Who Will Keep Research Data Infrastructure Open and Running?

JENNIFER GIBSON AND KAITLIN THANEY

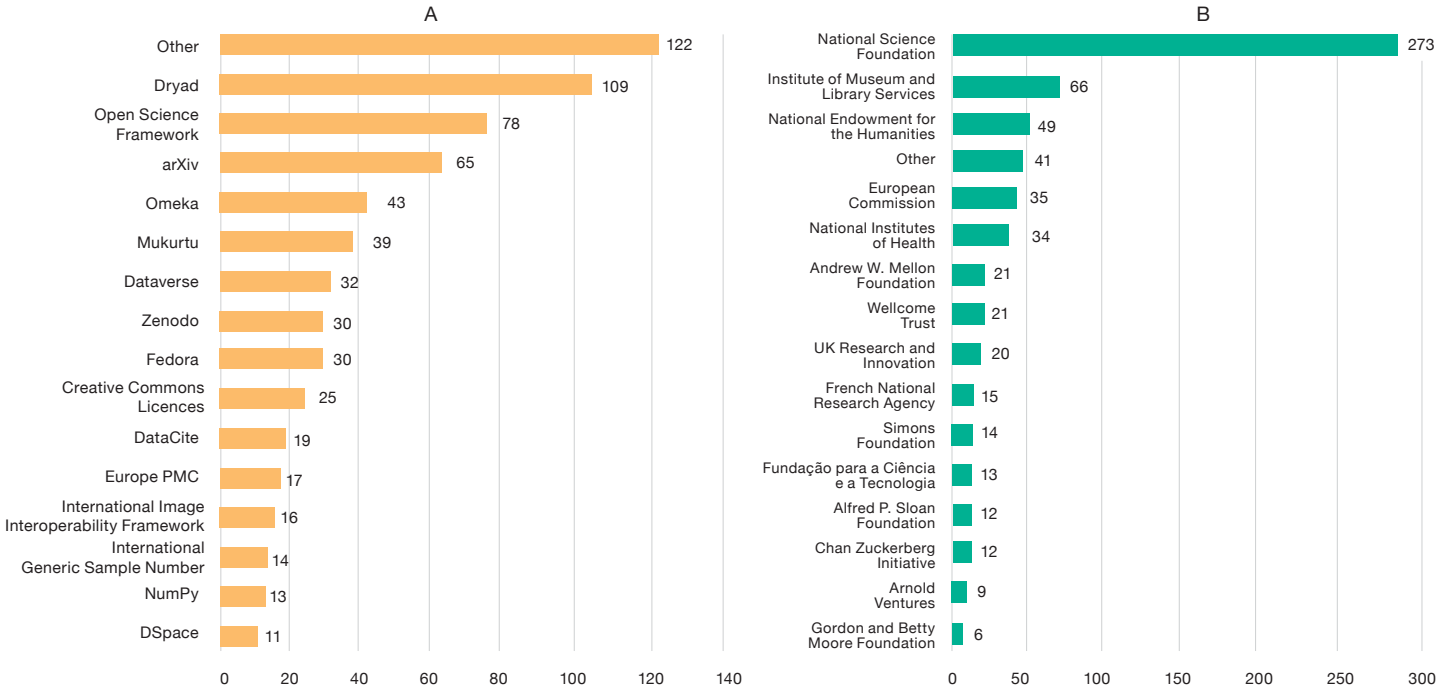
The global research enterprise relies on information infrastructure to power scientific discovery, medical breakthroughs, and evidence-based policymaking. But the data repositories, digital asset management services, and preservation systems that ensure research data remains open and accessible are often overlooked—until they disappear. Many of these tools and services are vulnerable to policy changes and funding cuts. Over the last 25 years, nearly 200 research data repositories have shut down permanently; more than half of those closures have happened since 2018.

Each closure represents lost knowledge and leads to broken links, bad citations, and a general inability to utilize and verify scientific findings. For example, without funding from the National Oceanic and Atmospheric Administration, the Alaska Earthquake Center has ceased providing real-time seismic data to inform tsunami warnings for the whole US West Coast. On topics as disparate as Gulf War illness or natural selection, when a repository goes dark, it can affect individuals or even entire research disciplines.

