



7874-108H-49 7874-108H-50 7874-108H-51 7874-108H-52 7874-108H-53 7874-108H-54 7874-108H-55 7874-108H-56

7874-108H-59 7874-108H-60 7874-108H-61 7874-108H-62 7874-108H-63 7874-108H-64 7874-108H-65

7874-108H-66 7874-108H-67 7874-108H-68 7874-108H-69 7874-108H-70 7874-108H-71 7874-108H-72

7874-108H-78 7874-108H-79 7874-108H-80 7874-108H-81 7874-108H-82 7874-108H-83 7874-108H-84 7874-108H-85 7874-108H-86 7874-108H-87

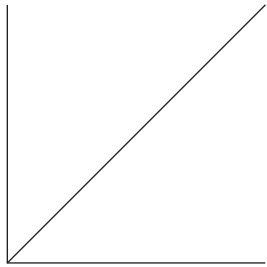
7874-108H-88 7874-108H-89 7874-108H-90 7874-108H-91 7874-108H-92 7874-108H-93 7874-108H-94 7874-108H-95 7874-108H-96 7874-108H-97

7874-108H-101 7874-108H-102 7874-108H-103 7874-108H-104 7874-108H-105

7874-108H-111 7874-108H-112 7874-108H-113 7874-108H-114 7874-108H-115 7874-108H-116 7874-108H-117

7874-108H-121 7874-108H-122 7874-108H-123 7874-108H-124 7874-108H-125





Scientists need to store all of their data, not just what's published.

THERE'S GOLD IN THOSE ARCHIVES

BY KARYN HEDE // PHOTOGRAPH BY FREDRIK BRODÉN

In fall 2003, Beth Chen, a graduate student at the Watson School of Biological Sciences, Cold Spring Harbor Laboratory, went on a treasure hunt. Her quest was to fill knowledge gaps in the neural circuitry of the roundworm *Caenorhabditis elegans*—that ubiquitous experimental-model organism. Chen indeed “discovered” several new neural synapses and neuromuscular junctions, but she did it without so much as lifting a pipette or looking through a microscope.

The secret of her success was an archive of Sydney Brenner’s work, a gold mine of many of the Nobel laureate’s laboratory notebooks and thousands of his electron-microscopy (EM) images, mostly unpublished, on *C. elegans* anatomy. Chen spent several months poring over a dozen laboratory notebooks and more than 10,000 electron micrographs at the Albert Einstein College of Medicine’s worm image archive, painstakingly reconstructing the neural connections that will inform her own research in the lab of Dmitri Chklovskii, an assistant professor at Cold Spring Harbor Laboratory and incoming group leader at HHMI’s Janelia Farm Research Campus.

Chen prevailed because the Brenner archive was safely in the hands of David H. Hall, director of the Center for *C. elegans* Anatomy at Einstein, after it had sat moldering in boxes at the Medical Research Council (MRC) Laboratory of Molecular Biology in Cambridge, England, for more than 15 years. Hall had long cajoled people

at MRC, convincing them to let him become the keeper of all that potentially useful worm image data.

There was a time when EM was a workhorse of biological research. But in the early 1980s, the genome revolution forced a radical change as scientists abandoned EM slides for DNA sequencing gels in a quest to get at the genetic secrets of *C. elegans*. Lost in this shift of resources was a massive amount of primary data, including maps of the complete neural circuitry of *C. elegans* collected mainly by Brenner and John G. White, now at the University of Wisconsin–Madison, while at MRC.

“By the mid-1990s, I was the only person left who could make sense of the records,” says Hall, an expert on *C. elegans* anatomy. “It would



HELPFUL HINTS FROM THE PROFESSIONALS

With more than 45 years’ experience helping scientists archive their life’s work, the American Institute of Physics has put together a short guide to assist scientists in maintaining their personal archives: <http://aip.org/history/source.htm> ¶ For biomedical researchers searching for a home for their archives, Paul Theerman, head of images and archives at the National Library of Medicine (NLM)’s History of Medicine Division, suggests first contacting the library at your home institution. If that fails, NLM will try to work with you to find a permanent home for them (<http://www.nlm.nih.gov/hmd/about/contactus.html>). ¶ For advice on storing primary data, HHMI has put together a guide to managing the scientific laboratory: www.hhmi.org/labmanagement ¶ For more information about the Jackson Laboratory’s Mouse Gene Expression Database, visit <http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml>

“THERE IS NO WAY WE CAN HAVE THE FUTURE TO KNOW WHAT WILL BE MOST IMPORTANT 10 OR 50 YEARS FROM NOW.”

all have been lost. I took a personal interest to make sure that didn't happen. My data sets and the MRC data sets were extremely complementary. It just made sense to put them together in one place."

