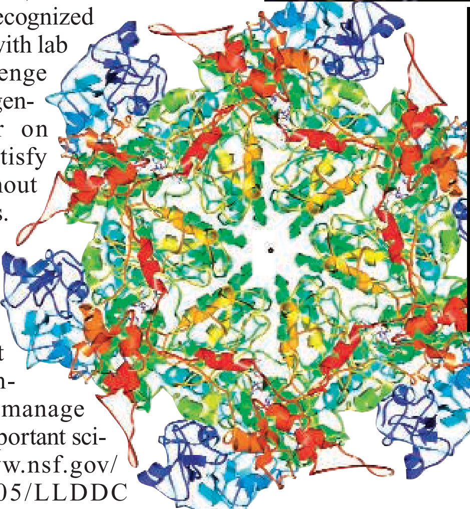
Boom in Digital Collections Makes a Muddle of Management

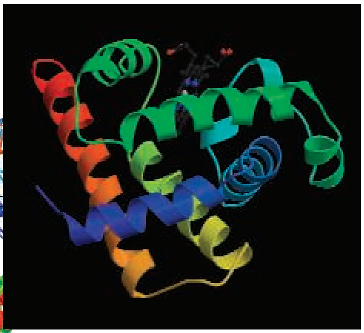
Electronic collections are a huge boon to scientists. But a new report says NSF needs to pay more attention to how they are funded and operated

Forget trays of preserved insects with their informational tags, as well as collections of rocks, fossils, and other samples from nature's treasure chest. Data have gone digital, and researchers from all walks of science—from climate modelers to systematists—are stockpiling their observations in newly created databases accessible to everyone through the World Wide Web. But as researchers head full speed into the digital world, the U.S. National Science Foundation (NSF) wants to ensure that they don't run out of gas along the digital highway and that the rules of the road are clear to everyone.

To date, NSF has not been tracking its total commitment to the increasing number of digital data collections. Yet once started, these collections require continued—and likely increasing—support. At issue too are policing data to maintain standards of data quality, formatting data for eventual incorporation into metacollections, and presenting the information in ever-more-sophisticated, yet understandable, displays. More students and researchers need to know how to use the information, and database management should be recognized as a career on a par with lab research. The challenge for NSF and other agencies (see sidebar on p. 189) is how to satisfy all these needs without busting their budgets.

Last week, NSF's oversight body, the National Science Board (NSB), approved a draft report that calls for a comprehensive plan to manage this increasingly important scientific asset (www.nsf.gov/nsb/meetings/2005/LLDDC_Comments.pdf). "The rate that the data are increasing is exponential," says board member Michael Rossmann, a structural biologist at Purdue University in West Lafayette, Indiana. Adds Anita Jones, a computer scientist at the University of Virginia in Charlottesville and former board member, "I am concerned about the growing bill." The board is eager for community input.





Data boom. As with other data, molecular structures have gone from simple (*right*) to complex (*left*) and require sophisticated storage technologies (*above*).

CREDITS (TOP TO BOTTOM): COURTESY OF STORAGE TEK; B. DELABARRE, A. T. BRUNGER, *J. MOL. BIOL.* **347**, 437 (2005)

A growing concern

Digital databases date back to the era of punch cards and computer tapes. In the 1970s, crystallographers agreed to deposit their data in the newly created Protein Data Bank (PDB) at Brookhaven National Laboratory in Upton, New York. The bank is now managed by the Research Collaboratory for Structural Bioinformatics located at Rutgers University in New Brunswick, New Jersey, and the University of California, San Diego. Each week a staff of 25 adds 100 new molecular structures to the 30,000 already deposited. About 10,000 individuals visit the database daily, says its head Helen Berman, who calls PDB "the center of the new biology."

Such a growing enterprise requires continued funding. PDB's annual budget has grown 200-fold since 1976, to about $6 million. Some $2 million comes from NSF, and eight other organizations chip in the rest. "It's money well spent," says NSB Chair Warren Washington. "We cannot afford to have these data sets lost or poorly handled."

A climate modeler at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, Washington knows how valuable long-term data sets can be for simulations and other research efforts. NSF provides about two-thirds of NCAR's $139 million annual budget, but NSF's contribution to its dozens of databases is harder to quantify, says Richard Anthes, president of the University Corporation for Atmospheric Research, which oversees NCAR. "It is on the order of about $10 million," he estimates.

NCAR databases contain an estimated 1.6 petabytes of oceanographic, climate, and other information. The size of its Scientific Computing Division data-support section doubled last year, says its manager Steve Worley, who adds, "I imagine that's happening for almost everybody. I don't see any end." As with other databases, new entries need to be formatted and incorporated into the existing databases, which are updated regularly to take advantage of the latest storage technology.

These two projects illustrate the growing importance—and expense—of keeping data accessible, possibly in perpetuity, to all who want to use them. NSF doesn't have a good handle on its portfolio, says NSB executive officer Michael Crosby, who guesses that the agency could be supporting "hundreds, even thousands," of digital data collections. They range from those built to suit an individual researcher's needs to ones that are essential to many disciplines. The mode of funding is equally haphazard, says NSB member and ecologist Daniel Simberloff of the University of Tennessee,

Published by AAAS

Canadian Report Calls for Data Agency

OTTAWA—Canada needs an agency dedicated to ensuring maximum access to the fruits of publicly funded research.

That's the conclusion of a task force formed by a bevy of scientific organizations, which last week urged government officials to create a national data preservation and management organization. Such an agency would craft a national strategy relating to the acquisition, maintenance, and dissemination of all types of research data, from published scientific material to electronic archives and databases. A blue-ribbon panel chaired by David Strong, president of the private University Canada West in Victoria, British Columbia, suggested that the Canada Foundation for Innovation, which helped back the 9-month study, take the first step by providing start-up money.

