life sciences data infrastructure, the loss of which would be felt across the entire field. The inaugural list of 37 key resources, announced in December 2022, included three flagship resources based in the UK: PomBase ([www.pombase.org](https://www.pombase.org)) (Carme et al. 2026), FlyBase ([flybase.org](https://flybase.org)) (Marygold et al. 2026), and the HUGO Gene Nomenclature Committee (HGNC) database ([www.genenames.org](https://www.genenames.org)) (Seal et al. 2026). All three have faced existential funding crises in 2024-2026.

The strategic importance of these resources derives from the biological systems they support and curate. Model organisms such as *Schizosaccharomyces pombe* (fission yeast), supported by PomBase, and *Drosophila melanogaster* (fruit fly), supported by FlyBase, occupy a central position in biomedical research. These organisms enable rapid and cost-effective mechanistic discoveries that underpin much of modern experimental biology. Moreover, because basic genetic and cellular processes are deeply conserved across evolution, the knowledge generated from model organisms and curated by knowledgebases is directly translatable, informing human gene function, disease mechanisms, and therapeutic targeting.

The HGNC provides an equally essential function, effectively maintaining a “dictionary” of genes for the human genome. As the sole worldwide authority for human gene nomenclature, HGNC assigns standardised symbols (e.g., *BRCA1*) and names to human genes. Consistent naming allows knowledge to be reliably connected across publications, databases, analytical tools, and clinical systems. As genomics becomes increasingly embedded in diagnosis, drug development and precision medicine, this shared language is indispensable. Through its sister project, the Vertebrate Gene Nomenclature Committee, this standardisation extends across major vertebrate species, supporting comparative and translational research worldwide.

Collectively, these resources form part of the fundamental framework for biological knowledge, and both the benefits and consequences of their loss have been discussed extensively (Oliver et al. 2016, The Alliance of Genome Resources Consortium 2019, Bellen et al. 2021, Hall et al. 2026, Marygold et al. 2026). They enable researchers to identify candidate genes, reduce duplicated experimental effort, interpret large datasets and connect findings within and between species; they accelerate discovery by compressing weeks or months of exploratory work into hours.

## Hidden work and ecosystem multiplier effects

The value of these resources is not confined to the information they contain. It emerges from the network of standards, annotations, identifiers, ontologies and cross-species connections they sustain. Specialist curators read and evaluate primary research to produce structured, standardised, and computable knowledge. They reconcile conflicting findings, develop and maintain controlled vocabularies (ontologies), identify errors, and connect discoveries across genes, species and biological processes. This expert interpretation creates the trusted bedrock of knowledge that allows biological information to be integrated, reused and analysed at scale.

Because this foundation is embedded directly into the global data ecosystem, its effects propagate far beyond the originating databases. PomBase and FlyBase are the largest contributors to Gene Ontology (GO) annotation quality control (QC) through the review and validation of experimental GO annotations (Feuermann et al. 2025). Their annotations, models and QC efforts flow into major international resources, including the GO database, UniProtKB, Ensembl and RNAcentral, where they support research across thousands of species. GO curation from PomBase and FlyBase is also routinely propagated manually to proteins in other species using phylogenetic methods. By this mechanism, both organisms directly support over a million downstream annotations each in UniProtKB (Feuermann et al. 2025), and many millions more by the fully automated TreeGrafter pipeline (Tang et al. 2019). The HGNC provides the authoritative naming system for human genes, assigning standardised symbols and names to more than 45,000 genes. These identifiers are propagated through Ensembl, NCBI, UniProtKB, OMIM, ClinVar, ClinGen and virtually every major clinical genetics resource. What appear to be specialist databases are, in practice, critical nodes in the global interconnected research infrastructure.

This same infrastructure also underpins the reliability of Artificial Intelligence (AI) in biology. Modern AI systems are trained on biological knowledge assembled from resources such as PomBase, FlyBase, and HGNC. Because these systems learn directly from curated annotations, errors or inconsistencies in the underlying data can be amplified at scale. Conversely, high-quality curation improves the reliability, interpretability and reproducibility of computational models. This results in a direct dependency: these resources do not simply record biological knowledge; they actively shape the knowledge that future AI systems are trained on.