And because repositories and other open science infrastructures are commonly designed to support transboundary research, their collapse can have compounding global effects. In early 2025, the United States Agency for International Development (USAID) suspended access to the Demographic and Health Surveys (DHS) Program databases, a repository containing decades of population, health, HIV, and nutrition data from more than 90 countries. Almost overnight, researchers in Malawi lost access to critical data informing antiretroviral therapy programs serving roughly one million HIV-positive patients; researchers in Nigeria had nowhere to store new data designed to identify causes of maternal deaths; and the release of complete data from a 2023–2024 key indicators survey in the Democratic Republic of the Congo was delayed for months. After USAID was dismantled in the first half of 2025, an emergency grant from the Gates Foundation restored access to existing DHS data and selected surveys. But this three-year support has not returned the program to its prior scale, leaving 23 countries with surveys still incomplete or unanalyzed.

The threat goes beyond federally funded data repositories. Upheavals to university research over the last year have put new budgetary pressures on the institutions that directly or indirectly support many open infrastructure solutions, like arXiv's open-access platform, dSPACE's repository software, and Dryad's open-data publishing

Figure 1. **TOP 15 OPEN INFRASTRUCTURE AWARD RECIPIENTS FROM 2001–2024 (A) AND TOP 15 FUNDERS OF TOTAL AWARDS (B) OVER THE SAME PERIOD.**



The total number of awards (direct, adjacent, and other) received is greater than the number of awards given because awards to multiple infrastructures are counted toward each infrastructure's total. Infrastructures that received fewer than 10 awards are grouped under "Other." Data for all figures comes from Invest in Open Infrastructure's 2025 State of Open Infrastructure *report. Sarah Lippincott and Lauren Collister conducted the analysis and generated the figures.*

platform—many of which already operate on razor-thin margins. Universities' cuts to open infrastructures may appear nominal on their balance sheets—$8,000 from one, $5,000 from another, $50,000 from another—but these seemingly modest reductions collectively dismantle the operating capacities of systems that would cost millions to rebuild.

The movement to support open science and research has mainly focused on breaking down cultural and institutional barriers that block researchers from participating in open scholarship. But there has been less communal thinking around the longevity and resilience of open research infrastructure—when and why it might fail, what risks failures represent for the broader community, and how to prevent them. This conversation urgently needs to begin today, to avoid the costs associated with further setbacks to scientific progress. As practitioners with many years of experience building and directing open research systems and services, we recognize this infrastructure as a utility that is essential to the success of the whole research community. What the community needs now is a plan to support this utility with sustained operational funding.

An infrastructure crisis

The economic returns on investments in open data infrastructure are remarkable and thoroughly documented. A comprehensive review found consistent evidence across studies showing substantial returns through efficiency gains, enhanced innovation, and economic growth. Open scientific infrastructures are also much less expensive than commercial equivalents: One study estimated savings of 87–94%.

Returns on investment in open infrastructure are more commonly analyzed on a project-by-project basis. For instance, the Human Genome Project, a foundational government investment in open research infrastructure, delivered returns of $141 for every dollar invested through new medicines, products, services, and employment. The European Bioinformatics Institute's openly accessible databases generate an estimated £1.3 billion annually in research value, with user time saved by not having to recreate datasets worth almost £6 billion per year. A 2005 study commissioned by NASA found that every $1 invested in open geospatial standards saved $1.19 in long-term operations and maintenance costs, compared to proprietary standards. For every dollar the federal government invested in the XSEDE research cyberinfrastructure between 2016 and 2022, it enabled research outcomes valued between

