THE CHRONICLE OF HIGHER EDUCATION

Information Technology

http://chronicle.com/weekly/v52/i42/42a03501.htm

From the issue dated June 23, 2006

Lost in a Sea of Science Data

Librarians are called in to archive huge amounts of information, but cultural and financial barriers stand in the way

By SCOTT CARLSON

Science is experiencing revolutionary changes thanks to digital technology, with computers generating a flood of valuable data for scientists to interpret.

But that flood could drown science.

Data from experiments conducted as recently as six months ago might be suddenly deemed important, but researchers might never find those numbers — or if they did, might not know what the numbers meant. Lost in some research assistant's computer, the data are often irretrievable or an indecipherable string of digits. That's a scenario increasingly familiar to scholars, says James M. Caruthers, a professor of chemical engineering at Purdue University.

"We are starting to die from data," he says bluntly.

To vet experiments, correct errors, or find new breakthroughs, scientists desperately need better ways to store and retrieve research data, Mr. Caruthers says, or "we are going to be more and more inefficient in the science that we do in the future."

Dealing with the "data deluge," as some researchers have called it, will be among the great challenges for science in the 21st century. Many in the field say that scientists should not be left to manage the data on their own.

Instead, librarians will have to step forward to define, categorize, and archive the voluminous and detailed streams of data generated in experiments. Already, librarians on some campuses — among them Purdue, the Johns Hopkins University, and the University of California at San Diego — are beginning to take on that role.

Some academics envision a day when scientists will deposit their data into broad archives that hold findings from many disciplines. Merely gathering data for databases, and leaving analysis for others, could become a common academic end in and of itself. And other scientists may spend their entire careers harvesting and analyzing the raw data collected by others.

But such science may be decades away. Scientists and librarians first have to overcome the challenges of archiving science data. Science data in many cases are vast and poorly organized, and sometimes stored using insecure and outdated technologies. Scientists often closely guard their data — the numbers can hold keys to competitive secrets or proprietary information, or reveal the embarrassments of experiments poorly executed.

A lot of attention is given to archiving data from "Big Science," like genomics and climate research, Mr. Caruthers says. The greater challenge is archiving data from "Small Science," like his work and the work of his colleagues.

Data from Big Science is highly organized on the front end — researchers define it before it even starts rolling off the machines — to make it easier to handle, to understand, and to archive. Small Science is "horribly heterogeneous," and far more vast. In time, Mr. Caruthers predicts, Small Science will generate two to three times more data than Big Science.

The concept of creating shared archives of raw data in those fields is still relatively new, and scientists are still debating who should do the storage — or whether data should be shared at all. But a few colleges and universities have begun experimenting with data libraries.

Those projects unite people who don't usually work together, says Clifford A. Lynch, director of the Coalition for Networked Information. "Scientists and scholars on one side and library and IT folks on the other are all feeling their way for the right roles for everybody."

"The big thing hanging over all of this is funding," he says, adding that agencies like the National Science Foundation are accustomed to supporting science projects and experiments, not infrastructure, like centralized archives.

The task and expense of archiving science data are so daunting that many institutions are wary of taking it on. "We haven't gotten into data repositories in large part because the amount of storage needed for that is huge," says Susan Gibbons, an associate dean of libraries at the University of Rochester.

"Some of these data sets are scary — many, many terabytes of information," she says. A terabyte is one trillion bytes, or the storage capacity of hundreds of DVD's. "You have to be really, really confident in the support of your institution to say, Give us your terabytes and we will find a way to preserve those for you."

At the Johns Hopkins University, librarians are starting with data from astronomy, where the need for an archive is clear: Images of space are massive, yet astronomers need them as a record of what's been observed in the sky. Some new telescopes will get a shot of the sky every few nights, yielding a couple terabytes of data every time. Those raw data are collected and archived by various agencies.

From those raw data, astronomers regularly manipulate the images to enhance aspects of them, then publish articles based on those resulting smaller images. Until Johns Hopkins began its effort, there was no archive for those manipulated images.