Taken together, these mechanisms form a single coupled system: expert curation generates structured knowledge from research results; that knowledge propagates through global databases and analytical pipelines; and those same structures increasingly define the datasets on which computational and AI systems depend. The impact is therefore multiplicative rather than additive, with improvements or failures in curation cascading across molecular biology, translational research and AI-driven discovery. These system-wide effects are achieved at relatively modest cost compared to the value they generate. Economic studies consistently demonstrate that biodata resources deliver exceptional returns on investment, with small sustained funding enabling disproportionately large downstream gains in research efficiency, reproducibility, and translational impact. Defunding them does not simply remove a database; it interrupts the flow of trusted knowledge through the scientific ecosystem.

## The crises

PomBase, FlyBase and the HGNC database have each faced existential funding crises between 2024 and 2026. PomBase recently lost its long-standing support from Wellcome following the discontinuation of its Biomedical Resource grants and has failed to secure replacement funding. It is now facing the real possibility of permanent closure. FlyBase lost funding for five curator positions when National Institute of Health (NIH) overseas subawards were terminated in 2025, surviving only through emergency support from Wellcome, the University of Cambridge and community donations. It has since launched a voluntary annual contribution model to fund staff and accelerated the transition of its data into the Alliance of Genome Resources ([www.alliancegenome.org](http://www.alliancegenome.org)) (The Alliance of Genome Resources Consortium 2024). HGNC endured a prolonged funding gap following the withdrawal of Wellcome Biomedical Resource grants, resulting in the temporary redeployment or loss of team members. It has since secured a new NIH award, but is now wholly dependent on US funding at a moment when that funding is under acute pressure. Although each crisis had different immediate causes, together they reveal a common vulnerability in the current UK funding model that sustains critical biological data infrastructure, one mirrored in funding systems elsewhere.

The databases discussed in this piece are not outliers, and this problem is not specific to the UK; for example, NIH support for model organism databases has been cut by roughly 50% since 2016 (Bellen et al. 2021), driving their consolidation within the Alliance of Genome Resources. The Arabidopsis Information Resource (TAIR), the central plant model organism database, lost its US National Science Foundation funding more than a decade ago and has survived only by moving

to a subscription model (Reiser et al. 2016). In 2024, the National Institute of Allergy and Infectious Diseases (NIAID) declined to renew the contract supporting VEuPathDB, the primary genomic resource for eukaryotic pathogens and their vectors, including malaria parasites and mosquitoes. The decision alarmed the global infectious disease community (Fernandez-Prada et al. 2024), and since 2025, the resource has moved towards a subscription model. The pattern is repeated across countries and disciplines: resources on which many researchers rely lose stable funding and are pushed towards paywalls, donations or one-off rescues, all of which weaken open and sustainable access to research data worldwide.

These crises point to another serious risk: the loss of curation expertise that would take decades to rebuild. The scientists who maintain these databases are domain specialists who have spent careers learning to reconcile conflicting annotations, resolve ambiguous nomenclature, develop standards and judge which submitted records are trustworthy. Training a replacement generation requires not just funding but continuity, time, and accumulated community trust. Without ongoing expert intervention, no new research is curated, vocabularies are not updated, and the curated layer, the interpretation, the cross-references, and the emergent knowledge begin to stagnate and degrade in value the moment the curators leave. Unlike physical infrastructure, expertise cannot simply be switched back on when funding returns.

## Structural failures

These crises are not isolated events but symptoms of a broader structural problem. Three factors have combined to create the current vulnerability.

**First, funder priorities have shifted.** This is illustrated by Wellcome's 2020 strategy, which moved away from long-term support for open-ended infrastructure, closing funding routes that had sustained many databases for years.

**Second, funding has not kept pace with need.** Around the world, support for biological data resources has typically been decreasing, even as research relies on them more. The UK shows the problem clearly: its principal funding mechanism for biological resources has remained flat at approximately £6-7 million per year for nearly two decades, and it has not compensated for the £16 million per year that Wellcome previously provided.

Over the same period, genomics, systems biology, AI and precision medicine have matured, and the associated biological data and knowledge have grown dramatically in volume and complexity. Research everywhere now depends far more on curated biological resources and the infrastructure that sustains them.

**Third, institutional incentives remain misaligned with infrastructure stewardship.** Universities are rewarded for papers and grants, while maintaining shared scientific infrastructure receives little recognition. The result is a persistent bias towards generating new research rather than sustaining the resources that make research possible.

Together, these pressures have left many essential databases dependent on a succession of temporary rescues rather than a stable, long-term funding model.