Figure 2. **ADJACENT AND DIRECT FUNDING OVER TIME (2001–2024) ACROSS ALL FUNDERS AND FOR ALL OPEN INFRASTRUCTURES.**

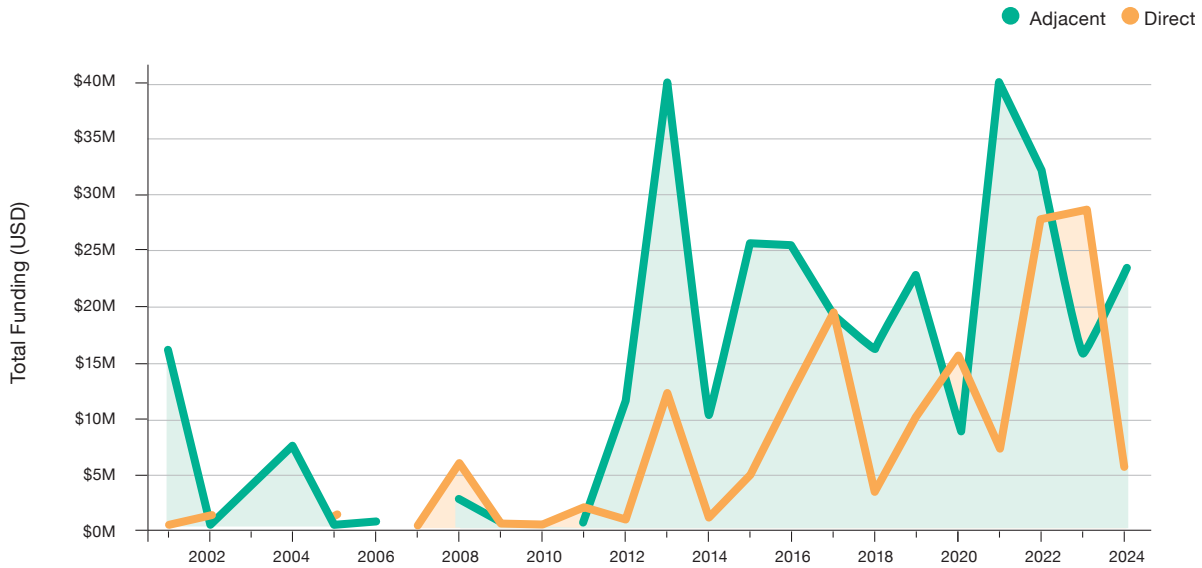


