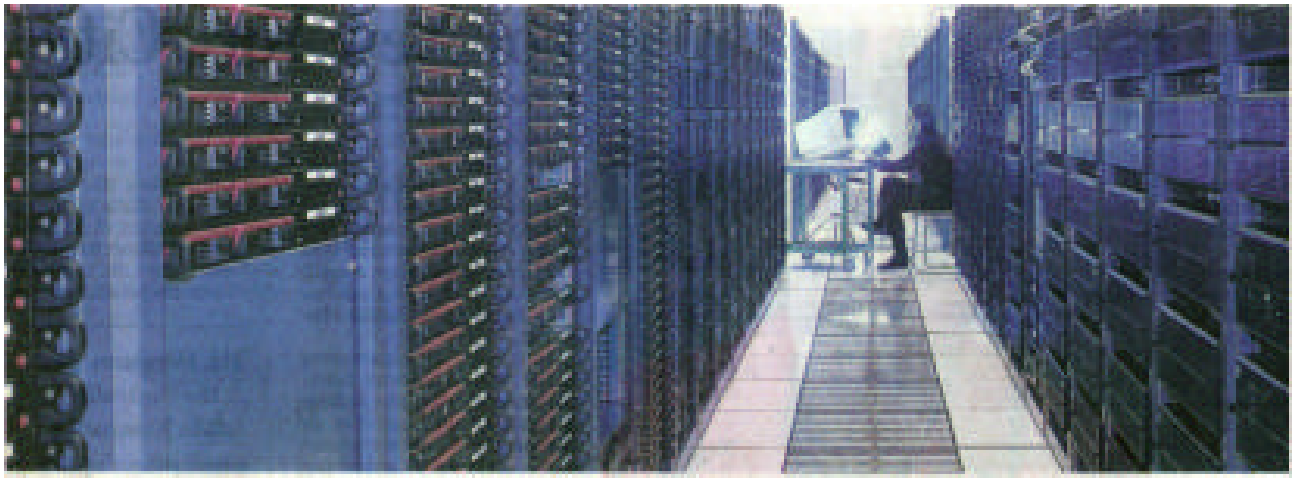
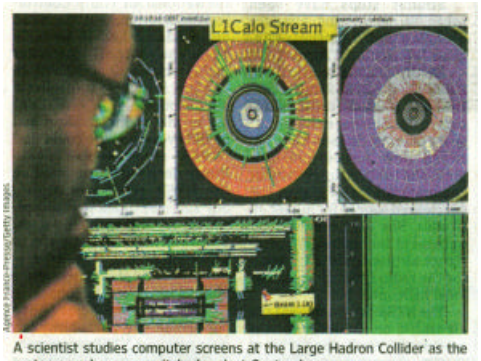
Wall Street Journal, 2009 August 28



The next generation of experiments, like the Large Hadron Collider, above, a powerful particle accelerator beneath the border of Switzerland and France, will be ev

A Data Deluge Swamps Science

As Paper Trails Fade, Digital Material Grows in Size and Complexity; How to Decipher Those 80



A scientist studies computer screens at the Large Hadron Collider as the proton smasher was switched on last September.

Oceans of Data: A selection of large digital archives

Large Hadron Collider: Will produce some 15 million gigabytes of data annually, enough for 1.7 million DVDs.

Sloan Digital Sky Survey: Includes data from the Sloan telescope in New Mexico related to 230 million celestial objects.

The Internet Archive: San Francisco-based archive holds almost two million gigabyes of Web files.

British Library Digital Lives: Repositories of personal digital files are being developed at the London-based library.

Protein Data Bank: The world-wide repository contains 3-D structures of large molecules and nucleic acids.

GenBank: Collects all publicly available DNA sequences at the National Institutes of Health.

Wall Street Journal, 2009 August 28

In a vault beneath the British Library here, Jeremy Leighton John grapples with a formidable challenge in digital life. Dr. John, the library's first curator of eManuscripts, is working on ways to archive the deluge of computer data swamping scientists so that future generations can authenticate today's discoveries and better understand the people who made them.

His task is only getting harder. Scientists who collaborate via email, Google, You-Tube, Flickr and Facebook are leaving fewer paper trails, while the information technologies that do document their complishments can be incomprehensible to other researchers and historans trying to read them.

Computer-intensive experiments and the software used to analyze their output generate millions of gigabytes of data that are stored or retrieved by electronic systems that quickly become obsolete.

"It would be tragic if there were no record of lives that were so influential," Dr. John says.