The research community has lately come full circle, however, because scientists are now eager to connect their molecular data to the detailed anatomical studies that Brenner, White, and their colleagues labored over for years. Researchers are flocking to access Hall's treasure trove of data; he gets 20 to 30 visitors and thousands of Web site hits per week. Hall is in the process of digitizing as many of the approximately 200,000 images as possible, with about 5,000 now available at his Web site (www.wormimage.org).

"We couldn't have done everything," says Brenner about cataloging the collection. "There was too much data there for the one project. But it testifies to the integrity of the result that it can be used over and over again. And it shows the importance of keeping primary data where others can use it."

Remarkably, no one besides Hall seems to have foreseen that the worm images would become so valuable. Although most scientists would agree that primary data should be saved, in some cases data can become outdated to the point that no one can interpret them.

"There will always be a need to go back and look at primary data," says John Spieth, group leader at the Genome Sequencing Center at Washington University in St. Louis School of Medicine. "There is no way we can have the foresight to know what will be



CONSTANCE CEPKO The challenge of storing the vast amount of data generated by her research has her searching for commercial solutions.

important 10 or 50 years from now."

But saving data goes way beyond collecting a pile of graduate student notebooks and theses on a dusty top shelf. A quiet crisis looms in many labs as the volume of data generated by large-scale science grows at an alarming rate. Individual laboratories are struggling to find efficient and economical ways to store and retrieve key data. Many researchers have coped alone thus far, but some are now looking to large centralized archiving systems to bear part of the burden. And the rapidity of technology development, for example, has prompted Spieth to resequence parts of the *C. elegans* genome rather than rely on decade-old sequence data produced by technology that is now considered antiquated. Reacquiring data may work for large centralized data centers, but in individual labs, changing technology has often meant keeping antiquated equipment around so that data are not lost.

"Computer hardware and software quickly become obsolete, so that unless you hold on to your old computers the data you backed up with them may become difficult if not impossible to recover," says Terrence J. Sejnowski, a computational neuroscientist and HHMI investigator at the Salk Institute for Biological Studies in La Jolla. "It's something we have to live with."

ARCHIVING LARGE DATABASES

Retaining all that material is easier said than done, however.

"It's a problem for everybody," says HHMI investigator Constance L. Cepko, a neurobiologist at Harvard Medical School who studies the structure and function of the eye in vertebrates. "In trying to link DNA clones, in-situ images, and microarray data, we can generate 30,000 data points in one experiment." She and her colleagues considered commercial data-management packages and high-tech start-up services for archiving such data, but none filled their needs. At present, an M.D.-Ph.D. student is setting up a customized relational database, but it is just a temporary solution.

Cepko says that because the volume of data her lab generates is rapidly filling servers, she is looking to a centralized archiving

FORESIGHT MAY BE NEAR

system, such as the Mouse Gene Expression Database at the Jackson Laboratory (TJL) in Bar Harbor, Maine, to take some of the data off her hands. TJL aims to make the database, funded by the National Institutes of Health, the leading archive of mouse genomic and proteomic data, and is actively soliciting and adding primary data to its curated, annotated database.

In much the same spirit, Sejnowski has an agreement with the San Diego Supercomputing Center, which maintains and archives all of his lab's large data sets. "You have to find a partner," he insists. "Data have become so unwieldy that managing them is too much for any one lab to handle on its own."

HHMI investigator Norbert Perrimon, who studies cell signaling at Harvard Medical School, found the solution to his data-management problems—at least, for the time being—by setting up a centralized public database to store the results of his lab's RNA interference screens in *Drosophila*. Its infrastructure was funded by a grant from the National Institutes of Health, which allowed him to hire two full-time programmers to get the job done.

But in the long run, the solution will depend on cheaper ways of storing data as well as being more selective, says Perrimon. "The issue that we are facing now is that we do not yet know what is worth keeping in these large-scale studies because the [RNAi] field is not very mature yet. We need to spend more time on data analysis to figure out what has real value in the data sets." So, for the time being, he is storing it all.

Paul W. Sternberg, an HHMI investigator at the California Institute of Technology, believes the answer may lie in more intelli-

gent searching. "My general feeling is that we know a lot more than we think we do in biology," he says. "We aren't taking full advantage of what already exists out there. Digital storage is cheap. We should be archiving and making retrievable unpublished primary data." He is working on systems that will allow scientists to combine primary data from disparate sources, allowing them to develop new hypotheses by combining what he calls "weak hints," which tend to be overlooked when sources are assessed individually.

In the March 10, 2006, issue of *Science*, Sternberg and colleagues described how to apply such a computational approach to integrating published data on how genes interact with each other in roundworms, fruit flies, and yeast. "We now know that mining published and available data is valuable," Sternberg says. "Imagine what we could do if we could access the likely larger amount of unpublished information."