Figure 3. **PROPORTION OF ADJACENT AND DIRECT FUNDING FOR OPEN DATA INFRASTRUCTURES, 2001–2024.**

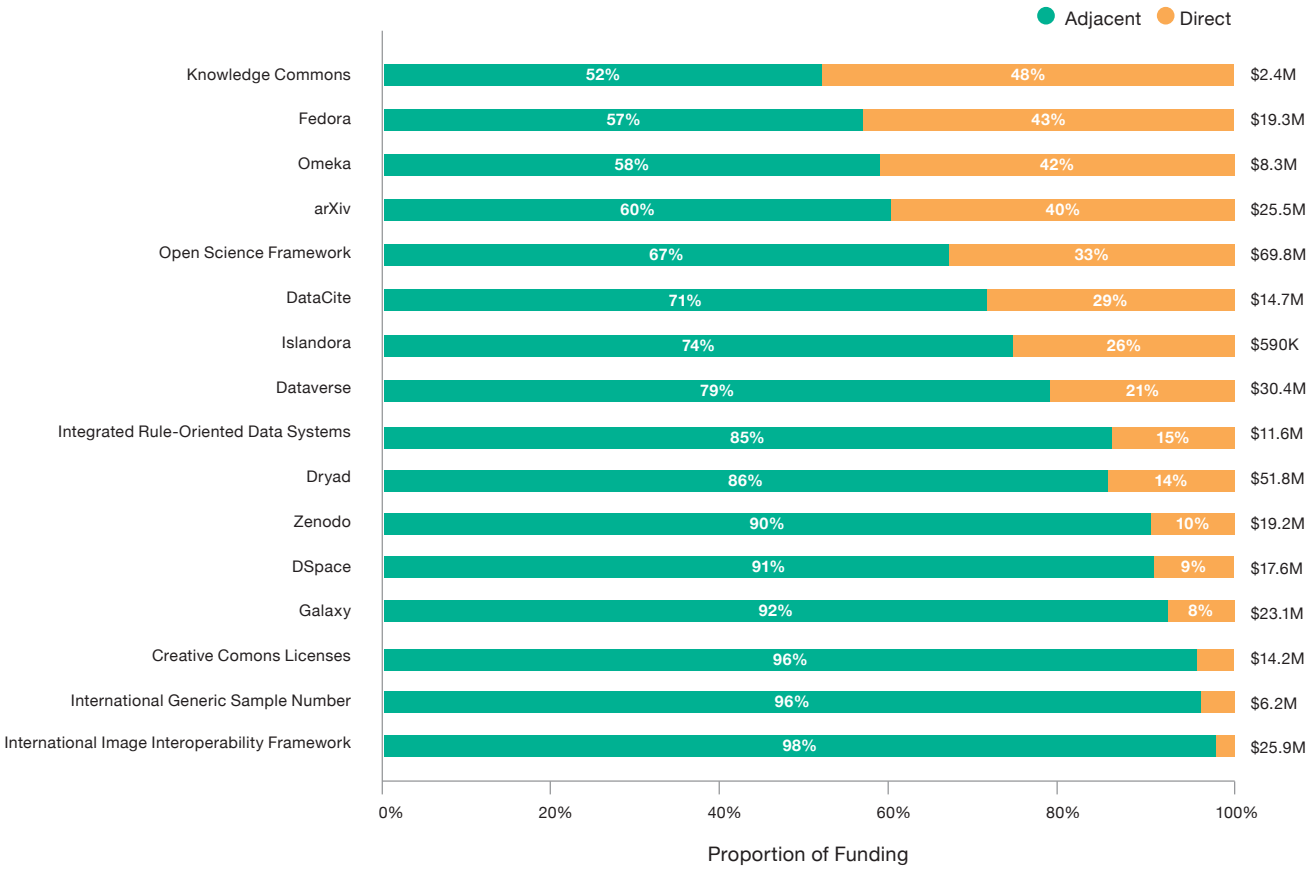
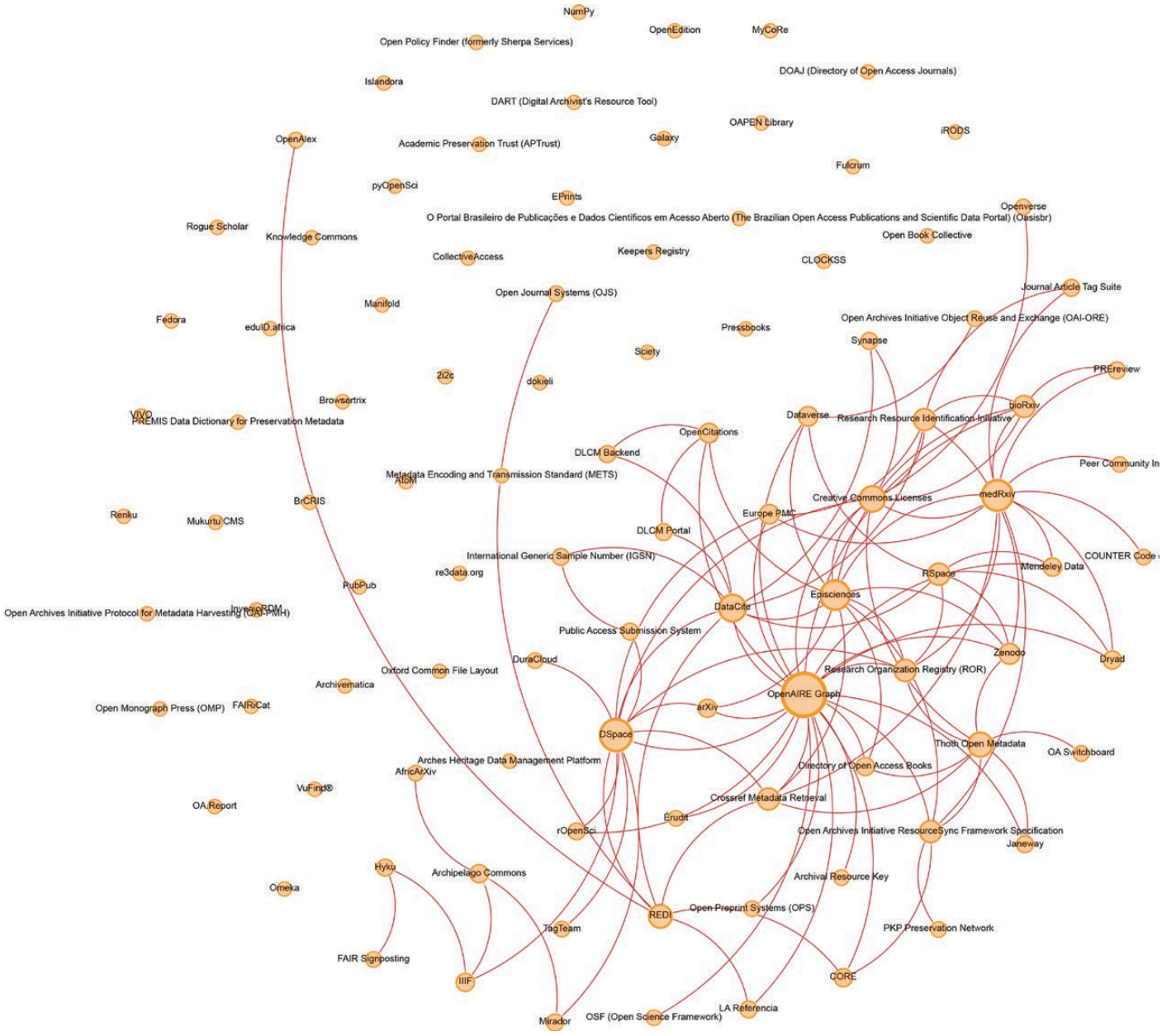


