Database under maintenance

Managing the growth in biomedical data requires coordinated strategies and a strong financial commitment by funders and institutions.

Everyone agrees: data from publicly funded research should be made available once the results are published, and sooner if possible. Yet the surge in biological data poses immense difficulties for storage, maintenance and curation, including a mounting price tag. These challenges can be met so long as data sustainability becomes a stated objective with a dedicated strategy, rather than an afterthought of experimental discovery.

Over the past two decades, the number of molecular biology databases has multiplied from fewer than 200 to nearly 1,700. The US National Center for Biotechnology Information (NCBI) holds about 20 petabytes of data, and this is expected to increase by around 50% this year alone. Rolf Apweiler, joint director of the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), estimates that biological data doubles every 12 to 18 months. The problem is compounded by the increasing complexity of data, which demands new data architectures, integration and high-quality curation.

Well-maintained data resources are essential for everything from routine queries to data mining and meta-analysis. A Correspondence in this issue by Jüri Reimand and colleagues (p 705), for example, underscores the dramatic impact that outdated gene annotations can have on pathway analysis. The US National Institutes of Health (NIH) plans to require that all research funding applications include data-management strategies with provisions for reuse and long-term preservation.

Many solutions to the explosive growth in data have evolved in an *ad hoc* manner in response to individual research projects and communities, creating problems in the long term.

The mechanisms used to fund databases are a good example. Outside of large centers, data repositories are often supported via standard grant applications. "We're funding resources as research projects," concedes Philip Bourne, associate director of data science at the NIH. A focus on innovation makes sense for launching resources, but not for maintaining them. Every grant renewal requires that developers add new functionality, which increases costs and can even compromise core utility. Short grant cycles also jeopardize stability. The Kyoto Encyclopedia of Genes and Genomes (KEGG), which receives over half a million page visits a month, faced an existential crisis in 2011 when it could no longer cobble together short grants as it had for the past 16 years.

More direct government funding of community resources is also vulnerable. After 14 years of support, the US National Science Foundation withdrew funding from the Arabidopsis Information Resource (TAIR). The National Human Genome Research Institute recently asked databases that it supports, including FlyBase and Online Mendelian Inheritance in Man, to develop new funding models within the next four years.

These highly curated resources may have a better chance of being financed by user fees than larger raw data repositories. When funding for TAIR ended in 2013, its curators formed a non-profit organization that adopted a tiered subscription model and has so far been successful in maintaining the database's quality and user base. KEGG now charges for its FTP service, which, in addition to some institutional and government support, enables free access to less heavy users. Tiered pay models can spread costs across countries and companies, but they risk excluding less well-off researchers and can impede data sharing.

The proliferation of smaller-scale databases in growing fields often results in duplicated effort, whereas centralized platforms like the NCBI and EBI are well placed to implement technological improvements such as automated annotation and data compression and to leverage their size to secure cheaper storage. Databases that serve smaller communities can consolidate or team with larger centers to exploit these benefits; the key challenge is to maintain enough flexibility that the needs and expertise of scientists in their fields are not lost. This must be tied to a long-term commitment to fund data curation.

Greater efficiency can also be achieved through better coordination within funding bodies and between data centers. The ELIXIR program, supported by the European Commission, was recently designed to implement data standards, reduce redundancy and provide scalable distributed infrastructure for data deposition and sharing across European countries. The NIH is also currently assessing how data is dealt with across its organization and exploring cloud-based alternatives in its Commons framework. These are welcome developments that give hope for data-driven and forward-looking solutions.

Apweiler estimates that the current global costs of maintaining public biomedical data repositories are under $300 million annually—a fraction of a percent of budgets for data generation. Meanwhile, their economic impact is estimated to be many times greater than the investment, and their effect on research is incalculable. With a combination of dedicated funding strategies and a robust coordination of effort, it should be possible to sustain the growth of these critical resources well into the future.