Usually, historians are hard-pressed to find any original source material about those who have shaped our civilization. In the Internet era, scholars of science might have too much. Never have so many people generated so much digital data or been able to lose so much of it so quickly, experts at the San Diego Supercomputer Center say. Computer users world-wide generate enough digital data every 15 minutes to fill the U.S. Library of Congress.

In fact, more technical data have been collected in the past year alone than in all previous years since science began, says Johns Hopkins astrophysicist Alexander Szalay, an authority on large data sets and their impct on science. "The data is doubling every year," Dr. Szalay says.

The problem is forcing historians to become scientists, and scientists to become archivists and curators. Digital records, unlike laboratory notebooks, can't be read without the proper hardware, software and passwords. Electronic copies are difficult to verify and are easy to alter or forge. Digital records "can be more direct, more immediate and more candid," Dr. John says. "But how can we demonstrate to people in the future that these are the real thing?"

Dr. John first encountered this archival problem nine years ago when the British Library received the working papers of William Hamilton, a leading evolutionary biologist who died in 2000. Among the 200 crates of handwritten letters, draft typescripts and lab notes, Dr. John discovered 26 cartons containing vintage floppy computer disks, reels of 9-track magnetic tape, stacks of 80-column punch cards, optical storage cards and punched paper tapes meant for computing devices dating to the 1960s.

These files likely contained crucial drafts of research papers, emails and other information that could illuminate an influential life of science, as recorded through 40 years of computing technology—as long as Dr. John can find a way to read them.

To extract the antiquated data required more than a password. Dr. John gradually assembled a collection of vintage computers, old tape drives and forensic data-recovery devices in a locked library sub-basement.

For more than a decade, policy makers and data experts have been debating the best way to preserve important digital records. "What you keep and how you pay for it are difficult issues," says Fran Berman, vice president of research at Rensselaer Polytechnic Institute in Troy, N.Y., who is co-chair of a federal commission studying the economics of data preservation.

The growing scale of new science projects, however, has university data custodians worried. "We are swimming in data these days, and people are overwhelmed," says digital curator Sayeed Choudhoury at Johns Hopkins University, the principal investigator for a national consortium of data preservationists called the Data Conservancy.

Consider a new computerized star atlas called the Sloan Digital Sky Survey. Using a telescope in New Mexico, the project in its first two days collected more data than gathered in all the previous history of astronomy, Dr. Choudhoury says. Its final data set catalogs 230 million celestial objects, encompassing 930,000 galaxies, 120,000 quasars and 225,000 stars, all encoded in 140 terabytes of digital data.

The next generation of experiments will be even more data-intensive. A new proton smasher near Geneva called the Large Hadron Collider is supposed to produce 15 million gigabytes of data annually—enough to fill more than 1.7 million DVDs every year. The Large Synoptic Survey Telescope, an astronomy program under construction in northern Chile slated to launch in 2016, will regularly image the entire sky, recording more than 30,000 gigabytes of data every night.

"Our ability to collect data now outstrips our ability to maintain it for the long run," says William Michener at the University of New Mexico, who leads a data-preservation network called DataONE. "We lose an awful lot of data that is collected with public funds."

Earlier this month, the U.S. National Science Foundation awarded $20 million to the Data Conservancy and another $20 million to the DataONE group to develop more effective data-preservation tools over the next five years, especially for researchers working on their own or in small teams.

"It is sexy to think about the big data sets, but a vast amount of data is contained in lots of really small data sets created by different researchers using different software," says Patricia Cruse, director of the digital-preservation program for the University of California system. "People retire and their knowledge about their data retires with them."

For future generations to get much use from 21st-century data, though, it won't be enough to simply archive email exchanges and file formats. "The problem is to actually capture the way scientists interact with the data," Dr. Szalay says. "Today's graduate students are starting to use instant messaging in their scientific work. We have to figure out how to capture these."

In the long run, no scientific data can outlast the storage media that contains it, unless it can be accurately recopied and reliably re-authenticated. Many computer CDs, DVDs and flash drives last only a decade or so. The best known star atlas, inscribed on a scroll discovered in Dunhuang, China, has survived for more than 1,000 years. It might have been traced from an even older star map.

Earlier this year, researchers at Keio University, Sharp Corp. and Kyoto University in Japan unveiled a memory chip designed to last for centuries. In April, physicists at the University of California, Berkeley, and the Lawrence Berkeley National Laboratory published the design of a digital device that could store data for a billion years, at least in theory.

"Digital information lasts forever—or five years," says RAND Corp. computer analyst Jeff Rothenberg, "whichever comes first."

Robert Lee Hotz shares video on this topic, recommended reading and responds to reader comments at WSJ.com/Currents. Email him at sciencejournal@wsj.com.