Figure 4. **NETWORKS OF RELIANCE AND INTEROPERABILITY BETWEEN OPEN INFRASTRUCTURE TOOLS, SERVICES, AND STANDARDS.**



$15 and $75 in scientific contributions and end products.

Perhaps the most powerful example emerged in October 2024, when the Nobel Prize in Chemistry was awarded for cracking the code of proteins—the molecules that "control and drive all the chemical reactions that together are the basis of life." That breakthrough would not have been possible without 40 years of open, human-curated protein data.

When data repositories and other open systems are supported, brilliant minds combining protein data with environmental, clinical, and behavioral data can generate medical, economic, and societal breakthroughs we cannot yet imagine. But few researchers really concern themselves with the operating models of the systems, tools, and services that make their ideas accessible to others.

Providers of open research infrastructure are funded by myriad sources: commitments from universities, research institutes, and libraries; grants from federal agencies and international scientific and philanthropic organizations; and donations. The *2025 State of Open Infrastructure Report* documented that 28 funders made 641 awards to 54 open infrastructures between 2001 and 2024.

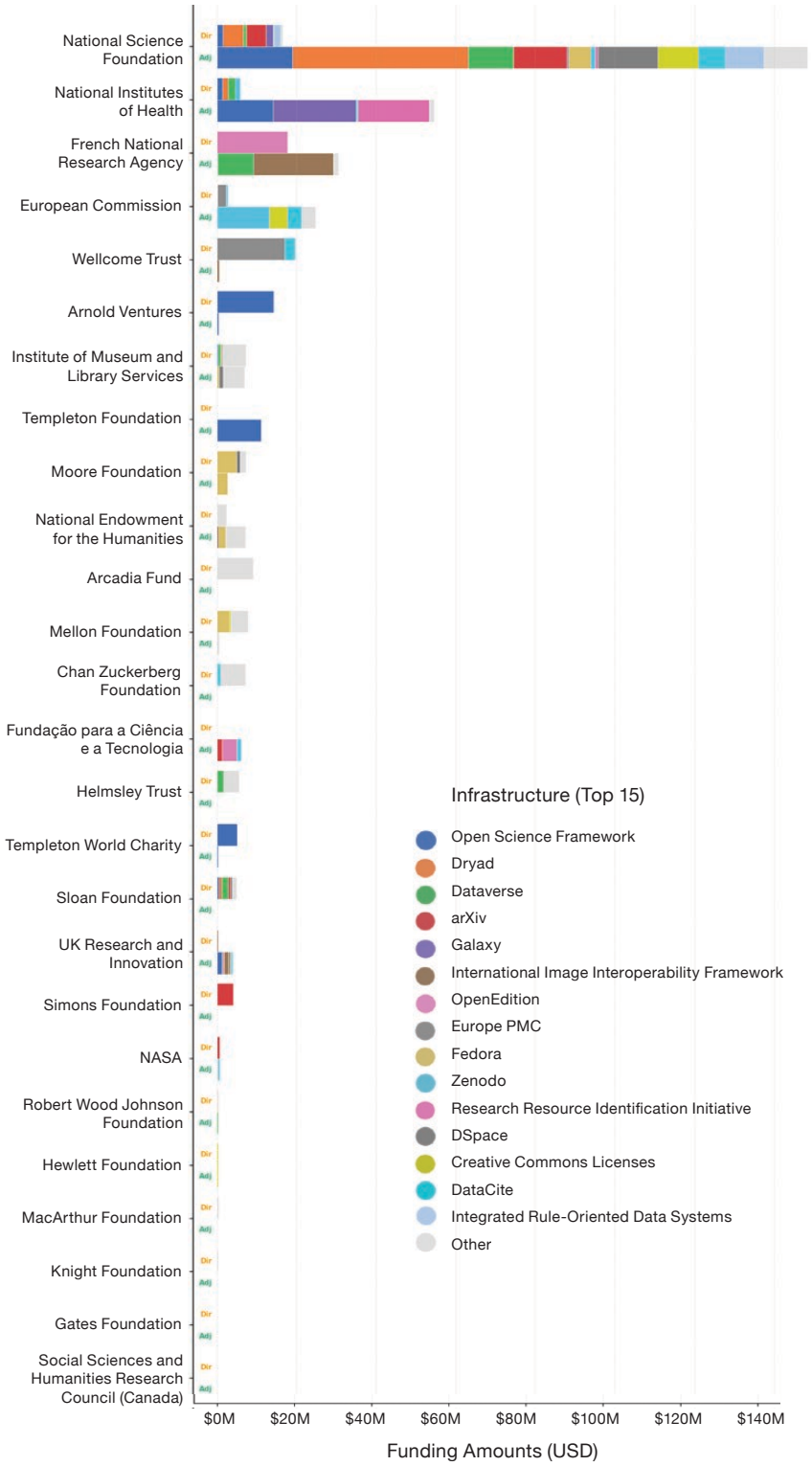
But a more nuanced breakdown of the finances reveals the true structural challenge: Funding that directly supports the operations, enhancements, and growth of particular infrastructures is dwarfed by funding granted to users of a particular system. In other words, users can secure funding to use a tool like Dryad or Zenodo to conduct research based on its holdings, but Dryad and Zenodo receive little to no financial support from that transaction. Researchers may cite an infrastructure as a necessary part of their work in a grant proposal, but funding flows only to the researcher and research project, not the repository. This disparity is defined as the difference between direct support (awards made directly to the infrastructure) and adjacent support (awards that rely on the infrastructure but are made to some other recipient). Adjacent support can therefore not be used for a repository's operations, development, or maintenance.