## Recommendations

Several priorities emerge from this commentary:

**Fund infrastructure as infrastructure.** Biological databases operate on long-term horizons that are misaligned with three- to five-year grant cycles and should be treated as research infrastructure, comparable in importance to telescopes and synchrotrons. Although alternative funding models have long been proposed, few have been tested or adopted at scale. Promising approaches, such as collective funding models in which agencies contribute a fixed fraction of their research budgets, or coordinated multi-funder support, must now move from proposal to pilot as an urgent priority.

**Assess systemic risk in funding decisions.** Funder portfolio reviews should explicitly consider the systemic risks posed by the weakening of core data infrastructure to the broader ecosystem of tools, databases, and AI systems built on top of it.

**Reward stewardship in research assessment.** These resources remain systematically undervalued because research assessment systems reward discovery more than stewardship. Research assessment criteria should explicitly recognise the strategic value of hosting and sustaining knowledgebases, thereby creating institutional incentives that currently do not exist.

## Conclusion

Curation is a foundational aspect of modern life science, shaping how biological insights are interpreted, connected and reused across the research ecosystem. Biodata resources, therefore, extend the value of research outputs far beyond individual publications by transforming static findings into reusable, interoperable knowledge that supports ongoing discovery. Yet this infrastructure remains fragile and is shrinking due to uneven and often precarious funding. Without more deliberate and sustained investment from funders, institutions, and policymakers, the life sciences risk overextending their ambitions while under-supporting the systems that make those ambitions achievable, ultimately limiting the pace and undermining the reliability of future discoveries.

## References

- Alliance of Genome Resources Consortium. 2019. The Alliance of Genome Resources: building a modern data ecosystem for model organism databases. *Genetics*. 213:1189–1196. doi:10.1534/genetics.119.302523.
- Alliance of Genome Resources Consortium. 2024. Updates to the Alliance of Genome Resources central infrastructure. *Genetics*. 227:iyae049. doi:10.1093/genetics/iyae049.

- 211 Bellen HJ, Hubbard EJA, Lehmann R, Madhani HD, Solnica-Krezel L, Southard-Smith EM.  
212 2021. Model organism databases are in jeopardy. *Development*. 148:dev200193.  
213 doi:10.1242/dev.200193.
- 214 Carme P, Rutherford K, Wood V. 2026. PomBase in 2026: expanding knowledge, modeling  
215 connections. *Genetics*. 232:iyag001. doi:10.1093/genetics/iyag001.
- 216 Fernandez-Prada C, Moretti N, do Monte-Neto R. 2025. Critical loss: the effects of VEuPathDB  
217 defunding on global health. *Lancet Microbe*. 6:100980. doi:10.1016/j.lanmic.2024.100980.
- 218 Feuermann M, Mi H, Gaudet P, Muruganujan A, Lewis SE, Ebert D, Mushayahama T, Gene  
219 Ontology Consortium, Thomas PD. 2025. A compendium of human gene functions derived from  
220 evolutionary modelling. *Nature*. 640:146–154. doi:10.1038/s41586-025-08592-0.
- 221 Hall N, Röper K, Wood V, Nurse P. 2026. Treat biological databases as infrastructure, not  
222 projects. *Nature*. 656:786. doi:10.1038/d41586-026-02575-5.
- 223 Marygold SJ, Attrill H, dos Santos G, Gramates LS, McLachlan A, Millburn G, Tabone CJ, Zaru  
224 R, FlyBase Consortium. 2026. FlyBase: development and metamorphosis into the Alliance.  
225 *Genetics*. iyag247. doi:10.1093/genetics/iyag247.
- 226 Oliver SG, Lock A, Harris MA, Nurse P, Wood V. 2016. Model organism databases: essential  
227 resources that need the support of both funders and users. *BMC Biol*. 14:49.  
228 doi:10.1186/s12915-016-0276-z.
- 229 Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, Mezheritsky Y, Vetushko A, Huala E.  
230 2016. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a  
231 case study of a subscription-based funding model. *Database (Oxford)*. 2016:baw018.  
232 doi:10.1093/database/baw018.
- 233 Seal RL, Braschi B, Gray K, McClay J, Tweedie S, Bruford EA. 2026. Genenames.org: the  
234 HGNC and PGNC resources in 2026. *Nucleic Acids Res*. Volume 54, Issue D1.  
235 doi:10.1093/nar/gkaf1329.
- 236 Tang H, Finn RD, Thomas PD. 2019. TreeGrafter: phylogenetic tree-based annotation of  
237 proteins with Gene Ontology terms and other annotations. *Bioinformatics*. 35(3):518–520.  
238 doi:10.1093/bioinformatics/bty625.