COMMENTARY

How do your data grow?

Scientists need to ensure that their results will be managed for the long haul. Maintaining data takes big organization, says **Clifford Lynch**.

Data can be 'big' in different ways. National and international projects such as the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva in Switzerland, or the Large Synoptic Survey Telescope planned for northern Chile, are frequently cited for the way they will challenge the state of the art in computation, networking and data storage. But research data can also be big by being of lasting significance — a clinical-trial result, or the observation of a unique event. Data can be big because of descriptive challenges that may require context such as the experimental set-up. Because digital data are so easily shared and replicated and so recombinable, they present tremendous reuse opportunities, accelerating investigations already under way and taking advantage of past investments in science.

To enable reuse, data must be well preserved. In some cases the effects of data loss are economic, because experiments have to be re-run. In other cases, data loss represents an opportunity lost forever. Funders now rightly view data as assets that they are underwriting and so seek the greatest pay-off for their investments. They demand that researchers and host institutions document and implement data-management and data-sharing plans that address the full life cycle of data — including what happens after a grant finishes. Host universities thus find themselves with legal and ethical obligations to provide a legacy of faculty data. Publishers must also identify the most effective ways to connect publications with data and preserve the scientific record.

Developing infrastructure

Managing the life cycle of scientific data presents many challenges. These include deciding responsibilities, funding, resource allocation, what data should be kept and for how long.

In a sense, landmark international projects like the LHC are the least problematic: the costs of data management are explicit in the budget and tend to be dominated by technology expenses that decline over time. These projects also include dedicated personnel; and, although the volume of data is often vast, the streams fit within well defined descriptive schemes.

But science's reliance on digital data extends far beyond these international projects. Funding programmes in Europe and the United States, for example, have invested substantially in common infrastructure for a more systematic reliance on data, networks and computation. And there are vast numbers of scientific research projects producing at most a few terabytes per year of big data, or data that can be aggregated into a big-data resource. Funding, support expertise and structuring the data for long-term management can be problematic for these projects. This has been shown in recent years by studies of faculty information management needs through a wide range of academic disciplines[1,2].

The challenges here are great, and will only be solved by focused effort and collaboration between funders, institutions and scientists.

Community standards for data description and exchange are crucial. These facilitate data reuse by making it easier to import, export, compare, combine and understand data. Standards also eliminate the need for each data creator to develop unique descriptive practices. They open the door to development of disciplinary repositories for specific classes of data and specialized software management tools. GenBank, the US National Institutes of Health (NIH) genetic sequence database, and the US National Virtual Observatory are good examples of what is possible here. In 2007, the US National Science Foundation, recognizing the importance of such standards, established the Community Based Data Interoperability Networks (INTEROP) funding programme for the development of tools, standards and data management best practices within specific disciplinary communities. INTEROP should make its first awards this autumn. Although many classes of scientific data aren't ready, or aren't appropriate, for standardization, well chosen investments in standardization show a consistently high pay-off[3].

> "The best stewardship of data will come from engagement with preservation institutions."

At the start of the data life cycle, individual scientists will have primary responsibility for stewardship. But longer term, data preservation can only be done by institutions. If data are to be consolidated or shared on a frequent basis, there is a lot to be said for moving to institutional control sooner rather than later. Scientists are not necessarily good data managers and can more fruitfully spend their time doing science. Moreover, it is unfair and unreasonable — and increasingly ineffective — to assign long-term information management tasks to a rotating staff of students and postdocs. Indeed, as specific data sets become distant from current research activities, stewardship can become a tax on scientific productivity.

Scientists need to act responsibly during their stewardship. This includes working through and honouring disciplinary standards. It also includes defining and recording appropriate metadata — such as experimental parameters and set-up — to allow for data interpretation. This is best done when the data are captured. Indeed, descriptive metadata are often integrated within the experimental design. Description includes tracing provenance — where the data came from, how they were derived, their dependence on other data and all changes made since their capture. Proper stewardship requires documenting the storage formats. These may be community standards, or they may be locally defined and often tied to locally developed software. It is desirable to keep versions of such software along with the data sets.

If data cannot survive in the short term, it is pointless to talk about long-term use. In a high-threat environment such as a major university's network, machines will often be compromised if updates aren't applied; this can mean data destruction or corruption. Disasters such as Hurricane Katrina, which destroyed labs and computing facilities, are important reminders that data need to be backed up frequently and comprehensively in diverse and distant locations. Appropriate use of IT services such as secure storage or hosting from the host institution may be valuable. In the longer term, digital data is at risk from various forms of technological obsolescence (particularly if locally held removable storage media are used). There is a need for new institutional services that can help with all these needs, handling traditional IT issues and information-management issues more familiar to librarians and archivists.

At some point, the primary copy needs to migrate to an institutional service. Today, these services are sparse. In the United Kingdom there are data services associated with several of the science-funding councils. Both NASA and the European Space Agency have planetary

D. ALLISON

science archives into which they place mission data. And in the United States, for example, there are also other focused archives connected to some disciplines, including the collections at the NIH's National Center for Biotechnology Information, the social science archive of the Inter-University Consortium for Political and Social Research based at the University of Michigan in Ann Arbor, and the Protein Data Bank, which holds structural data for proteins and nucleic acids. These are somewhat mature and have relatively stable funding.

New disciplinary repositories are also springing up, and some universities are setting up broad-based multidisciplinary repository services, usually working through the campus research library, to manage their faculties' research data. The National Science Foundation is preparing to make its first awards under an Office of Cyberinfrastructure programme called Datanet that will invest around US$100 million over the next five years for building data-stewardship capabilities; the grants will go to large university-led consortia. There are possible roles here for publishers and scholarly societies, but at present it seems as if in most disciplines, leadership will fall to stewardship services run by universities and government agencies.

These newer institutionalized data stewardship services — whether structured along university or disciplinary lines — are still immature. The handing over of data for deposit is not simple or well defined, and necessary community standards are lacking. Funding models are sketchy. Although stewardship needs to be funded, funding agencies are not eager to pay. Educational institutions are equally reluctant to make open-ended commitments. Perhaps, ultimately, this can be factored into overhead cost negotiations. Effective structures are needed to manage limited resources; not everything can be preserved forever, and we need methods for prioritization.

Ultimately, the best stewardship of data will come from disciplinary engagement with preservation institutions. General-purpose data management as provided by universities through their research libraries will have its limits. Where there is no natural locus of disciplinary stewardship, universities will need to establish consortia to enable disciplines to create and sustain such engagement[4].

The time is right for scientists to take stock of the institutionalized data services that are available or under development, to understand how these institutions are governed and financed, and to make choices about the best strategies for their disciplines. Can a discipline-oriented solution work? If a university-based system seems more practical, what can be done to expedite the move to university consortia strategies? As the volume of data, and the need to manage it grows, disciplinary consensus leadership will be very powerful factors in addressing the challenges ahead. ■

Clifford Lynch is the executive director of the Coalition for Networked Information, 21 Dupont Circle, Washington DC 20036, USA, and an adjunct professor at the School of Information, University of California, Berkeley, California, 94720-4600, USA.
e-mail: cliff@cni.org

1. www.lib.umn.edu/about/scieval/documents.html
2. www.library.ucsb.edu/informatics/documents.html
3. www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/
4. ARL Workshop on New Collaborative Relationships Report to the National Science Foundation. *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering* (2006).

See Editorial, page 1.
Join the discussion at http://tinyurl.com/6eedyu.