Another structural challenge is the mismatch between the support funders are willing to provide and the support these services actually need. Infrastructure providers frequently cite a need for operational and adoption support, but the *2025 State of Open Infrastructure Report* found that 65% of total funding goes to research and development efforts—discrete innovation projects or feature enhancements. Out of the vast ecosystem of funders for open infrastructure, only a handful support the bulk of direct operational expenses.

The result is a stealthily compounding problem: Infrastructures invoked as fundamental to the conduct of modern research at scale are being funded on the periphery. Users can secure grants to conduct research built entirely on these holdings, while the infrastructure itself receives comparatively little, and in some cases nothing at all. When usage puts demands on infrastructure without providing corresponding financial support, each new research project that names a repository in its abstract may deepen the sustainability risk. High citation is not the same as adequate support, and mistaking one for the other may be one of the most consequential gaps in how we currently fund open science.

Infrastructure isn't just about innovation; it's a utility communities rely on. In the construction of a laboratory, an institution might spend a million dollars on construction, but everyone understands annual operating funds are needed for utilities, maintenance, and repairs. Data infrastructure is comparable. Building it requires ongoing yearly investment in scalable storage, curation and quality control, format migration, security updates, and user support.

Figure 5. **ADJACENT AND DIRECT FUNDING BY FUNDER AND INFRASTRUCTURE BY AMOUNT (USD), 2001–2024.**



These totals do not reflect awards in which multiple infrastructures are named as it is not generally possible to determine how this money is distributed between infrastructures.