Sternberg believes this idea also extends to updating that laboratory mainstay, the lab notebook. "The new generation is more comfortable with electronic notebooks," he says. One of his graduate students keeps a personal blog on the lab's private intranet for recording observations and ideas. "I would have kept that kind of thing in a margin of my [paper] notebook," says Sternberg. "But then how would I ever find it again? In digital form, you can search and organize thoughts and ideas—and have instant recall."

A COMPLETE RECORD

In his 1965 Nobel Prize address, physicist Richard P. Feynman revealed one of the rarely uttered secrets of scientists. "We have a



When Articles and Data Go AWOL



JEREMY NATHANS If you're counting on your published articles serving as a record of your research, he warns, think again.

Jeremy Nathans, an HHMI investigator at the Johns Hopkins University School of Medicine, remembers searching for an article he considers to be a landmark publication. Citation in hand, he figured the fastest way to find it would be a quick PubMed search to link to the original article, which appeared in the journal *Nature* in 1978. He found nothing. The article had been missed in the process of adding pre-computer-era articles to the PubMed database, which includes citations and abstracts for virtually all published biomedical literature. Eventually, Nathans tracked down the article by contacting the author, who scanned the original print document and sent a grainy PDF file. But Nathans was

still left with an uneasy feeling. Because scientists rely so heavily on PubMed searches, he reasoned, if it doesn't appear there "it's as if it had never existed." (*Nature* has since added that particular article to its electronic archive.) Research results can also disappear when they are relegated to the ranks of "supplemental data" when a journal article is published. These data are only available online, and do not always print out along with the main article. "A lot of us believe that the best way to store data is by publishing them," says Nathans. "But now journals are telling us to put so much in supplemental data, and that gets divorced from the published article." "This issue of supplemental data is becoming

habit,” he observed, “in writing articles published in scientific journals to make the work as finished as possible to cover all the tracks, to not worry about the blind alleys or to describe how you had the wrong idea first.”

Scientists are so acculturated to think of published literature as the ultimate archive of their life’s work that they sometimes overlook the need to save the many other pieces. But science historians and archivists are often highly interested in, say, bumps in the road, which are often hidden in, or missing from, the clean and logical progression of ideas presented in the scientific literature. “Scientists have trouble understanding us,” says R. Joseph Anderson, an archivist at the American Institute of Physics. “We want all their documents that are likely to have historical value. It’s a matter of keeping a complete record.”

Thus, scientists should keep materials such as early versions of manuscripts, correspondence, photos, minutes of scientific meetings, and especially correspondence and lab notebooks, which not only help scientific colleagues in their research but may also help science historians glean the thought processes that go into developing science policy.

“People think of archives as quaint,” says Clare Flemming, curator of research collections at The Explorers Club, in New York City. “What curators at scientific collections do isn’t splashy, but



NORBERT PERRIMON A grant funded the centralized public database he set up to manage his RNA interference data.

when you think about it, material without provenance is meaningless. If someone comes along later and disagrees with your result, and no one can find the data, what happens then? The whole foundation of science is predicated on information being able to be duplicated.”

Still, “It’s hard to know what’s going to be helpful in the future,” says Miriam Spectre, an archivist who organized and described the Barbara McClintock Papers at the American Philosophical Society in Philadelphia. Spectre says that while most of the 1983 Nobel laureate’s laboratory notebooks describing her seminal work on transposable genetic elements survived, McClintock destroyed most of her correspondence before she died. “We sure would like to have had that,” Spectre says. ■

bigger and bigger,” says Edwin Sequeira, policy coordinator for PubMed Central, an electronic complement to PubMed that offers free access to full-text journal articles at the National Library of Medicine. “I see it as an economic decision not to put all of the data into print, but I would argue that if the data are important enough to include at all, they are an integral part of an article and should be treated as such.” ¶ Further, says Sequeira, not all journal publishers provide supplemental data when sending their articles for archiving. If a publisher goes out of business, there’s no guarantee that those types of materials in its possession will survive. He thinks that as long as scientists are providing such supplemental

materials, they should make sure the journals are supplying them to PubMed Central along with the article they complement. ¶ Traditionally, publishers have relied on libraries to maintain long-term archives, but in the digital age that role is in transition. Librarians, publishers, and the scientific community are grappling with how libraries will maintain the role of storing published articles and their supplemental data in the digital age. ¶ One potential solution is now being explored by a consortium organized by Stanford University Libraries. The system, called LOCKSS (<http://lockss.stanford.edu>) collects newly published content from participating publishers by using a Web

crawler that compares the content it has collected with the same content collected by other LOCKSS users and repairs any discrepancies. The system, initiated by a small team of librarians and engineers, provides a mechanism to guarantee libraries long-term access to complete content by making multiple copies of published data stored at all participating sites. If one site has a technical problem, data can be restored from any of the other sites. Some scientific publishers have begun to buy into the system, which is still in its infancy. To date, 80 major research libraries in the United States and 25 in Europe, as well as others scattered around the world, are participating. (continued on page 56)