"That's the data that you want the community to be able to inspect, because that's the data that scientists are using to make their interpretations," says Robert J. Hanisch, an astronomer with the Space Telescope Science Institute whose office is on the Johns Hopkins campus. Researchers might also use the enhanced images to make new discoveries.

But journal editors are not interested in collecting the images because they see it as a hassle and not part of journals' central missions. And researchers may or may not keep the files. "It's totally ad hoc," Mr. Hanisch says. "Some people have their data, and some people don't."

Costs for the data-library project at Johns Hopkins aren't yet clear. "One of the major rationales for the pilot project we are starting is to understand the business model and the long-term operational costs," Mr. Hanisch says.

Card Catalog of the Future

Purdue librarians were encouraged to tackle the challenge of science data by James L. Mullins, dean of libraries there. Mr. Mullins had worked at the Massachusetts Institute of Technology, where various archiving projects are under way as part of a well-known project called DSpace. Although librarians are working with scientists and technology staff members to apply for grants, the archiving project for now is supported almost entirely by Purdue.

Purdue's data-repository model defies some traditional conceptions of an archive.

The data will not be stored in a central location on campus, like books stored in the stacks. Instead, the data will reside on the hard drives of faculty members, on departmental servers, or on the TeraGrid, a large-scale computing project run by a handful of institutions, including Purdue.

D. Scott Brandt, an associate dean at Purdue's library who is directing the project, describes it as a "distributed institutional repository." When it comes to data repositories, "the libraries don't think that they should be in the business of storage," Mr. Brandt says. They would rather leave that to the information-technology department.

Purdue librarians will do what librarians do best: consult with the scientists and examine the data to

produce metadata — that is, information that will tell others how to find and interpret the data. Think of meta-data as card-catalog information for the data, Mr. Brandt says.

The metadata will be published in an online catalog that the public can see and search. In recording data, "our commitment would be to Purdue first," Mr. Brandt says, but the catalog could list data housed at other institutions or be combined with other catalogs. Mr. Caruthers is one of the professors Mr. Brandt and other librarians are working with at Purdue. A researcher in plastics, Mr. Caruthers says his research group "burns as much computer time as anyone." He is assembling huge high-definition screens on which he can display large amounts of data at one time.

With a data deluge coming, he confesses to wanting to sell his colleagues on data repositories, but he wants to sell them softly. "If you start a project that is supposed to be the universal database for a particular discipline," he says, "it is a project that will be doomed to failure because no one is going to be the first person to put data into the database."

"You need to start with the natural social unit, the research group," he says. Make participation voluntary, low in cost, and risk-free. Show researchers that they can retrieve old data that would be valuable to them, yet hide data from competitors.

Researchers are more comfortable keeping data close by, so let them store the data on a server in their department, he says, rather than a centralized server, even if it is more secure and better managed. In trying to enlist researchers to join an archive, "what matters is not the truth but the perception" of whether their data are secure, he says.

Purdue has made more progress than most institutions on data repositories. But sitting at a table in conversation with Mr. Caruthers and Mr. Brandt, one can see how differing ideas about data archiving still need to be hammered out, and how difficult the job will be.

Mr. Caruthers has worked closely with industry and, like many scientists, has worked with data that hold secrets to key discoveries. If data are stored in an archive, researchers should be allowed to keep all or part of it secret, he says.

Except for the metadata, Mr. Brandt interjects. "The data that describes what your data is and what it can do would be public — or should be public," he says.

Mr. Caruthers shrugs skeptically. Even the simplest metadata describe what a researcher is working on, and can provide advantages to competitors. "With some of our commercial arrangements, what's important to them is not the data but the metadata. The data is just a string of numbers."

Is Purdue's the Right Answer?

Purdue's distributed model responds to technical, financial, and social realities, but observers of

information technology and archiving wonder if it is the right solution. If professors are archiving data on machines in their offices or departments, who makes sure those machines are backed up or up to date? What's to stop a professor from storing important data on a machine that lapses into obsolescence?

Mr. Lynch, of the Coalition for Networked Information, says he understands why Purdue is pursuing distributed archiving, but adds, "I'm uneasy about some of the ramifications about that kind of approach."