These costs are far from enormous. The annual operating budget for Dryad, an open data publishing platform that ensures persistent access to over 50 million files and hosts evidence for an estimated 50,000 published reports from the last 20 years, is less than $2 million. Allocating 10–15% of research grants directly to data infrastructure would transform the ecosystem.

But beyond simply increasing the proportion and amount of direct funding, there are other actions stakeholders across the ecosystem can take to stabilize the funding model for open infrastructures. First, direct support for research infrastructure should be a specified line item in the budgets of all research institutions, much like facilities and administrative costs. Data infrastructure support should be explicit when institutions receive major research grants, making the dependency—and the investment—visible to all stakeholders.

Second, philanthropic organizations should move beyond one-off grants toward sustained operational funding. This funding includes more strategic and creative forms of support, like pooling amounts from multiple funders to increase investment, share risk, reduce siloed decisionmaking, and direct investment in business planning at the repository level to ensure sustainability.

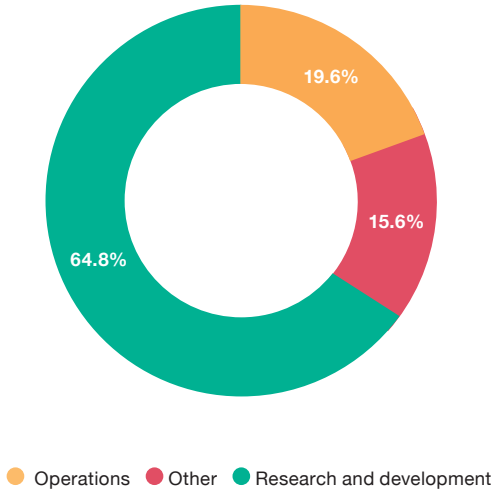
Third, universities and university associations need to coordinate open infrastructure investment and divestment decisions. Universities can leverage existing networks and consortia like the Big Ten Academic Alliance, nonprofit membership organizations like Lyrasis, and library-to-library collaborations like the Partnership for Academic Library Collaboration and Innovation and the Private Academic Library Network of Indiana to reconstitute purchasing strategies, collective investment, and processes that support open infrastructures.

Finally, no essential resource should depend on single-source funding. Open infrastructure providers should strive to diversify funding sources to the extent possible. For some infrastructures, this aim might mean combining unrestricted and service-based fees with grants. Other providers could take different approaches. But providers should have a shared focus on eliminating critical dependencies on individual revenue streams.

Protecting an investment

A convergence of forces is putting unprecedented pressure on research infrastructure. The fallout from cuts to federal research funding and capacity has put institutions across the enterprise on edge, revealing the tightly networked and interdependent nature of the ecosystem and the fundamental instability of relying on short-term, single-source funding. At the same time, federal agencies continue to create or uphold mandates for open science, requiring that research data be made openly accessible but providing no sustained funding for the repositories and systems that make this possible.

Figure 6. **DISTRIBUTION OF DIRECT FUNDING BASED ON AMOUNT (USD) TO INFRASTRUCTURES BY CATEGORY, 2001–2024.**



Meanwhile, the artificial intelligence boom is creating massive, unforeseen demand for the knowledge troves held in research infrastructures. Large language models are built on open training data; without the repositories, governance, and systems that ensure data quality, accessibility, and longevity, AI will become a storyteller rather than an engine of scientific progress.

At this inflection point, a more strategic approach to ensuring the resilience and longevity of the knowledge systems that accelerate discovery and translation is necessary. If the federal government fails to act, it risks losing out on its own investments in open infrastructure. European countries are actively building alternatives to United States–based research infrastructure for sovereignty and security reasons. Between 2021 and the present, the European Union has invested €1.2 billion in 177 research infrastructure projects, much of it in open science systems like the European Open Science Cloud. Likewise, powerful commercial providers can easily strip remaining, community-embedded solutions from the enterprise, perpetuating issues of pricing, monopoly power, and control over the scientific agenda by consolidating services under their own umbrellas. A healthy system needs both commercial and noncommercial providers to maintain accountability.

The infrastructure that makes science possible is not separate from science itself. To realize the benefits of research in the modern age, that infrastructure must be supported and sustained for the future.

Jennifer Gibson *is executive director of Dryad.* ***Kaitlin Thaney*** *is executive director of Invest in Open Infrastructure.*