The initial questions to be examined include many of those addressed in a draft report from the oversight body of the U.S. National Science Foundation, such as standardization of format, training, and funding for databases (see main text). Proponents hope that federal legislators will create a statutory agency—called Data Canada—with a $2.5-million-a-year budget to investigate a "central data preservation and management facility and a series of access and service nodes located in research institutions" across the country. The panel didn't speculate on how much it would cost to create and operate such a system. –WAYNE KONDRO

Wayne Kondro is a freelance writer in Ottawa.

Knoxville. "What we are asking NSF to do is come up with a single strategy" for evaluating and prioritizing these projects, he says. Part of that strategy should include criteria to determine continued support. There should also be guidelines about the right balance between data maintained and the acquisition of new data, says Rossmann.

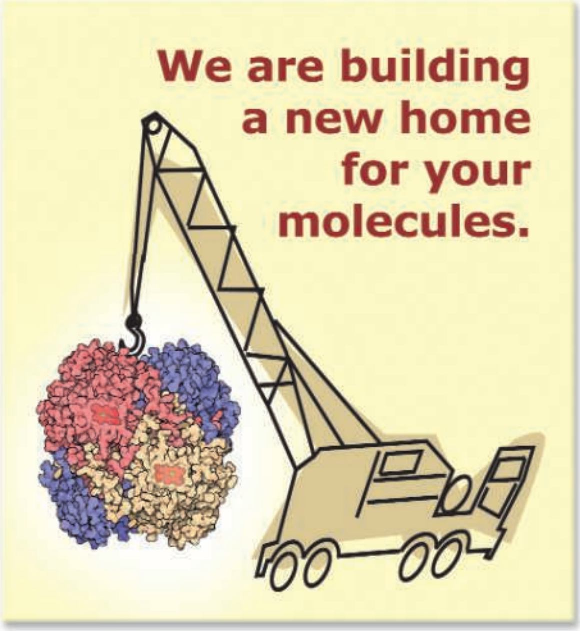
Projects such as PDB and the NCAR collection illustrate how a decision years ago to support a database can have significant, long-term implications for NSF's budget. "Clearly the current trend is to spend a large proportion [of NSF's database support] on maintaining databases," says Rossmann. When times are tight, however, that emphasis could mean fewer research awards.

Data-rich but poor

That doesn't mean database managers are feeling flush, however. "We have money troubles all the time," NCAR's Worley says, citing his desire to incorporate data from different collections into a single, seamless data resource. But that goal has taken a back seat to maintaining what's already on hand. Likewise, a compendium of *Arabidopsis* data at the Carnegie Institution Department of Plant Biology in Stanford, California, and the National Center for Genome Resources in Santa Fe, New Mexico, recently received about $3 million less from NSF than the almost $11 million its managers had requested for the next 5 years. "We ended up having to give up a lot of innovative stuff," says *Arabidopsis* Information Resource lead investigator Seung Yon Rhee, a Carnegie plant biologist.

One solution to funding shortfalls is to find other backers. Both NCAR and PDB supplement NSF's contribution with money from other federal agencies and international organizations. In other cases, host institutions are expected to cover costs for maintenance and upkeep. That's been the approach taken by NSF's Biological Research Collections program, which has helped keep natural history specimens in good shape but which now limits awards to one-time support of specific goals and projects.

To make NSF's money go further, biological research collections program manager Mark Farmer spends about half of his $4.5 million budget on a new long-term digital data collection—a "virtual" natural history museum with a portal that will provide desktop access to the world's preserved plants, animals, rocks, and so on. At the same time, museums and universities have agreed to bear the cost of operations for their collections, including keeping the links current and the original specimens in good shape. "We don't want to get into the business of paying for permanent staff at an institution," says Farmer.



Avoiding obsolescence. To be useful, digital databases require constant improvements to data storage, quality, and accessibility.

CREDIT: PDB

The science board's goal, says Simberloff, is "to make sure that the data collections we are funding are of the highest quality, that standards for storage and access are good." Toward that end, its report asks NSF to tally up all databases under its wing and to establish consistent rules to evaluate and fund them. That may include clarifying who is in charge of policing the data and requiring a database management plan covering the kind of data to be included, the standards for quality, and the criteria for what will be archived.

Key human resources issues also need to be addressed, says NSF program director Chris Greer. One big issue is encouraging database managers to develop new ways to disseminate the information more broadly. Greer cites the PDB's "Molecule of the Month," which provides online images and lay-language summaries of a protein's structure, function, and relevance to human health, as an excellent example of outreach to students.

A second issue is preparing undergraduates, graduate students, and postdoctoral fellows to take advantage of all these databases. NSF's 2006 budget request, now pending, includes a new program to expand competence in computing and other skills needed by 21st century scientists. Greer says that even more focused training programs may be needed.

Finally, Greer and others say that those who maintain these databases should be recognized as credible scientists whose work warrants tenure and other career advancements. "They are collectively an outstanding resource," Greer says. Toward that end, PDB's Berman says she encourages her employees to write research papers and speak at conferences and offers opportunities for career advancement. "It's very important to keep them motivated," she points out.

The science board's report sends NSF a signal that there's work to be done. "The NSF strategy and policies have not kept pace" with what's needed, the report points out. But Berman is optimistic that NSF will catch up. "The best thing is that NSF is now prepared to think about this."

–ELIZABETH PENNISI

Published by AAAS