"Some of this data is going to outlive their projects, and it's going to have to go to the custody of central administrative entities, like the library," he says. "In the long run, it's hard for me to see how libraries or disciplinary data archives operating at a national or international scale aren't going to have to take responsibility for archiving the data."

But Purdue's distributed model is in line with conclusions in "Towards 2020 Science," a report written by distinguished scientists and issued by Microsoft Research this year. The report states that a big storage facility is not the best solution for data archiving.

"We believe that attempts to solve the issues of scientific data management by building large, centralized, archival repositories are both dangerous and unworkable," the report says. "They are dangerous because the construction of data collection or the survival of one's data is at the mercy of a specific administrative or financial structure; unworkable because of scale, and also because scientists naturally favor autonomy and wish to keep control over their information."

Data archiving is not a short-term problem, the report notes. In 2020, academe and science organizations will still grapple with the difficulties of managing and archiving huge volumes of data.

But the report also predicts that by 2020, new kinds of science and different models of publishing will emerge. Some journals may someday be composed primarily of raw data, for instance.

Perhaps a future scientist will routinely search a database to gather data from experiments in, say, climate studies, geological surveys, forestry projects, and crop yields. He or she might combine and crunch the data to generate breakthroughs in studies of the environment, global warming, or agriculture.

Those sorts of analyses, founded on data from disparate disciplines, could also lead to discoveries in medicine, astronomy, marine biology, anthropology, economics — virtually any field that seeks to understand the big picture of how things work.

"We are seeing more and more examples where the starting point for new research and new ideas is digital data that was produced by someone else," says Chris L. Greer, a cyberinfrastructure adviser for the National Science Foundation. "People find ways to use this information that the original researchers didn't think of."

Hard to Handle

The possibility that scientists may one day archive and share data is one of the reasons Sylvie M. Brouder has gotten involved with the data-repository projects at Purdue.

An agronomist who studies water quality in agriculture, she works in a discipline that requires decades of data collection to generate sound conclusions. Compared with other disciplines in science, the amount of data she collects is relatively small and manageable, which is why the library has started its pilot project with her. But it is still a lot for one researcher to handle.

She sends research assistants out to dozens of sites every day, to read sensors and collect water samples — information from which is entered on a series of computers in her laboratory. She gives a visitor a tour of storage rooms, which smell like grain, in the basement of her building.

Just one of the rooms, with ceilings that extend up a dozen feet, is filled top to bottom with wooden boxes, and in each box are scores of little bags of dirt. Each little bag represents numerous data points.

"I spend all my time getting data," with far less time to analyze it, she says. But she knows there are people in government, at corporations, and in academe who would be interested in what she has collected.

"Making it available to other people to do their own research would save years of research and literally millions of dollars," she says. Likewise, if archives were online, she could draw from data collected by other agronomists, the U.S. Geological Survey, and other state and federal agencies. Those data could be used to "scale up" her studies of watersheds from plots on farms, to the whole state of Indiana, to an entire region.

"Scaling up is what we want to do these days," she says.

But, like scientists in other fields, she finds it difficult to manage, decipher, and even store data. Some government agencies, such as the Geological Survey, post data online. But the data are often just numbers, she says, with no metadata attached to tell you what they are.

"No one has told them how to do it," she says "What they put up there is just literally what they pull off sensors. You can look at it, but if you want to figure out what it means, you have to figure out who posted it and who was responsible for collecting it."

Even the legacy of data from her own institution is confounding, she says. Her office is crowded with file cabinets filled with water-quality data that she inherited from other researchers. In the old days, when a scientist would retire, a new scientist in the same field would get hired and inherit reams of data, research assistants, and even the previous researcher's desk and chair.

Today, years might pass before a new scientist is hired to fill an empty position, and it's up to that person

to scrounge around the department for whatever data remain. Meanwhile, she says, amid the deluge of potentially valuable data, administrators are constantly pressuring professors to keep servers and offices cleaned of clutter.

"The word from up top is: Pitch it," she says. "As a practice right now, we are throwing out information."

http://chronicle.com
Section: Information Technology
Volume 52, Issue 42, Page A35
